

Prosody, Phonology and Phonetics

Chunsheng Yang *Editor*

The Acquisition of Chinese as a Second Language Pronunciation

Segments and Prosody

 Springer

Prosody, Phonology and Phonetics

Series Editors

Daniel J. Hirst, CNRS Laboratoire Parole et Langage, Aix-en-Provence, France

Hongwei Ding, School of Foreign Languages, Shanghai Jiao Tong University,
Shanghai, China

Qiuwu Ma, School of Foreign Languages, Tongji University, Shanghai, China

The series will publish studies in the general area of Speech Prosody with a particular (but non-exclusive) focus on the importance of phonetics and phonology in this field. The topic of speech prosody is today a far larger area of research than is often realised. The number of papers on the topic presented at large international conferences such as Interspeech and ICPhS is considerable and regularly increasing. The proposed book series would be the natural place to publish extended versions of papers presented at the Speech Prosody Conferences, in particular the papers presented in Special Sessions at the conference. This could potentially involve the publication of 3 or 4 volumes every two years ensuring a stable future for the book series. If such publications are produced fairly rapidly, they will in turn provide a strong incentive for the organisation of other special sessions at future Speech Prosody conferences.

More information about this series at <http://www.springer.com/series/11951>

Chunsheng Yang
Editor

The Acquisition of Chinese as a Second Language Pronunciation

Segments and Prosody

 Springer

Editor

Chunsheng Yang
Department of Literatures,
Cultures and Languages
University of Connecticut
Storrs, CT, USA

ISSN 2197-8700

ISSN 2197-8719 (electronic)

Prosody, Phonology and Phonetics

ISBN 978-981-15-3808-7

ISBN 978-981-15-3809-4 (eBook)

<https://doi.org/10.1007/978-981-15-3809-4>

© Springer Nature Singapore Pte Ltd. 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.

The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Introduction

This volume emerged from the 32nd North American Conference on Chinese Linguistics (NACCL-32) held at the University of Connecticut. NACCL-32 was originally scheduled for April 24–26, 2020, which was postponed to September 18–20, 2020, due to the breakout of the COVID-19 pandemic in the US in March 2020. In September 2020, the conference was moved online to ensure the safety and health of all participants. The contributors of this book were not limited to NACCL-32 participants; experts and researchers in Chinese as a second language (CSL) phonetics, phonology and pronunciation in both the US and China were invited to contribute. While all of us had to juggle life and work during this unprecedentedly challenging time, all contributors worked really hard to follow the original plan of publication in all phases of this long process, hence enabling the edited volume to be published in a timely manner. Therefore, my heartfelt thanks and gratitude go to every author and coauthor for their work, punctuality, and, most importantly, contribution to an emerging field, namely, the acquisition of CSL pronunciation.

Research on Chinese as a second language (CSL) has gained great momentum in recent years. Studies on the acquisition of various aspects of CSL have flourished in both journals (Everson 1998; Lee et al. 2010; Li 2012; Jin 1994; So and Best 2010; Wang et al. 2004; Zhang 2010; Yang 2013, 2014; Yang and Chan 2010; Zhao 2012; Ke and Li 2011; Yuan 2010, among many others), and (edited) books (Everson and Xiao 2011; Han 2014; He and Xiao 2008; Jiang 2014; Ke 2018; Shei et al. 2019; Wen 2012; Wen and Jiang 2019; Yuan and Li 2019). Pronunciation is one of the most important components of a language which overlaps with all other aspects of a language (lexicon, semantics, syntax, and pragmatics). Worth noting is that, except for a few books on tone and prosodic acquisition mentioned above, there are no edited volumes or monographs exclusively devoted to CSL pronunciation. While journal articles and chapters in edited volumes/monographs touch upon some topics in CSL pronunciation, many of them, such as those on tone processing (Lee et al. 2010; So and Best 2010; Wang et al. 2004), are motivated more by theoretical issues, such as cross-linguistic speech perception, than by applied or pedagogical concerns in CSL instruction, as Jiang (2014) correctly points out.

Books on Chinese phonetics and phonology primarily examine theoretical issues, such as Chen (2000), Duanmu (2007, 2009, 2016), Wang and Norval (2013), and

Zhang (2017). While a few volumes concern the second language acquisition of Mandarin phonetics and phonology, these books either only discuss tone acquisition (Yang 2015; Zhang 2018) or only touch upon limited issues of prosodic acquisition (Yang 2016). To this end, an edited volume devoted to CSL pronunciation is in dire need so that CSL researchers, graduate students, and advanced undergraduate students can have a reliable and comprehensive source to refer to.

One of the major guiding principles when editing this volume is the integration of theory, practice and pedagogy. For this purpose, all chapters contain a section on pedagogical implications to put the research findings in perspective so that CSL instructors and practitioners can make research-informed decisions in their teaching practice.

This book consists of fourteen chapters covering a wide range of issues on CSL pronunciation. The fourteen chapters are grouped into three parts. Part I, consisting of seven chapters, concerns tones and segments in L2 Mandarin. The chapter by Min Liu and Rongru Chen analyzes the error types of rhotic onset /ɹ/ produced by Indonesian learners of Mandarin, such as taps, plosives, trills, fricatives, affricates, zero onset and laterals. Acoustic realization of the correctly produced rhotic onset /ɹ/ by the L2 Indonesian learners was found to have a stronger lowering effect on the F3 of the following vowels /a, ə, u/, suggesting a higher degree of rhoticity in the correct L2 productions. Such detailed error/acoustic analyses of segment production by L2 Mandarin learners are rather limited and, therefore, will be very helpful in assisting L2 learners in noticing their issues in production and enabling instructors to come up with ways to improve on L2 learners' segmental production.

As can be expected, a volume on L2 Chinese pronunciation will inevitably include more chapters on tone acquisition than segmental acquisition. Six chapters in the book are devoted to tone processing, acquisition and pedagogy. The chapter by Chenqing Song is a welcome attempt at interpreting advanced learners' tone errors within the framework of Optimality Theory. A distinction of tone sandhi and coarticulation was proposed to explain why Tone 2 was erroneously produced as a low tone in the T2-T4 sequence. The chapter by Chunsheng Yang examines the categorical perception of three tone continua, T1-T3, T2-T3, and T4-T3, and finds that there exist the T2-T3, T1-T3, and T4-T3 (only to some extent) continua in Mandarin Chinese, especially for L2 listeners. This means that Tone 1, Tone 2, and Tone 4, if produced differently than the correct forms (for example when the overall F0 contour for Tone 1 was lowered or the turning point in Tone 2 was delayed), would be likely to be perceived as Tone 3. This study adds to our understanding of one of the most difficult tones in Mandarin Chinese, namely Tone 3. Yadong Xu and Kevin Russell's chapter investigates the interference of two dialectal tone systems, Lanzhou dialect tones and Mandarin tones, on word recognition, by Lanzhou dialect speakers (but dominate standard Mandarin speakers). While only unidirectional effect from the dominant language (standard Mandarin) to the non-dominant language (Lanzhou dialect) was found, this study shows how languages/dialects influence each other and how power relationship of languages/dialects shapes language acquisition and processing.

The last three chapters in this part are on tone training and pedagogy. The chapter by Yingjie Li and Goun Lee is a high variability phonetic training study on tones

focusing on beginning level of learners in both monosyllabic and disyllabic stimuli. It was found that the disyllabic stimuli are more helpful in that they provide tone coarticulation variability. More future studies along the same line should be pursued on intermediate and even advanced learners in CSL field. Jiang Liu's chapter is an innovative attempt to integrate pronunciation teaching, specifically tone learning, in the L2 Chinese curriculum, namely, recently learned vocabulary was used as the stimuli to train CFL learners' perception and production of Chinese tones. The findings are very promising, showing that tone identification and word repetition tasks improve learners' pronunciation in disyllabic words both immediately after the training session and in a delayed test. The chapter by Nan Meng is a pedagogical exploration on the difference of blocked and random practice on tone error correction. The random practice, namely practice in more diverse contexts, is found to help learners internalize the tone acquisition. Implications of the study go beyond word lexical tone training and support the incorporation of diverse contexts in perceptual and pronunciation training.

The second part consists of two chapters on CSL prosodic acquisition. The chapter by Jing Yang and Bei Yang investigates the duration of disyllabic words produced by Russian learners of Mandarin Chinese. Main differences were found at the second syllable in the disyllabic words at the sentence-medial position, likely due to their prosodic phrasing; learners did acquire the final lengthening, indicating final lengthening is more of a language universal feature. Zhen Qin's chapter examines the effect of Mandarin learning experience on L2 learners' phonological knowledge of T3 sandhi in word production among Korean-speaking L2 learners. The findings suggest that experienced Korean-speaking L2 learners were better in using their phonological knowledge of T3 sandhi than less experienced learners in producing pseudo and novel words, but not real words. The findings shed lights on L2 Learners' underlying mechanism of using tones, and provide pedagogical implications for Mandarin teaching in classroom setting.

The third part contains five chapters on CSL intelligibility, comprehensibility, accentedness, and fluency.

The chapter by Kaidi Chen and Chunsheng Yang considers the effects of F0 (i.e., natural F0 versus flattened F0) on the intelligibility of Mandarin speech by L2 Mandarin learners from different proficiency levels in quiet and white noise conditions when controlling for sentence context. The findings confirm the effects of F0, listening environment, and proficiency level on intelligibility and highlight the importance of tone accuracy in L2 Mandarin teaching and learning. The chapter by Chunsheng Yang, Jing Chu, Si Chen and Yi Xu explores the effects of segments, intonation and rhythm on the perception of L2 accentedness and comprehensibility. Results of the Chinese native judges' ratings showed that segments contribute more to the perception of L2 accentedness and comprehensibility than intonation and rhythm, and that intonation contributed more to L2 perception than rhythm. It was also found that accentedness ratings highly correlated with comprehensibility judgment. The findings of this study confirm what some recent studies have found regarding the contribution of segments and prosody to L2 perception, but differ from some

previous studies in regards to the relationship between L2 accentedness and comprehensibility. Eric Pelzl's chapter is different from the other chapters in this volume in that it is a review chapter. Pelzl successfully makes a distinction of different types of pronunciation "errors", namely, accent-shifted pronunciation, systematic error, and unsystematic error. Pelzl points out that the first two types of "errors" are relatively easier to be understood by listeners (or listeners may adapt to such "errors"), listeners may not be able to adapt to the third type of errors, and the worst case scenario will be that listeners would just ignore these errors, such as the unsystematic tone errors. The relationship between tone errors and comprehensibility is also discussed, namely tone errors impact the speed and efficiency of word recognition. The last two chapters of the volume concern L2 Chinese fluency. Yuyun Lei's chapter finds that the amount and rate features of speech, as well as silent pausing features, were significantly correlated with oral proficiency ratings, and these features could also distinguish among the assessed levels. The results suggest that the amount and rate of speech and silent pausing features could be reasonably selected as proxies of fluency. The last chapter by Yu Liu focuses specifically on the role of vocabulary on L2 speaking fluency. The findings show significant correlations among vocabulary size and all three facets of utterance fluency: speed fluency (speech rate, mean length of runs), breakdown fluency (mean length of silent pauses, number of silent pauses), and repair fluency (number of disfluencies). However, among all fluency measures, only speech rate was significantly correlated to lexical retrieval speed. Moreover, stimulated recall responses revealed that around two-third of the disfluencies were reported to be caused by vocabulary-related issues. Liu's chapter confirmed that efficient task related lexical access was crucial for producing fluent speech in a second language.

While this volume has covered a full array of topics in CSL pronunciation, it is worth noting that the acquisition of CSL pronunciation is still an emerging field and there is a long way to go. Below are some directions that future studies on CSL pronunciation can pursue: (1) studies on the intelligibility, comprehensibility and accentedness should be expanded to L2 learners who are tonal language speakers, such as Thai and African language speakers, to test the generalizability of the findings in the chapter by Yang et al. of this volume; (2) function load or error gravity of segments, such as vowels and consonants, needs to be examined in Mandarin Chinese so that teaching priority can be set up, if needed; (3) pronunciation assessment is almost an uncharted area in CSL and a lot has to be done; (4) task-based form-focused pronunciation teaching is another promising field (see Gurzynski-Weiss et al. 2017 and articles in the same issue); (5) types of errors in L2 Chinese pronunciation and their relationship to L2 Chinese accentedness, intelligibility and comprehensibility (see Pelzl's chapter of this volume); (6) interdisciplinary collaboration between L2 speech researchers and CSL practitioners so that CSL instructors can make research-informed decisions in their pedagogy; and (7) the use of technology in pronunciation teaching and learning, such as telecollaborative communications (cf. Luo and Yang 2018) and automatic speech recognition and assessment.

References

- Chen, M. Y. (2000). *Tone Sandhi: Patterns across Chinese dialects*. Cambridge University Press.
- Duanmu, S. (2007). *The phonology of standard Chinese* (2nd ed.). Oxford University Press.
- Duanmu, S. (2009). *Syllable structure: The limits of variation*. Oxford University Press.
- Duanmu, S. (2016). *A theory of phonological features*. Oxford University Press.
- Everson, M. E. (1998). Word recognition among learners of Chinese as a foreign language: Investigating the relationship between naming and knowing. *The Modern Language Journal*, 82(2), 194–204.
- Everson, M. E., & Xiao, Y. (eds.) (2011). *Teaching Chinese as a foreign language: Theories and applications* (2nd ed.). Boston, MA: Cheng & Tsui Company.
- Grurzynski-Weiss, L., Long, A. Y., & Solon, M. (2017). TBLT and L2 pronunciation: Do the benefits of task extend beyond grammar and lexis? *Studies in Second Language Acquisition*, 39(2), 213–224.
- Han, Z. (Ed.) (2014). *Studies in second language acquisition of Chinese*. Bristol, Buffalo, Toronto: Multilingual Matters.
- He, A. W., & Xiao, Y. (Eds.) (2008). *Chinese as a heritage language: Fostering rooted world citizenry*. Honolulu: NFLRC/University of Hawaii Press.
- Jiang, N. (Ed.) (2014). *Advances in Chinese as second language: Acquisition and processing*. Cambridge Scholars Publishing.
- Jin, H. G. (1994). Topic-prominence and subject-prominence in L2 acquisition: Evidence of English-to-Chinese typological transfer. *Language Learning*, 44(1), 101–122.
- Ke, C. (Ed.) (2018) *The Routledge handbook of Chinese second language acquisition*. Routledge.
- Ke, C., & Li, Y. A. (2011). Chinese as a foreign language. *Journal of Chinese Linguistics*, 39(1), 177–238.
- Lee, C.-Y., Tao, L., & Bond, Z. S. (2010). Identification of acoustically modified Mandarin tones by non-native listeners. *Language and Speech*, 53(2), 217–243.
- Li, S. (2012). The effects of input-based practice on pragmatic development of requests in L2 Chinese. *Language Learning*, 62(2), 403–438.
- Luo, H., & Yang, C. (2018). Twenty years of telecollaborative practice: Implications for teaching Chinese as a foreign language. *Computer-Assisted Language Learning*, 1, 1–26. DOI: 10.1080/09588221.2017.1420083.
- Shei, C., Zikpi, M. E. M., & Chao, D.-L. (Eds.) (2019). *The Routledge handbook of Chinese language teaching*. Routledge.
- So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech*, 53(2), 273–293.
- Wang, J., & Norval, S. (1997). *Studies in Chinese phonology*. Walter de Gruyter.
- Wang, Y., Behne, D. M., Jongman, A., & Sereno, J. A. (2004). The role of linguistic experience in the hemispheric processing of lexical tone. *Applied Psycholinguistics*, 25(3), 449–466.
- Wen, X. (2012). *Chinese as a second language acquisition and instruction*. Peking University Press.
- Wen, X., & Jiang, X. (Eds.) (2019). *Studies on learning and teaching Chinese as a second language*. London, UK: Routledge Taylor & Francis Group.
- Yang, B. (2015). *Perception and production of Mandarin tones by native speakers and L2 learners*. Verlag Berlin Heidelberg: Springer.
- Yang, C. (2013). Acquiring the pitch patterns of L2 Mandarin Chinese. *Chinese as a Second Language Research*, 2(2), 221–242.
- Yang, C. (2014). Acquisition of Mandarin lexical tones: The effect of global pitch tendency. *Contemporary Foreign Language Studies*, 12(4), 81–95.
- Yang, C. (2016). *The Acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice*. John Benjamins Publishing Company.
- Yang, C., & Chan, M. K. M. (2010). The perception of Mandarin tones and intonation by American learners. *Journal Chinese Language Teachers Association*, 45(1), 7–36.
- Yuan, B. (2010). Domain-wide or variable-dependent vulnerability of the semantics-syntax interface in L2 acquisition? Evidence from “Wh” words used as existential polarity words in L2 Chinese grammars. *Second Language Research*, 26(2), 219–260.

- Yuan, F., & Li, S. (Eds.) (2019). *Classroom research on Chinese as a second language*. Routledge.
- Zhang, H. (Ed.) (2010). Phonological universals and tone acquisition. *Journal of the Chinese Language Teachers Association*, 45(1), 39–65.
- Zhang, H. (2018). *Second language acquisition of Mandarin Chinese tones: Beyond first-language transfer*. Brill.
- Zhang, H. (2017). *Syntax-phonology interface: Argumentation from Tone Sandhi in Chinese dialects*. Routledge.
- Zhao, L. X. (2012). Interpretation of Chinese overt and null embedded arguments by English-speaking learners. *Second Language Research*, 28(2), 169–190.

Contents

Segments and Tones

Production of the Mandarin Rhotic Onset /ɹ/ by Indonesian Learners of Mandarin	3
Min Liu and Rongru Chen	
What is in the Final Stage of Inter-Language? Tone Errors and Phonological Constraints in Spontaneous Speech in Very Advanced Learners of Mandarin	21
Chenqing Song	
Categorical Perception of Mandarin Tones by Native and Second Language Speakers	55
Chunsheng Yang	
What if What You Think is the Opposite of What I Say? Evidence from Putonghua/Lanzhou Bidialectal Speakers on the Online Processing of Two Tonal Systems	75
Yadong Xu and Kevin Russell	
The Effect of Perceptual Training on Teaching Mandarin Chinese Tones	107
Yingjie Li and Goun Lee	
Tone Category Learning Should Serve Tone Word Learning: An Experiment of Integrating Pronunciation Teaching in the L2 Chinese Curriculum	141
Jiang Liu and Cheng Xiao	
“Repeat After Me”: Is There a Better Way to Correct Tone Errors in Teaching Mandarin Chinese as a Second Language?	163
Nan Meng	

Prosody

Duration of Disyllabic Words Produced by Russian Learners of Chinese	177
Jing Yang and Bei Yang	

How Does Mandarin Learning Experience Modulate Second-Language Learners' Phonological Knowledge of Tone 3 Sandhi in Word Production?	195
Zhen Qin	

Intelligibility, Comprehensibility, Accentedness, and Fluency

The Effect of Fundamental Frequency on Mandarin Intelligibility by L2 Learners in Quiet and Noise Environments: A Pilot Study	213
Kaidi Chen and Chunsheng Yang	

Effects of Segments, Intonation and Rhythm on the Perception of L2 Accentedness and Comprehensibility	233
Chunsheng Yang, Jing Chu, Si Chen, and Yi Xu	

Foreign Accent in Second Language Mandarin Chinese	257
Eric Pelzl	

Exploring Fluency and Disfluency Features of Oral Performances in Chinese as a Second Language	281
Yuyun Lei	

The Role of Vocabulary Knowledge in Second Language Speaking Fluency: A Mixed-Methods Study	305
Yu Liu	

Editor and Contributors

About the Editor

Chunsheng Yang is an Associate Professor of Chinese and applied linguistics at the University of Connecticut in the United States. His main research areas are the acquisition of second language prosody, Chinese linguistics, sociophonetics, computer-assisted language learning, and applied linguistics in general.

Contributors

Kaidi Chen is a Ph.D. student in Applied Linguistics and Discourse Studies at the University of Connecticut in the United States. His primary research interests are speech production and perception, sociophonetics, bilingualism, and language teaching pedagogy.

Rongru Chen is a linguistic undergraduate at the College of Chinese Language and Culture at Jinan University in Guangzhou, China. Her main research interest is speech production and perception.

Si Chen is an Assistant Professor of Chinese and bilingual studies at the Hong Kong Polytechnic University. Her research interests are phonetics, phonology, statistical modeling, second language acquisition, speech disorder, and human-robot interaction.

Jing Chu is an Associate Professor of Chinese at Bohai University. Her research areas are Chinese philology, Chinese linguistics, and Chinese language teaching.

Goun Lee is an Adjunct Assistant Professor in the Department of English Language and Literature at Sungkyunkwan University in Seoul, South Korea. Her main research areas are acoustic and auditory phonetics, and second language phonetics and phonology.

Yuyun Lei is a Ph.D. candidate of East Asian Languages and Cultures at the University of Illinois at Urbana-Champaign in the United States. She is also a Flagship Instructor of Chinese at the University of Mississippi. Her main research interests are language testing and assessment, Chinese language acquisition and pedagogy, and computer-assisted language learning.

Yingjie Li is a Senior Instructor and Chinese Program Coordinator at the University of Colorado Boulder. Her main research interests lie in computer-assisted language learning, acquisition of Chinese pronunciation, and teaching Chinese as a foreign language in general.

Jiang Liu is an Assistant Professor of Chinese and Linguistics at the University of South Carolina, United States. His primary research areas are psycholinguistics, second language acquisition, phonetics and phonology. Before joining in the Department of Languages, Literatures and Cultures at the University of South Carolina, he served as the Associate Director of Chinese Flagship Program at the University of Minnesota for three years.

Min Liu is a Lecturer at the College of Chinese Language and Culture at Jinan University in Guangzhou, China. Her main research areas are the acquisition of second language speech, bilingual and bidialectal speech processing, and Chinese prosody, in particular, tone and intonation.

Yu Liu is an Assistant Professor of Chinese at Brigham Young University in the United States. Her research interests are Chinese linguistics, second language acquisition, and language assessment.

Nan Meng is an Associate Professor in Residence of Chinese and applied linguistics at the University of Connecticut in the United States. Her main research areas are the acquisition of second language, Chinese pedagogy, language socialization, and teacher education.

Eric Pelzl is a Postdoctoral Research Fellow at Penn State University. His main research interests include the perception and lexical processing of novel speech sounds by first and second language listeners.

Zhen Qin is an Assistant Professor at the Division of Humanities, The Hong Kong University of Science and Technology. Adopting phonetic and psycholinguistic approaches to second language learning, his research focuses on how Chinese tones are perceived and processed by native speakers and adult second-language learners.

Kevin Russell is an associate professor at the University of Manitoba. His research interests are dispersed into various areas: laboratory phonology, the phonology-phonetic interface, morphology, computational linguistics. His research encompasses various typologically distinct languages, such as American Sign Language and indigenous languages of the Americas including Guaraní and Cree.

Chenqing Song is an Associate Professor of Chinese and linguistics at SUNY-Binghamton University in the United States. Her main research areas are Chinese phonology, historical linguistics, poetic prosody, and Chinese language teaching.

Cheng Xiao is a Ph.D. student in the Linguistics Program at the University of South Carolina, United States. Her research interests include second language acquisition, speech perception and production, and psycholinguistics.

Yadong Xu is a Ph.D. student at the University of Manitoba. Her primary research interest lies in the morphology-syntax interface of Algonquian languages, a family of polysynthetic languages that is native to the land of North America. She has gained an interest in psycholinguistics, bilingual word recognition, and tone perception since working on her 2nd General's Paper.

Yi Xu is a Professor of speech sciences at University College London, UK. His research is primarily concerned with the basic mechanisms of speech production and perception in connected discourse, especially in terms of how multiple layers of communicative meanings can be encoded through a common process of articulation.

Bei Yang is an Associate Professor of (Chinese) linguistics at Sun Yat-sen University in China. Her main research areas are L1/L2 acquisition, Chinese dialectology, psycholinguistics, instrumental phonetics, and phonology.

Chunsheng Yang is an Associate Professor of Chinese and applied linguistics at the University of Connecticut in the United States. His main research areas are the acquisition of second language prosody, Chinese linguistics, sociophonetics, computer-assisted language learning, and applied linguistics in general.

Jing Yang is a Ph.D. candidate of (Chinese) linguistics at Sun Yat-sen University in China. Her main research areas are L2 phonetics & phonology, and the syntax of Chinese dialects.

Segments and Tones

Production of the Mandarin Rhotic Onset /ɹ/ by Indonesian Learners of Mandarin



Min Liu and Rongru Chen

Abstract The Mandarin rhotic /ɹ/ is difficult to acquire for many second language learners, including Indonesian learners of Mandarin. However, few studies have investigated the error patterns of the Mandarin rhotic /ɹ/ produced by Indonesian learners using objective acoustic methods. The present study empirically examined the differences in acoustic realization of the Mandarin rhotic onset /ɹ/ by L1 Mandarin native speakers and L2 Indonesian learners, so as to reveal the error types made by Indonesian learners. Through a production experiment, we found that Indonesian learners with intermediate Mandarin level had an overall error rate of about 20% in producing the Mandarin rhotic onset /ɹ/. The error types included taps, plosives, trills, fricatives, affricates, zero onset and laterals. And the speech errors were more likely to occur when the rhotic onset /ɹ/ was followed by a vowel /u/ than by other vowels. The correct acoustic realization of the rhotic onset /ɹ/ by the Indonesian learners resembles that of the native productions, but with a higher degree of rhoticity, as indicated by the stronger lowering effect of the rhotic onset /ɹ/ on the F3 of the following vowels /a, ə, u/ in the correct L2 productions relative to the standard L1 productions.

Keywords Indonesian learners · Mandarin native speakers · Rhotic onset /ɹ/ · Speech error types · Acoustic features

1 Introduction

Rhotic r-sounds are widely existent among world's languages. According to Maddieson (1984), about 75% of the world's languages have at least one r-sound in their consonant inventory. Phonologically, the various r-sounds within a language are often grouped into one class—rhotics (Widdison 1997). Phonetically, however, rhotics exhibit great variance both within and across languages (Ladefoged and Maddieson 1996; Lindau 1985). As a result, the rhotic r-sounds have been reported to be difficult

M. Liu (✉) · R. Chen
College of Chinese Language and Culture, Jinan University, Guangzhou, China
e-mail: nwliumin@gmail.com

to acquire for both L1 and L2 learners across languages (see, e.g., Boyce et al. 2016; McLeod 2007).

Mandarin has a rhotic /ɹ/ sound. Studies have shown that the Mandarin rhotic /ɹ/ is difficult to acquire for many second language learners of Mandarin (Liao and Shi 1985), including Indonesian learners. Based on impressionistic observations, research showed that there is a relatively high error rate of rhotic onset /ɹ/ productions by Indonesian learners of Mandarin (e.g., Deng 2011; Li 2013). However, few studies have investigated the erroneous patterns of the rhotic onset /ɹ/ productions from Indonesian learners of Mandarin in detail with objective acoustic methods. A dramatic increase of Indonesian learners of Mandarin over these years clearly highlights the need for such a study.

Mandarin rhotic onset /ɹ/ can be described in terms of the place of articulation and the manner of articulation. The place of articulation of the Mandarin rhotic onset /ɹ/ is rather consistent. The majority of researchers held that the place of articulation of the Mandarin rhotic onset /ɹ/ consists with that of the Mandarin consonants /tʂ, tʂʰ, ʂ/, being postalveolar with the tongue tip touching against the back of the alveolar bridge (Fu 1956; Lin 2007). Other researchers considered the Mandarin rhotic onset /ɹ/ as retroflex (Chao 1968; Duanmu 2007) or alveopalatal (Bao and Zheng 2011). Different from the place of articulation, the manner of articulation of the Mandarin rhotic onset /ɹ/ varies greatly among studies. It has been reported that the Mandarin rhotic onset /ɹ/ could be realized as voiced fricatives (e.g., Duanmu 2007; Gao 1940; Wu and Lin 1989), approximants (e.g., Chao 1968; Wang 1983), semivowels (e.g., Fu 1956; Zhu 1982) and several free variations. Liao and Shi (1987) categorized the realization of the Mandarin rhotic onset /ɹ/ into voiced fricatives and approximants. Miao et al. (2007) concluded that the manner of articulation of the Mandarin rhotic onset /ɹ/ includes voiceless/voiced fricatives and voiced approximants. Ran and Shi (2008) discovered several forms of realization of the Mandarin rhotic onset /ɹ/ in their electropalatographic data, including semivowels, vowels and voiced fricatives. Based on our observation on the Mandarin rhotic onset /ɹ/ productions by Mandarin native speakers, we agree with the view that the Mandarin rhotic onset /ɹ/ can be realized in the form of several free variations. From different speakers and under different syllabic contexts, we obtained five types of realization of the Mandarin rhotic onset /ɹ/: approximants, voiceless fricatives, voiced fricatives, vowels and semivowels. As that of many other languages, the Mandarin rhotic onset /ɹ/ has been acoustically characterized by a very low third formant frequency, which is generally lower than adjacent vowels (Delattre and Freeman 1968; Liao and Shi 1987).

In the field of second language acquisition, it has been widely accepted that the acquisition of an L2 phonological category can be affected by similar L1 segmental constellations that are close to L2 in L1 phonological space. Current theoretical models such as the Perceptual Assimilation Model-L2 (PAM-L2) and the Speech Learning Model (SLM) all somehow attribute the difficulties L2 learners confronted with in L2 acquisition to discrepancies between the L1 and L2 phonological categories (Best 1995; Best and Tyler 2007; Flege 1995). To investigate the Mandarin rhotic onset /ɹ/ productions by Indonesian learners of Mandarin, it is therefore needed

to first examine the rhotic r-sounds in Indonesian. Indonesian also has one rhotic r-sound, which is represented by the phoneme /r/. Generally, the Indonesian /r/ has been described as an apical post-dental trill [r] (Andi-Pallawa and Fiptar Abdi Alam 2013; Moeliono and Grimes 1995). It would sometimes be realized as a tap [ɾ], such as in intervocalic position (Soderberg and Olson 2008). The trill [r] and tap [ɾ] can be acoustically indicated by the number of contacts of the articulators involved. In the spectrogram, two or more contacts would be noticeable for the trill, whereas only one contact would be noticeable for the tap (Ball and Muller 2005; Hualde 2005).

A few studies have investigated the production of the Mandarin rhotic onset /ɹ/ by Indonesian learners of Mandarin. It was reported that Indonesian learners easily misproduced the Mandarin rhotic onset /ɹ/ as trill (e.g., Deng 2011; Li 2013; Zhang 2016). And the speech error type was consistently found in Indonesian learners at all Mandarin levels, from the beginner-level to the advanced-level learners. This has been taken as evidence of negative transfer from L1 to L2 in the L2 rhotic onset /ɹ/ production. In addition to the trill, Zhang (2016) also revealed the lateral as an error type of the Mandarin rhotic onset /ɹ/ by Indonesian learners of Mandarin. Note that these studies were not specifically designed to look into the production of the Mandarin rhotic onset /ɹ/ by Indonesian learners of Mandarin. Each study contained very limited number of r-syllables. And the detection of the speech errors was mainly based on subjective impressionistic observations of the researchers rather than objective criteria. Therefore, the existing studies may not be sufficient to show the full picture of the Mandarin rhotic onset /ɹ/ production by Indonesian learners of Mandarin. In the present study, we exhaustively investigated the production of all the Mandarin syllables containing a rhotic onset /ɹ/ by the Indonesian learners of Mandarin with more objective acoustic analyses. We examined the differences in acoustic realization of the Mandarin rhotic onset /ɹ/ between L1 (native speakers of Mandarin) and L2 speakers (Indonesian learners of Mandarin), to reveal the types of error made by L2 Indonesian learners in producing the Mandarin rhotic onset /ɹ/. By doing so, we aimed to reveal the source of speech errors and make pedagogical implications for teaching Mandarin rhotic onset /ɹ/ to Indonesian learners.

2 Method

2.1 Participants

Nineteen L2 Indonesian learners of Mandarin (13 females and 6 males) with an intermediate Mandarin level and six L1 Mandarin native speakers (3 females and 3 males) were recruited and paid to participate in the experiment. They were all college students from Jinan University in Guangzhou, China. The age of the L2 learners ranged from 19 to 25 years old ($M \pm SD$: 21.5 ± 1.93), and that of the native speakers ranged from 21 to 28 years old ($M \pm SD$: 24.5 ± 2.93). The Mandarin level of the L2 learners was assessed from their performance in the HSK test (Hanyu

Shuiping Kaoshi). Those who passed the HSK level 3 or level 4 test are generally considered to achieve an intermediate Mandarin level. We tested the intermediate-level L2 learners rather than the beginner-level L2 learners, because on the one hand, the language skills of the beginner-level L2 learners change rapidly and their production errors can be difficult to characterize; on the other hand, the intermediate-level L2 learners might show more persistent errors in L2 acquisition. For these speakers, Indonesian (L1), not Mandarin, was their primary language of daily communication at the time of testing. The native Mandarin speakers (L1 speakers) we recruited were all from Northern China. They all achieved the 1B level in the Putonghua Shuiping Ceshi (PSC), indicating that they have native proficiency in Mandarin without regional accents. Neither the L1 nor L2 participants had reported any speech or hearing disorders. Informed consent was obtained from all the participants before the experiment.

2.2 *Stimuli*

To have a complete and comprehensive investigation, we included all the 34 monosyllables containing the rhotic onset /ɹ/ in Mandarin for production, which covers all the possible combinations of rhymes and tones with the rhotic onset /ɹ/. The full list of monosyllables grouped by the following vowel (i.e., medial or nucleus of the rhyme of a syllable) is presented in Table 1. In addition to these experimental stimuli, we included one-third (17) monosyllabic stimuli with other consonant onsets as fillers. In total, there were 51 monosyllabic stimuli in the experiment.

2.3 *Recording*

The recordings took place in a soundproof recording booth at the Phonetics Lab of Jinan University. Participants wore a head-mounted microphone, and they were presented with the stimuli on the computer screen in a random order. Note that the stimuli were presented in the form of Pinyin (i.e., the official romanization system for Mandarin) to the L2 learners (Indonesian learners of Mandarin), in case they have not acquired the corresponding characters of some monosyllables. Given the reading habit of the L1 speakers (Mandarin native speakers), characters, instead of Pinyin, were presented to them. Participants were asked to produce each stimulus once without repetition. The recordings were made using the Adobe Audition CC software (Adobe, San Jose, CA; Breen and Breen 2015), with a 44.1 kHz sampling rate and 16-bit resolution. Participants completed a practice session before the test session, to familiarize themselves with the procedure. Instructions were given to participants orally by the experimenter before the experiment.

Table 1 Full list of monosyllables grouped by the following vowel for the production experiment

Vowel	Syllable	Vowel	Syllable
[ɿ]	rì	[u]	rú
[a]	rán		rǔ
	rǎn		rù
	ráo		ruá
	rǎo		ruán
	rào		ruǎn
	rāng		ruí
	ráng		ruǐ
	rǎng		ruì
	ràng		ruó
[ə]	rě		ruò
	rè		rún
	rén		rùn
	rěn		róng
	rèn		rǒng
	rēng		
	réng		
	róu		
	ròu		

2.4 Data Analysis

We collected 587 valid monosyllabic stimuli containing the rhotic onset /ʅ/ for L2 learners and 175 valid stimuli for L1 speakers. These stimuli were annotated. Based on the annotation information, we identified the error types that these L2 learners had when producing the rhotic onset /ʅ/ in Mandarin. The corresponding acoustic measures were further extracted and statistically analyzed for certain types of productions.

2.4.1 Annotation

The annotation of the recorded stimuli was conducted in Praat (Boersma and Weenink 2020). All stimuli were first forced aligned with two tiers “syllable” and “phones” in the TextGrid file in the Montreal Forced Aligner software (McAuliffe et al. 2017). Their accuracy was then manually checked by the second author in Praat (Boersma and Weenink 2020). A third tier “attribute” was added to annotate information about the manner of articulation of the rhotic onset /ʅ/ productions, and this information was used to identify the erroneous productions later on. We relied on both the perception of the segment and the acoustic features to make judgments on the attribute of the /ʅ/

productions. As the rhotic onset /ɹ/ can be realized in the form of several types of free variations, we identified the correct productions with the criteria that the productions did not show different manners of articulation compared to the standard productions of the native Mandarin speakers, and the productions were still perceived as a Mandarin rhotic /ɹ/ by native Mandarin speakers. Otherwise, the productions would be considered as erroneous productions. For example, a r-sound with an aperiodic high-frequency noisy spectrogram and a hearing sense of frication different from the native /ɹ/ productions would be recognized as a fricative and annotated with “F”. We specifically annotated the contact information of the rhotic onset /ɹ/ productions, that is, the number of contacts of the tongue tip against the back of the alveolar ridge. According to Hualde (2005), and also Ball and Muller (2005), the contact information can indicate whether the r-sound is realized as a tap or trill, which we assumed that the L2 learners in this study were very likely to do. A tap is usually characterized with one contact, whereas a trill is characterized with two or more contacts in the spectrogram. We applied this rule into the annotation of the contact number. Figure 1 presents an example of such annotation. One other point worth mentioning is that in the annotation of the “attribute”, we came across productions that were slightly displaced or unnatural, such as a relatively backer /ɹ/. These productions were not specifically classified and were treated as correct productions in our study, in contrast to the erroneous productions concerning changes in the manner of articulation.

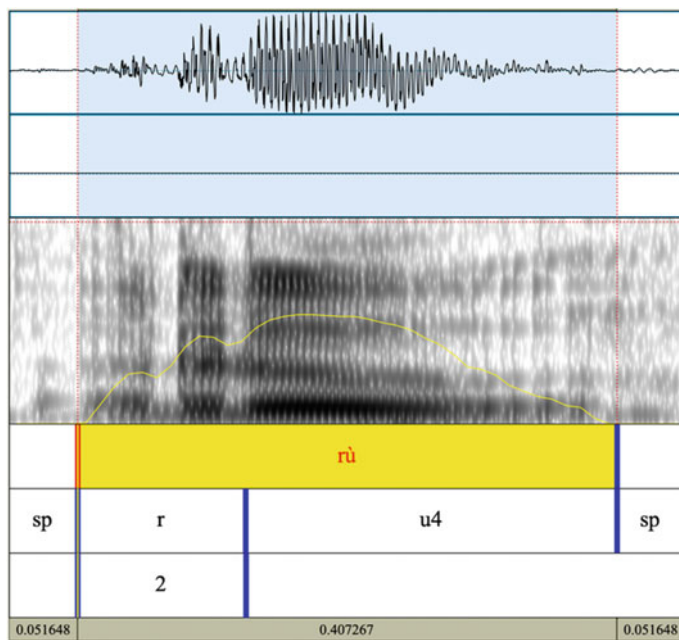


Fig. 1 An annotation example in Praat

2.4.2 Acoustic Measurements

One of the main goals of this study was to detect the distinctive acoustic features of the rhotic onset /ɹ/ produced by L2 Indonesian learners and L1 Mandarin native speakers. To do this, we first identified all the error types that L2 learners held for the Mandarin rhotic onset /ɹ/. As mentioned earlier, the error types can be obtained from the annotation information in the “attribute” tier. We noticed that the error types of the rhotic onset /ɹ/ mainly fell in the categories of taps, trills, plosives, affricates and fricatives. After identifying all the error types, we measured different acoustic parameters for different types of erroneous productions and for the correct /ɹ/ productions from L2 Indonesian learners in Praat (Boersma and Weenink 2020). Detailed acoustic parameters measured can be found in Table 2. We also extracted the same parameter (formant frequency here) for the L1 Mandarin native productions as for the correct L2 productions, so that comparisons could be made between the correct /ɹ/ productions by L2 learners and the standard /ɹ/ productions by L1 speakers.

It should be noted that we measured the formant frequency of the rhotic onset /ɹ/ productions in a particular way. As we obtained various types of free variations for the rhotic onset /ɹ/ such as voiceless fricatives, approximants, semivowels and vowels, we had to classify the productions based on the manners of articulation to get the accurate formant frequencies. However, the data points for each type of allophones were far from balanced. We therefore decided to take alternatives and treated all the rhotic onset /ɹ/ productions as a whole. Previous studies have shown that the rhotic consonants exhibit strong coarticulatory effects on the adjacent vowels (Recasens and Pallarès 1999). Instead of measuring the formant frequency of the rhotic consonant directly, we measured the F1, F2 and F3 values of the following vowels at two time points, one near the consonant offset (10 ms after the consonant offset/vowel onset) and one at the vowel midpoint. The formant frequency at the near-consonant point should be affected by that of the rhotic consonant, whereas the formant frequency at the vowel midpoint should remain unaffected. We calculated the difference of the formant frequency between the two points ($\Delta F = F_{\text{vowel midpoint}} - F_{\text{near-consonant point}}$). The difference of the rhotic productions between L1 and L2 speakers could then be seen through the different effects they exerted on the following vowels’ formants.

Table 2 Acoustic parameters measured for different production types

Production type	Acoustic parameter
Trill/tap	Contact number; formant frequency
Plosive/affricate	Voice onset time (VOT)
Fricative	Spectral center of gravity
Correct production	Formant frequency

2.4.3 Statistical Analyses

We analyzed the error types the L2 learners had when producing the rhotic onset /ɹ/ in Mandarin. We will report the distribution of different types of erroneous productions as well as that of the correct productions by the Indonesian learners of Mandarin with intermediate Mandarin level. The individual difference among speakers, as well as the vowel condition difference regarding the distribution of the error types, will also be reported.

Next, we conducted statistical analyses for the acoustic parameters we measured for the correct productions and the erroneous productions by L2 speakers, in comparison with the acoustic parameters measured for L1 speakers. However, regrettably, not enough data points were available for us to conduct valid statistical analyses for each type of erroneous productions. We had to restrict our analyses to the correct productions and the native productions. We ran several independent samples *t*-tests for the formant frequency difference (ΔF) of the first three formants between the correct productions by L2 learners and the standard productions by L1 speakers, to examine the acoustic differences between L2 and L1 productions. The degrees of freedom would be adjusted using the Welch–Satterthwaite method when the homogeneity of variance was violated.

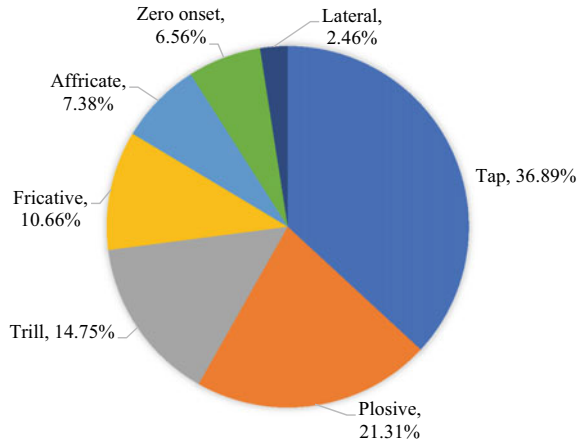
3 Results

3.1 Error Types

Of the 587 monosyllables containing the rhotic onset /ɹ/ productions collected from the L2 learners, we found 465 correct productions and 122 erroneous productions. The erroneous productions account for 20.78% of the whole stimuli, consistent with the error rate of the Mandarin rhotic onset /ɹ/ perception by L2 Indonesian learners reported in Wang (2008). The erroneous productions were further divided into seven types, according to the manners of articulation. We calculated the number of stimuli for each error type and the corresponding error rate (i.e., the percentage of a type of erroneous productions over all the erroneous productions). Arranged the error rates in descending order, the hierarchy went like this: taps (45 stimuli, 36.89%) > plosives (26 stimuli, 21.31%) > trills (18 stimuli, 14.75%) > fricatives (13 stimuli, 10.66%) > affricates (9 stimuli, 7.38%) > zero onset (8 stimuli, 6.56%) > laterals (3 stimuli, 2.45%). Clearly, the intermediate-level Indonesian learners of Mandarin easily mispronounced the Mandarin rhotic onset /ɹ/ as taps, plosives and trills, among others. Figure 2 shows the frequency distribution (in percentage) of different types of erroneous productions by these L2 speakers.

One might wonder whether each L2 speaker showed similar error types. In fact, the seven types of erroneous productions described above did not distribute evenly among speakers. Some error types were rather common across speakers, while others

Fig. 2 Frequency distribution of different types of erroneous productions by L2 speakers



were more speaker-specific. Figure 3 shows the frequency distribution of the correct productions as well as the different types of erroneous productions by each L2 speaker.

Out of the 19 L2 speakers, 6 speakers did not make any erroneous productions. This was judged by the criteria that the productions did not show different manners of articulation compared to the standard productions of the Mandarin native speakers, and the productions were still perceived as a Mandarin rhotic onset /ʅ/ by Mandarin native speakers. However, some of these /ʅ/ productions seemed to be over-stressed and sounded unnatural. This was a common problem for almost all speakers among their correct productions.

Other than that, it could be observed that the tap was the most widely distributed error type (in 10 out of 19 speakers), followed by the fricative and affricate (in 5 out of 19 speakers). The trill and plosive types were similarly populated (in 4 out of 19 speakers). Next was the error type of zero onset (in 3 out of 19 speakers), and

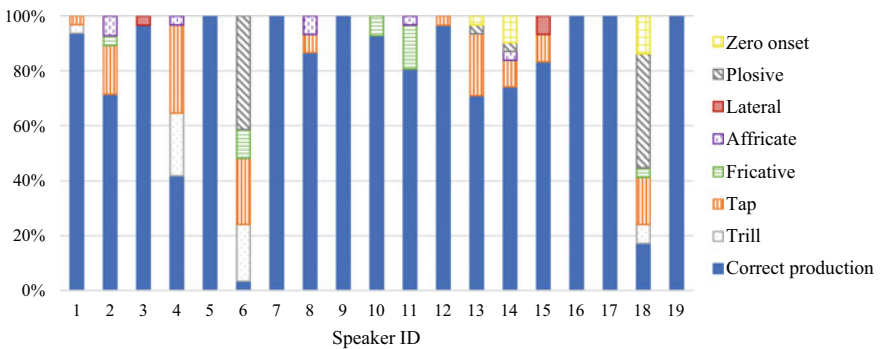


Fig. 3 Frequency distribution of different types of productions by each L2 speaker

Table 3 Frequency distribution of each type of errors occurred in different vowel conditions

Error type	Following vowel			
	/a/ (%)	/ə/ (%)	/ɨ/ (%)	/u/ (%)
Lateral	66.67	33.33	0.00	0.00
Tap	26.67	13.33	4.44	55.56
Trill	33.33	27.78	5.56	33.33
Fricative	23.08	23.08	7.69	46.15
Plosive	23.08	23.08	3.85	50.00
Affricate	0.00	55.56	0.00	44.44
Zero onset	0.00	37.50	0.00	62.50

the least widely distributed error type was the lateral, which only occurred in two speakers' productions.

We also explored if the error types of the rhotic onset /ɹ/ productions by Indonesian learners of Mandarin correlate with the following vowel types (see Table 3 for a summary). It appeared that the rhotic onset /ɹ/ was most likely to be mispronounced in the /u/ vowel condition than in other vowel conditions, and the /ɨ/ condition showed the fewest errors. The error types, taps, fricatives, plosives as well as zero onset occurred most frequently in the /u/ vowel condition. The error-type lateral co-occurred with a following vowel /a/ more often. Laterals did not appear in the /ɨ/ and /u/ conditions, and affricates and zero-onset productions were only found in the /ə/ and /u/ conditions. Trills, plosives and fricatives occurred in all the four vowel conditions.

It should be kept in mind that the data points for each vowel condition were quite different in our study, with the most stimuli for the /u/ condition and the fewest stimuli for the /ɨ/ condition. However, we exhaustively explored all the possible monosyllables containing the rhotic onset /ɹ/ in Mandarin. Biased as it might be due to the unbalanced data points for each vowel condition, what we reported here should represent the typical distribution of the erroneous productions of the rhotic onset /ɹ/ produced by the Indonesian learners of Mandarin in relation to the following vowels.

3.2 Acoustic Results

Since there were not enough data points for each type of erroneous productions, we could not perform reliable statistical analyses for these data. In the following, we will only report the statistical results for the correct productions by L2 speakers and for the standard productions by L1 speakers. About 80% productions of the Mandarin rhotic onset /ɹ/ by the L2 speakers were correct productions. However, as L2 productions, they did not sound completely natural. We thus performed acoustic

analyses to compare the acoustic realization of the correct productions by L2 speakers and that of the standard productions by L1 speakers.

We measured the F1, F2 and F3 values of the vowels following the rhotic onset /ʅ/ at two time points, one near the consonant offset (10 ms after the consonant offset/vowel onset) and one at the vowel midpoint, for the correct productions by L2 speakers and the standard productions by L1 speakers. We further calculated the formant frequency difference between the two points ($\Delta F = F_{\text{vowel midpoint}} - F_{\text{near-consonant point}}$). Independent samples *t*-tests were performed for ΔF in F1, F2 and F3 between L2 correct productions and L1 standard productions. For syllables with a diphthong, there can be noticeable vowel change within the syllable, which could easily distort the formant value of vowels. We thus discarded these syllables in the analysis. As a result, only syllables with a monophthong or with an additional nasal coda were included in the final analysis (see Table 4 for the list). In the end, we had 289 valid stimuli for the L2 speakers and 94 stimuli for the L1 speakers for the formant analyses. To eliminate influences from the physiological factors, the formant data were normalized with the Nearey method (Nearey 1977) using the formula:

$$F_{n[V]}^* = \text{anti-log}(\log(F_{n[V]}) - \text{MEAN}_{\log})$$

where $F_{n[V]}^*$ is the normalized value for $F_{n[V]}$, formant *n* of vowel *V*, and MEAN_{\log} is the log-mean of all F1s, F2s and F3s for the speaker under investigation. As vowels intrinsically differ in their formant frequencies, we will present the results by the vowel type in the following (see Fig. 4).

As can be seen from Fig. 4, ΔF showed similar patterns in the vowels /a/, /ə/ and /u/. Both L1 and L2 speakers had a positive ΔF in F1 and F3, but a negative ΔF in F2 in all three vowel conditions. This indicates that the F1 and F3 of vowels /a, ə, u/ increased from the near-consonant point to the vowel midpoint, whereas the F2 decreased from the near-consonant point to the vowel midpoint.

Table 4 Stimuli list for the formant analysis

Vowel	Syllable	Vowel	Syllable
[ɿ]	rì	[u]	rú
			rǔ
			rù
			róng
			rǒng
[ə]	rě	[a]	rán
	rè		rǎn
	rén		rāng
	rěn		ráng
	rèn		rǎng
	rēng		ràng
	réng		

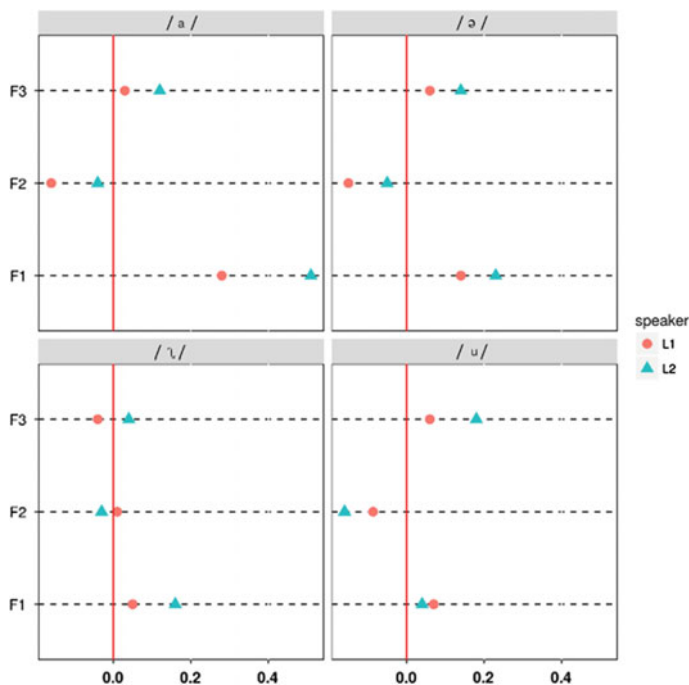


Fig. 4 ΔF for the first three formants of different vowels following the rhotic onset /ɹ/ produced by L1 and L2 speakers. Note that all vowel formants and hence ΔF were normalized with the Nearey (1977) method

Statistical analyses showed that the F3 of the L2 speakers showed a more significant rising from the near-consonant point to the vowel midpoint than the F3 of the L1 speakers across the three vowels (all $ps < 0.05$). Viewed from the perspective of the influence of the rhotic onset /ɹ/ on the near-consonant part, we can see that the rhotic onset /ɹ/ produced by the L2 speakers had a more significant lowering effect on the F3 of the vowels /a, ə, u/ than those produced by the L1 speakers did. For the vowels /a/ and /ə/, compared to L1 speakers, the rhotic onset /ɹ/ produced by the L2 speakers had a weaker rising effect on the F2 formants (/a/: $t(123.44) = -5.08, p < 0.05$; /ə/: $t(140) = -2.88, p = 0.005$), but a stronger lowering effect on the F1 formants (/a/: $t(121.68) = -3.19, p = 0.002$; /ə/: $t(131.83) = -2.60, p = 0.01$). For the vowel /u/, however, no significant difference was found for the ΔF in F1 ($t(92.12) = -0.88, p = 0.38$) and F2 ($t(93) = -1.15, p = 0.25$) between L1 and L2 speakers.

A pattern distinctive from that of the /a, ə, u/ was found in the /ɪ/ condition. In general, the ΔF centered around the zero line, and the values were close to zero. Statistical analyses showed no significant difference for ΔF in F2 between L1 and L2 speakers ($t(18) = 0.49, p = 0.63$). Nor was there a significant difference for ΔF in F3 between L1 and L2 speakers ($t(18) = -1.75, p = 0.098$). Exceptional was the

ΔF in F1 of the /ɹ/ produced by L2 speakers. The rhotic onset /ɹ/ produced by the L2 speakers showed a significant lowering effect on the F1 of the following vowel /ɹ/ ($t(13.59) = -2.48, p = 0.03$), compared to the L1 speakers.

4 General Discussion

The present study empirically examined the differences in acoustic realization of the Mandarin rhotic onset /ɹ/ between L1 (native speakers of Mandarin) and L2 speakers (Indonesian learners of Mandarin), and revealed the types of error made by L2 Indonesian learners with intermediate Mandarin level in producing Mandarin rhotic onset /ɹ/. It was found that the Indonesian learners of Mandarin we recruited had an overall error rate of about 20% in producing the Mandarin rhotic onset /ɹ/ sounds. The erroneous productions can be further divided into seven types according to the manners of articulation: taps (36.89%), plosives (21.31%), trills (14.75%), fricatives (10.66%), affricates (7.38%), zero onset (6.56%) and laterals (2.46%). The numbers in the brackets indicate the proportion of each type of errors over all erroneous productions. It can be seen that taps, plosives and trills accounted for the majority of the speech errors made by Indonesian learners of Mandarin in producing the Mandarin rhotic onset /ɹ/. And the speech errors were more likely to occur when the rhotic onset /ɹ/ was followed by a vowel /u/ than by other vowels, with the error type most likely to be a tap.

The previous investigations on the error types of the Mandarin rhotic onset /ɹ/ produced by Indonesian learners were primarily based on impressionistic observations (e.g., Deng 2011; Li 2013; Zhang 2016). With very limited numbers of r-syllables covered, these studies all reported trill as the main error type for the Mandarin rhotic onset /ɹ/ productions by the Indonesian learners of Mandarin, irrespective of the Mandarin level of the Indonesian learners. The lateral was recognized as an error type merely in Zhang (2016). In our study, we exhaustively explored the production of all the monosyllables containing a rhotic onset /ɹ/ in Mandarin by the Indonesian learners with an intermediate Mandarin level. Based on the acoustic cues as well as the auditory perception, we found seven error types that Indonesian learners made when producing the rhotic onset /ɹ/ in Mandarin. Our study discovered new error types for Mandarin onset /ɹ/ produced by L2 Indonesian learners which has not been reported in previous studies. Among the top four error types, three of them (taps, plosives and affricates) were newly discovered in our study. The tap, rather than the trill, was recognized as the most common error type.

As mentioned earlier, Mandarin rhotic onset /ɹ/ is subject to much variation and could still be perceived as a /ɹ/ (Miao et al. 2007; Ran and Shi 2008). However, the tap and trill are not included in these free variations. The Mandarin rhotic /ɹ/ class has no room for the tap and trill. In contrast, both of them exist in the Indonesian consonant inventory. The Indonesian rhotic /ɹ/ has been typically described as a trill, with the tap being a conditioned variant (Moeliono and Grimes 1995; Soderberg and Olson 2008). It appears that Indonesian learners of Mandarin assimilated the L2

phonological category (Mandarin rhotic onset /ɹ/) to their L1 phonological category (Indonesian rhotic /r/: tap/trill), causing speech errors in the production of the L2 rhotic onset /ɹ/. This result supports the PAM-L2 model and the SLM model (Best 1995; Best and Tyler 2007; Flege 1995). As for the error-type plosive, it exists in the consonant inventory of both Mandarin and Indonesian. We found that the average voice onset time (VOT) of the erroneous plosive productions by Indonesian learners of Mandarin was much longer than that of the typical Mandarin plosives (cf. Zhou and Zheng 2008), but rather close to the VOT of the typical Indonesian plosives reported in the literature (Hardjono 2011). Indonesian learners tended to produce the plosives that exist in their native language as a substitute to the Mandarin rhotic /ɹ/ sometimes. It might as well be considered as a negative transfer from their native language.

We further performed acoustic analyses to compare the acoustic realization of the correct productions by Indonesian learners of Mandarin (L2 speakers) and that of the standard productions by Mandarin native speakers (L1 speakers). It was found that the Mandarin rhotic onset /ɹ/ produced by both L1 and L2 speakers had a lowering effect on the F3 of the following vowels /a, ə, u/, but not on the F3 of the following vowel /ɿ/. Also, the Mandarin rhotic onset /ɹ/ produced by both L1 and L2 speakers had a lowering effect on the F1 of the following vowels /a, ə, ɿ/ and a rising effect on the F2 of the following vowel /a, ə/. The lowering effects for F1 and F3 were stronger in the L2 productions than L1 productions, whereas the rising effects for F2 were weaker in the L2 productions than L1 productions. Generally, the Mandarin rhotic onset /ɹ/ did not affect the formants of the /ɿ/ vowel much, except the F1 of the vowel /ɿ/ by the L2 speakers. This result echoes the view that the place of articulation of the Mandarin /ɿ/ vowel is almost identical to that of the Mandarin rhotic onset /ɹ/ (Liao and Shi 1987). In all the other vowel conditions (/a, ə, u/), we consistently found a lowering effect of the rhotic onset /ɹ/ on the F3 of the vowels, regardless of the speaker groups. This confirmed the role of a low third formant frequency in characterizing the rhotic r-sounds (Delattre and Freeman 1968; Liao and Shi 1987). Previous studies have shown that F3 is positively correlated with the dorsopalatal contact degree and with the palatal constriction narrowing (Fant 1960; Recasens and Pallarès 1995). The stronger lowering effect of the rhotic onset /ɹ/ on the F3 of the vowels /a, ə, u/ in the L2 productions relative to the L1 productions seems to indicate that the rhotic onset /ɹ/ produced by the Indonesian learners of Mandarin had a greater level of constriction and a higher degree of rhoticity, which could account for the unnaturalness in their correct productions. Moreover, we found a lowering effect of the rhotic onset /ɹ/ on the F1 and a rising effect on the F2 of the vowels /a, ə/ across the speaker groups. These effects might result from the specific configuration of the formant frequencies of the rhotic onset /ɹ/ and the vowels involved. More research is needed to explore whether they are defining features of the Mandarin rhotic onset /ɹ/. Overall, the correct acoustic realization of the rhotic onset /ɹ/ by the Indonesian learners of Mandarin resembles that of the native productions, but with a higher degree of rhoticity.

Our study provides insights into the language pedagogy in teaching Mandarin rhotic onset /ɹ/ to Indonesian learners. Indonesian learners are likely to assimilate the

Mandarin rhotic onset /ʃ/ to the Indonesian rhotic /ʃ/ class (tap/trill) in the Mandarin production. They should be instructed to pay special attention to the different manners of articulation of the two rhotic systems. Also, Indonesian learners should learn to adjust the contact degree of their articulators to avoid the problem of over rhoticity in producing Mandarin rhotic onset /ʃ/.

5 Conclusion

To conclude, Indonesian learners of Mandarin with intermediate Mandarin level had an overall error rate of about 20% in producing the Mandarin rhotic onset /ʃ/ sounds. The error types included taps, plosives, trills, fricatives, affricates, zero onset and laterals. And the speech errors were more likely to occur when the rhotic onset /ʃ/ was followed by a vowel /u/ than by other vowels. Though the acoustic realization of the correct rhotic onset /ʃ/ produced by the Indonesian learners of Mandarin resembles that of the Mandarin native productions, there was a stronger lowering effect of the rhotic onset /ʃ/ on the F3 of the following vowels /a, ə, u/ in the correct L2 productions relative to the standard L1 productions, which suggest a higher degree of rhoticity in the correct L2 productions.

Acknowledgements This research was funded by Grant-GD19YYY06 from the Guangdong Planning Office of Philosophy and Social Science to ML. We thank Dr. Matthew Faytak for his helpful comments and suggestions on the acoustic measurement methods.

References

- Andi-Pallawa, B., & Fiptar Abdi Alam, A. (2013). A comparative analysis between English and Indonesian phonological systems. *International Journal of English Language Education*, 1(3). <https://doi.org/10.5296/ijele.v1i3.3892>.
- Ball, M. J., & Muller, N. (2005). Articulation: Consonant manner types. *Phonetics for communication disorders* (pp. 63–73). New York: Psychology Press.
- Bao, H., & Zheng, Y. (2011). Putonghua Dongtai Ewei Yanjiu. [An electropalatography study on the articulation of Standard Chinese]. *Journal of School of Chinese Language and Culture Nanjing Normal University*, 03, 1–11. (in Chinese).
- Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). Baltimore, MD: York Press.
- Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Language experience in second language speech learning* (pp. 13–34). Amsterdam: John Benjamins Publishing.
- Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer [Computer program]. Version 6.1.16, retrieved 6 June 2020 from <http://www.praat.org/>.
- Boyce, S. E., Hamilton, S. M., & Rivera-Campos, A. (2016). Acquiring rhoticity across languages: An ultrasound study of differentiating tongue movements. *Clinical Linguistics & Phonetics*, 30(3–5), 174–201. <https://doi.org/10.3109/02699206.2015.1127999>.

- Breen, C., & Breen, C. (2015). Review: Adobe audition CC a solid upgrade hampered by subscription pricing. Retrieved 13 March 2015, from <http://www.macworld.com/article/2043340/adobe-audition-cc-solid-upgrade-hampered-by-subscription-pricing.html>.
- Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley: University of California Press.
- Delattre, P., & Freeman, D. C. (1968). A dialect study of American r's X-ray motion picture. *Linguistics: An Interdisciplinary Journal of the Language Sciences*, 6(44), 29–68.
- Deng, X. (2011). *Yinni Mianlan Huayi he Feihuayi Xuesheng Putonghua Yuyin Xide Pianwu Bijiao*. [The comparison of errors in Mandarin acquisition between Chinese students and non-Chinese students of Medan, Indonesia]. M.A thesis, Jinan University.
- Duanmu, S. (2007). *The phonology of Standard Chinese*. Oxford: OUP Oxford.
- Fant, G. (1960). *Acoustic theory of speech production*. The Hague, The Netherlands: Mouton.
- Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research*. Timonium, MD: York press.
- Fu, M. (1956). Beijinghua de Yinwei he Pinyin Zimu. [The phoneme and pinyin alphabets of Beijing Mandarin]. *Studies of the Chinese Language*, 05, 3–12. (in Chinese).
- Gao, B. (1940). *Zhongguo Yinyunxue Yanjiu*. [On the phonology of Chinese]. Beijing: Tsinghua University Press. (in Chinese).
- Hardjono, F. (2011). *Stop consonant characteristics: VOT and voicing in American-born-Indonesian children's stop consonants*. Ph.D. dissertation, The Ohio State University.
- Hualde, J. I. (2005). *The sounds of Spanish*. Cambridge: Cambridge University Press.
- Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford Cambridge, MA: Blackwell.
- Li, S. (2013). Yinni Liuxuesheng zai Hanyu Shengmu Yunmu shang de Yuyin Pianwu Diaocha Fenxi. [Analysis of errors of initials and finals made by Indonesian learners of Mandarin]. M.A thesis, Chongqing Normal University. (in Chinese).
- Liao, R. & Shi, F. (1985). Duiwai Hanyu Jiaoxue Zhong r Shengmu Yinzhi de Shiyang Yanjiu. [An experimental study on the sound quality of r initials in teaching Chinese as a foreign language]. In: *Proceeding of the 1st international conference on Chinese language pedagogy* (pp 236–256). (in Chinese).
- Liao, R., & Shi, F. (1987). Hanyu Putonghua r Shengmu Yinzhi de Shiyang Yanjiu. [An experimental study on the articulatory manner of the Mandarin initial r]. *Studies in Language and Linguistics*, 02, 146–160. (in Chinese).
- Lin, Y.-H. (2007). *The sounds of Chinese*. Cambridge: Cambridge University Press.
- Lindau, M. (1985). The story of /r/. In V. Fromkin (Ed.), *Phonetic linguistics* (pp. 157–168). Orlando, Florida: Academic Press.
- Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner [Computer program]. Version 0.9.0, retrieved 17 January 2017 from <http://montrealcorpus.github.io/Montreal-Forced-Aligner/>.
- McLeod, S. (2007). *The international guide to speech acquisition*. Clifton Park, NY, USA: Thomson Delmar Learning.
- Miao, G., Hao, Y. & Lv, S. (2007). Putonghua Fuyin [r] de Shengxue Tezheng Yanjiu. [A study on acoustic features of the Mandarin initial [r]]. In: *Proceeding of the 9th national conference on man-machine speech communication*, 296-299. (in Chinese).
- Moeliono, A. M., & Grimes, C. E. (1995). Indonesian (Malay). In D. T. Tryon (Ed.), *Comparative Austronesian dictionary: An introduction to Austronesian studies, Part 1, Fascicle 1 (Trends in Linguistics. Documentation 10)* (pp. 443–457). Berlin: Mouton de Gruyter.
- Nearey, T. M. (1977). *Phonetic feature systems for vowels*. Dissertation, University of Alberta.
- Ran, Q. & Shi, F. (2008). Beijinghua r Shengmu de Bianti ji Yinwei de Juhe Chengdu. [The variations of the initial r in Beijing Mandarin and the degree of convergence of phonemes]. In: *Proceedings of the Nanjing conference on Chinese phonology*, (pp 450–464). (in Chinese).

- Recasens, D., & Pallarès, M. D. (1995). Articulatory-acoustic correlations and coarticulatory resistance in consonants. *European Journal of Disorders and Communication*, 30, 203–212.
- Recasens, D., & Pallarès, M. D. (1999). A study of /r/ and /r/ in the light of the “DAC” coarticulation model. *Journal of Phonetics*, 27(2), 143–169. <https://doi.org/10.1006/jpho.1999.0092>.
- Soderberg, C. D., & Olson, K. S. (2008). Indonesian. *Journal of the International Phonetic Association*, 38(2), 209–213. <https://doi.org/10.1017/S0025100308003320>.
- Wang, G. (2008). Yinni Liuxuesheng Hanyu Shengmu Ganzhi Shiyan Yanjiu. [A perception experiment on Chinese initials by Indonesian learners of Chinese]. *Language Teaching and Linguistic Studies*, 05, 32–38. (in Chinese).
- Wang, L. (1983). Zailun Rimu de Yinzhì, Jianlun Putonghua Shengmu Biao. [Revisit the articulation manner of the initial r and Putonghua initial alphabets]. *Zhongguo Yuwen*, 03, 20–23. (in Chinese).
- Widdison, K. A. (1997). Variability in lingual vibrants: Changes in the story of /r/. *Language & Communication*, 17(3), 187–193. [https://doi.org/10.1016/S0271-5309\(97\)00011-6](https://doi.org/10.1016/S0271-5309(97)00011-6).
- Wu, Z. & Lin, M. (1989). *Shiyan Yuyinxue Gaiyao (Chuban)*. [Introduction on experimental phonetics (First edition)]. Beijing: China Social Sciences Press. (in Chinese).
- Zhang, J. (2016). *Yindunixiya Xuesheng Xide Hanyu Yuyin de Pianwu Fenxi-Yi Wanlong Guoji Waiyu Xueyuan Weili*. [An analysis on Indonesian students’ acquisition errors of Chinese Phonemes-Take bandung international foreign language college as an example]. M. A thesis, Jilin University. (in Chinese).
- Zhou, X. & Zheng, Y. (2008). Jiyu EPG de Putonghua Fuyin VOT Tongji Fenxi. [Statistical analyses on VOT of the Mandarin consonants based on EPG]. In: *Proceedings of the Nanjing conference on Chinese phonology*. (in Chinese).
- Zhu, X. (1982). Guanyu Putonghua Rimu de Yinzhì [On the articulation features of the r initial]. *Current Research in Chinese Linguistics*, 03, 19–21. (in Chinese).

What is in the Final Stage of Inter-Language? Tone Errors and Phonological Constraints in Spontaneous Speech in Very Advanced Learners of Mandarin



Chenqing Song

Abstract Following the research direction proposed in Zhang (Journal of Chinese Language Teachers Association 45(1):39–65, 2010, The second language acquisition of Mandarin Chinese tones by English, Japanese and Korean speakers. [Doctoral Dissertation], University of North Carolina-Chapel Hill, 2013, Second language acquisition of Mandarin Chinese tones—Beyond first language transfer, Brill, 2018), this study examines the tone errors and substitution patterns in spontaneous connected speech produced by L2 learners who have progressed further into advanced or superior levels of proficiency. Based on the distinct patterns of errors and substitution patterns in their speech samples, effects of phonological universals in the format of constraints, including tone markedness scales (TMS), tone-position constraints (TPC) and the obligatory contour principle (OCP), are studied. Comparisons are made between the findings about tonal acquisition made in previous studies on lower-level learners and the higher-level learners in this study. With these error data, analyses and comparisons, I argue that some effects of the above-mentioned universals are still visible (TMS and T4-T4 OCP) while others are masked. It is a special configuration of the coarticulation rule applying to T2-T4 and T2-T1 combinations that really distinguishes the tonal system found in these very advanced learners from that of the other learners and that of the native speakers of Mandarin. Pedagogical practices that are designed to re-configure this rule will allow learners at this stage of tone acquisition to proceed into native-like speech production.

Keywords Phonological constraints · Tone errors in production · Advanced L2 learners

C. Song (✉)
SUNY-Binghamton University, Binghamton, NY, USA
e-mail: csong@binghamton.edu

© Springer Nature Singapore Pte Ltd. 2021
C. Yang (ed.), *The Acquisition of Chinese as a Second Language Pronunciation*,
Prosody, Phonology and Phonetics, https://doi.org/10.1007/978-981-15-3809-4_2

1 Introduction

Mandarin tones have been studied from the aspect of phonology, phonetics, L1 and L2 acquisition. There are still unsolved problems, but generally researchers accept the following basic properties of the tones in Mandarin. There are four tone types for each full syllable and changing the tone will change the meaning of a morpheme, which is normally monosyllabic. The tones on single syllables may be transcribed and represented in different ways. Some of the most used ones are shown in Table 1.

The 1–5 scale was introduced in Chao (1930), in which 1 represents the lowest pitch point and 5 the highest. The HL system (following the tradition of Autosegmental Phonology, Goldsmith (1976)) uses H to represent a high tone, and L for a low tone and a contour tone is a combination of more than one-level tones. An M is sometimes used to indicate a mid-tone. The register-component tone representation is following the proposal in Bao (1999), in which register and contour of the tones are separate branches from the tone bearing unit (TBU), and the proposal made in Yip (1980), which uses [\pm upper] for the registers of tones. In this way of representation, *h* and *l* are terminal tone segments that are high and low, respectively. In this paper, tone category names and the HL system will be used to represent tones as they are already sufficient for the purpose of discussion here. Besides the four tones, Mandarin has a weak tone on reduced syllables, which is often called the “neutral tone.” I am going to label it as Tn for the ease of recognition among other category labels.

When two or more tones are together, one or more of them may change the pitch value, a phenomenon called “tone sandhi.” Mandarin’s most widely discussed tone sandhi rule is the third tone sandhi (T3 sandhi), which states that a T3 will be pronounced as a high rising (MH) when it is followed by another T3. The resulting high rising is believed to be identical to the citation form of T2 but some researchers argue that the two are still distinctive (Zee 1980; Kratochvil 1987; among others). In this paper, I will label the output of the T3 sandhi as T5. When preceded by tones of a different category, T3 is pronounced as a low, slightly falling tone (ML), which is going to be labeled as HT3. This process is known as the half T3 sandhi. There are a couple of other sandhi rules, including some that are lexically restricted and a T2 Sandhi (Chao 1968) which occurs to T2s in some trisyllabic expressions. The domain of sandhi rules is a topic of prolific research, but it is not going to be touched upon in this paper.

Table 1 Basic information of the four tones in Mandarin

Category label	1–5 scale	HL system	Register + component tones	Example
T1	55	H	[+ U, hh]	mā “mother”
T2	35	MH	[+ U, lh]	má “hemp”
T3	214/21	L	[–U, ll]	mǎ “horse”
T4	51	HL	[+ U, hl]	mà “to scold”

Studies on the learning of Mandarin tones among L2 learners have produced rich insights into various issues. By studying the tone production and errors, researchers tried to identify the sources of the errors. Many of them (White 1981; Shen 1989; Miracle 1989; Guo 1993; Chen 1997; Sun 1998; Q.H. Chen 2000) focus on L1 transfer effects or the impact of L1 on the mastering of Mandarin tones. The complexity of the Mandarin tonal system and particular tonal features are also found to cause errors (Shen 1989; Miracle 1989; Elliot 1991, Hao 2012; among others). Some studies (Leather 1990; Elliot 1991; Guo 1993; Wang et al. 1999; Wang 2006; Hao 2012; Yang 2015; among others) combine production and perception, arguing that the relationship between the two plays a role in the accurate/inaccurate production of tones. In terms of research design, there are experimental studies, in which subjects are recruited to perform production or perception tasks with target tones¹ and combinations being solicited. Only a few studies are longitudinal and checked learners' performance at different developmental stages (e.g., Guo and Tao 2008). Yang (2011) investigates tone errors in the greater context of Mandarin phonology and argues that the tone errors are the results of "superimposition of the L1 English utterance-level prosody over tone production by L2 learners." Among the above-mentioned research, most are done on English native speakers. Zhang (2010, 2013, 2018) diverges from the other studies on the topic by focusing on the universal phonological constraints under the framework of the Optimality Theory (OT) (Prince and Smolensky 1993; McCarthy and Prince 1993, 1995; among others). By comparing the experimental data from three different learner groups (English native speakers, Korean native speakers and Japanese native speakers), L1 transfer effects are isolated from the possible effects of universal constraints. There are issues remaining in the study of the topic. As Zhang (2013) points out, research that bridges general linguistic theories and language teaching is still scarce. The data used for research are often isolated words, phrases and sentences with little discourse contexts. In experimental settings, tasks are often reading and repeating, which cannot reflect the proficiency of learners, an ability that can only be properly assessed in natural speech. Q.H. Chen (2000) tries to avoid this problem by using natural connected speech but found that tones in connected speech are hard to judge for accuracy out of context because tones (even produced by native speakers) in connected speech may be drastically different from their citation forms. Moreover, most existing literature on the L2 tonal acquisition is done on learners of low or intermediate proficiency levels, due to the limited availability of advanced-level or superior-level speakers in the past years.

¹Target tone refers to the intended tone category that a speaker is trying to pronounce. In some studies, the target tones were provided to the speakers who read words/expressions from a list. In the current study, it is identified in the sample speech according to the lexical and contextual information. In connected spontaneous speech when a L2 learner's proficiency is low, and pronunciation errors as well as other errors are abundant, it is often difficult to identify the target tones. In the current study, the speech samples from the four speakers are very clear with almost no grammatical errors, lexical errors or segmental errors. Therefore, it is not difficult to know which word a speaker intended to say even if the tones were not pronounced accurately. Through the identification of words in the speech samples, the target tones were inferred.

The literature survey reveals gaps left unfilled by the previous studies and thus calls for a study in which spontaneous connected speech production of advanced or superior-level learners will be collected, examined and analyzed at the level of phonological acquisition. This is the approach and goal of this study. I will partially follow the framework proposed in Zhang (2010, 2013, 2018), especially focusing on three types of constraints: TMS, TPC and OCP. Tonal markedness scale (TMS) is a universal and phonetically grounded constraint, which ranks some tones as more marked than others. The proposed ranking or scale of the tones according to their markedness is *Rising >> *Falling >> *Level (Ohala 1978; Hyman and VanBik 2004). To L2 phonology, a more marked tone is often acquired later than a less marked one. Tone position constraints (TPC) are a set of constraints that states that certain tones are more marked or disfavored in certain positions. For example, Zhang (2004) proposes that “phrase-final syllables and syllables in shorter words are the preferred bearers of contour tones, even though they are usually not privileged for other phonological contrasts.” Obligatory contour principle (OCP) is a constraint that states certain consecutive identical segments/tones/features are banned or disfavored (Goldsmith 1976).

Through controlled experiments done with three L1 groups (English, Korean and Japanese native speakers) and comparisons among the data obtained from the three groups, Zhang was able to make specific claims about the L1 transfer effects. Since only English native speakers’ speech is used in this study, no cross-linguistics or L1 transfer claims will be made. Due to the length restriction of this paper, OT theory (and the corresponding OT L2 theory) will be brought into discussion at the conceptual level where rankings of relevant constraints are proposed but not at the technical level where all constraints are ranked in tableaux to output the attested forms.

2 Research Questions, Method and Subjects

Advanced-level (or superior-level) L2 learners of Mandarin Chinese are often not abundant, which explains why most previous studies focused on learners who are at lower levels of proficiency. In experiment-based studies, the sample size requirement compels researchers to choose subjects of study who are beginner or intermediate level learners. Not many studies have revealed the tonal productions among advanced or superior-level learners of Mandarin Chinese. As a consequence, many questions remain unanswered with regard to the tonal production of this group. To that end, the current study attempts to address some of these questions:

1. What kind of tone errors do advanced/superior learners make?
2. Do the errors they make fall into some common categories? If so, what are the categories of these errors?
3. Are these errors similar to or different from the errors identified from previous studies done mostly on learners of lower proficiency levels?

4. If so, can phonological universals (TMS, TPC and OCP) explain these errors? To what extent, do they explain these errors?
5. What do these errors reveal to us about the development of the inter-language (IL) at this very final stage of acquisition?

The significance of the answers to these questions extends beyond the understanding of Mandarin tonal acquisition. If the learners are at a very high level of proficiency, then their IL systems are very close to that of the native speakers. Then, the differences between the IL systems and the native speakers' system should provide clues about the Mandarin phonological system itself. As previously stated, the present study follows the theoretical standpoint in Zhang (2010, 2013, 2018), which views IL as an L2 phonological system, in which phonological constraints interact in different rankings to produce the observed outputs (see the OT framework mentioned above). Along this line of thinking, the crucial question to ask in addition to the five listed above is what differences exist among constraints and their rankings, which set apart the IL system and the native speakers' system. If the OT theoretical framework is adequate for L2 acquisition study, and our analysis is accurate, then it is predicted that there will only be a few constraints and minor differences in their rankings between the advanced learners and the native speakers. The results from the current study, to be presented in this paper, provide support to this hypothesis.

To answer these research questions, this paper uses spontaneous connected speech from four L2 speakers that are publicly available through YouTube channels. These speech samples were produced not for educational or research purposes but purely for informative and recreational ones. In other words, the speakers focused on the message over the forms. These speakers' pronunciation of the Mandarin tones is very close to that of the native speakers' with an error rate between 1.5% and 6.3%.² They are able to sustain a speech in Chinese for an extended period of time, as demonstrated by their YouTube videos, on a good variety of topics from education to politics, from daily life to economy. Their speech in Chinese is both fluent and accurate, facilitated by an extended vocabulary and a solid mastery of grammar. There is almost no grammatical error or misuse of lexical items in their speech. In a single video continuous shot with minimal post-shooting editing or revision, these speakers are able to elaborate on a topic for over ten minutes, weaving discourses that are highly consistent, coherent, culturally appropriate and functionally adequate. Using the *ACTFL Performance Descriptors for Language Learners* (ACTFL 2012) to evaluate these speakers, in the presentational mode of communication, two ACTFL OPI trained evaluators (including the author of the paper) independently assessed and confirmed that they fall into the advanced range (or higher) properly by demonstrating the following performance.

²The term "error rate" is defined as the ratio of total number of incorrect tones in all syllables in the sample for each speaker. The method of how the accuracy of tonal production was assessed will be explained in a later section of this paper. Initially, nine speakers were included in the pool. Five of them turned out to have much higher overall tone error rates and were assessed as lower in proficiency level using the ACTFL Performance Descriptor. So these five were eliminated from the research subject pool. The four remaining speakers included in this study are those whose tone error rate is smaller than 10% and whose performance overall is at or above advanced level.

Descriptions of the speech of the subjects:

- **Functions:** They all produced narrations and descriptions on both familiar topics (such as learning a foreign language), and unfamiliar topics (such as Covid-19 pandemic). In their speech, they constructed well-supported arguments, including details of evidence in support of a point of view.
- **Context/content:** In their speech, they covered content areas that are of both personal and general interest. Also, there is some evidence showing that they are able to dive into more abstract notions (such as freedom or identity).
- **Text type:** They were producing paragraphs that are organized and detailed.
- **Language control:** Native speakers of Chinese with no training or experience working with nonnative speakers would have no difficulty understanding their speech. They master the grammar and syntax well with few mistakes. They use special constructions such as the Ba-construction or the resultatives accurately both in terms of forms and in terms of functions.
- **Vocabulary:** They used a good variety of vocabulary that are suitable for the topic and the contexts.
- **Communication strategies:** Although self-correction is not abundant (since they are already very accurate), there is clear evidence for elaboration, clarification and circumlocution.
- **Cultural awareness:** These speech samples are delivered in culturally appropriate manners and the speakers demonstrated cultural knowledge in their presentation of the topics.

Although evaluators are not able to conduct full OPI interviews and to test all the aspects of their speech performance to establish a performance ceiling, the floor of their proficiency is at the ACTFL advanced level. Without probing into a higher level, it is unknown where the ceiling of their proficiency is.

There are advantages and disadvantages using the spontaneous connected speech samples from publicly accessible platforms such as YouTube. The main advantage of using these spontaneous speech samples is that they reflect the natural status of the L2 language. These clips were produced for non-research purposes, with an intention that centers on the message (content) rather than linguistic forms, including the tones. The speakers each had a personal channel on YouTube, where they release such videos frequently covering topics from life to study, from society to politics. The samples were taken from their channels randomly. To ensure the consistency of performance, the selected samples are produced within a year of time for each speaker. The downside of using these spontaneous speech samples comes from a few directions. First, the quality of the sound is not ideal. The recording equipment, recording environment and the recording skills all affected the quality of the sound. These clips are definitely not research materials that can be used for acoustic analysis. Therefore, in the present study, trained native speakers' judgments are employed to assess the data and mark errors. This method is appropriate with regard to the data but may have missed important nuances in tonal production. The second issue with spontaneous connected speech concerns the representativeness of the data, an issue raised by many SLA researchers (see C. Chaudron 2003 for general discussions of

Table 2 Proportions of the four tones in each speakers' samples

	J (%)	M (%)	F (%)	X (%)
T1	15.9	17.1	17.0	17.4
T2	19.8	18.0	17.6	16.5
T3	20.7	19.2	18.6	20.9
T4	30.7	32.9	34.0	31.4
Tn	11.0	10.6	10.7	11.4

SLA data collections). It is well known that L2 speakers avoid the structures, words and sounds that they do not master well. In spontaneous speech, we may not observe what they do not do well, and studying what is present in the samples may not reveal the entire picture of their L2 status. In the present study, this issue is not very serious because each of the four tones are so abundant that it is impossible to completely avoid using a particular tone or a tonal combination. Moreover, the overall proportions of the four tones stay rather consistently across each speech clip and across speakers,³ as shown in Table 2.

The third issue is not a problem but rather a disadvantage of spontaneous speech data collection compared to experimental data collection. In spontaneous speech, the forms/problems under investigation may not appear frequently enough, resulting in insufficient data for further analysis. To solve this problem, a large amount of spontaneous speech must be collected and analyzed in order to collect enough relevant data to answer the research questions, while in experimental studies, the target forms are designed to be solicited and produced by the subjects. In the present study, since the speakers are fluent and the length of their speech clips are long enough, there are enough tones of each category and each combination for the purpose of production error study. These speakers also produce nearly error-free speeches, so that transcribing and scoring errors are not as difficult as working on speeches from lower proficiency level L2 speakers of Chinese. As the data will show in the next section, not only is the overall number of errors in each speaker's tonal production small, but the types of errors are limited to a very small number of tone categories and tone combinations. Moreover, the error substitutions are almost all from the Mandarin tone inventory. The four speakers are all native speakers of English, but they learned Mandarin in different settings (e.g., different Chinese programs, teachers, etc.). The following table lists the basic information about these speakers (Table 3).

In Zhang (2010, 2013, 2018), subjects who speak different first languages form different groups in experimental settings. This allows Zhang to compare data from the different groups and isolate the L1 transfer effect from the influences from other factors including linguistic universals and pedagogical choices. In the present study, due to the limited availability of the speech samples from speakers at this proficiency

³There are not many well-reported studies on the proportions of tones in native speaker's natural speech. The estimates have been that T4 has the highest proportion, and L2 the lowest. FLA studies such as (Van de Wijer and Sloos 2014) used high frequency lexical words to calculate the proportion of the tones in L1 development and reached similar conclusions.

Table 3 Basic information about the four subjects^a

Subject	Gender	Age	Native language	Studied and/or lived in China
J	M	20–25	English (Canadian)	10 years
M	M	31	English (American)	9 years
F	M	30	English (British)	10 years
X	F	26	English (American)	8 years

^aThis information was obtained from the speakers' public social media accounts such as YouTube channel, Bilibili.com channel or Weibo account

level, the four speakers included in this research all speak English as their native language. In future studies, more speech samples from speakers of other languages should be included to draw a more complete picture of the tone errors and error patterns.

The four speakers each have many videos on various topics on their YouTube channels. Clips (total length of 20 min or so) from each speaker were randomly selected⁴ and transcribed into Chinese characters and pinyin by four experienced Chinese language teachers (8-10 years of teaching experience). Transcriptions were reviewed by at least one other transcriber. The speech of these four speakers is very close to that of the native speakers, so the transcribers found only one or two words in each of the video clips that they did not agree on or they could not decide what they heard. These words were marked with a few possible interpretations. Then two Chinese phonologists (including the author) listened to the video clips and assessed the tonal productions of each syllable and marked them as either “correct” or “incorrect,” based solely on their language intuition in that connected speech context. In other words, the evaluations were not based on citation form or any standard form. If marked as “incorrect,” the substitute tone was transcribed using the 1-5 scale tonal marking system (Chao 1930, 1948). Each of the phonologists did so independently and did so at least twice, with a minimum of one month between each transcription. The results from the two phonologists have an overall 98.3% rate of agreements. Where the two phonologists disagreed, a third phonologist was invited to make an independent judgment. A final discussion of the three phonologists resolved most disagreements except in two cases. These two cases were excluded from the data calculation. The following table shows the basic information of the video clips and the transcriptions (Table 4).

After tone errors were identified and transcribed, tonal contexts were also annotated for further analysis. Using these raw materials, here are the actions taken to answer the research questions raised above.

1. What kind of tone errors do these learners make?

Action 1: Check the tonal errors found in our transcriptions. Identify the target tonal categories and the substitutions.

⁴In these speakers' channels, the lengths of the video clips are very different from each other. Very short and very long ones were excluded from the total before the random selection.

Table 4 Basic information of the speech samples

Speaker	Topics	Total length	Total syllables	Total incorrect tones
J	Cultural comparison Language learning Personal identity Personal history	21'42''	4670	117 (2.5%)
M	Covid-19 Language learning	22'36''	5694	88 (1.5%)
F	Covid-19 Brexit	24'17''	3884	106 (2.7%)
X	Covid-19 Language learning	18'45''	4733	299 (6.3%)

2. Do the errors they make fall into some common categories? If so, what are the categories of these errors?

Action 2: Based on the results from step 1, compare data from different speakers.

3. Are these errors similar to or different from the errors identified from previous studies done mostly on learners of lower proficiency levels?

Action 3: Based on the results from Step 2, compare the findings among the four speakers in this study to the findings in previous studies.

4. If so, do phonological universals (TMS, TPC and OCP) explain these errors? To what extent, do they explain these errors?

Action 4: The results from Step 2 and 3 will be discussed to identify the causes of the errors, following part of the framework laid out in Zhang (2010, 2013), in which OPC, TMS and TPC are the three types of constraints among the universal phonological constraints.

5. What information do these errors reveal to us about the development of the IL at this very final stage of acquisition?

Action 5: Rather than a separate step, it is more of a co-occurring action with Action 4. During the discussions about the causes of the errors, native speakers' tonal phonology will be brought up and used as a point of reference. By doing so, we will learn about the stage of development of IL and project the changes that are needed to further approximate the L1 tonal phonology of Mandarin Chinese. The universal phonological constraints and their ranking (and re-ranking) will be used as the theoretical tools in the discussion as well.

3 Data and Analysis

3.1 TMS

In the previous section, research questions and actions have been laid out. In this section, Actions 1 to 4 will be carried out. Action 5 will be done in Sect. 4 of the paper. Tables 5 and 6 show the errors in terms of tonal categories for each of the four speakers. In Table 5, the target tone categories are listed in the first row. In each speaker’s row in each cell, the first number is the total number of errors in this target

Table 5 Tone errors in each tone category and their proportion in the total errors made by each speaker

	T1	T2	HT3	T4	FT3	T5	Tn	Total
J	18 15.4%	80 68.4%	6 5.1%	13 11.1%	0 0%	0 0%	0 0%	117 100%
M	0 0%	64 72.7%	4 4.5%	18 20.5%	0 0%	2 2.3%	0 0%	88 100%
F	12 11.3%	73 68.9%	3 2.7%	15 14.2%	1 0.9%	1 0.9%	1 0.9%	106 100%
X	13 4.3%	169 56.5%	7 2.4%	108 36.1%	1 0.3%	0 0%	1 0.3%	299 100%
Overall	43 7.0%	386 63.3%	20 3.3%	154 25.2%	2 0.3%	3 0.5%	2 0.3%	610 100%

Table 6 Tone errors in each tone category and their proportion in the total number of targets

		T1	T2	HT3 + FT3	T4	T5	Tn	Total
J	Total target no. of errors	741	924	966	1434	93	512	4670
		18	80	6	13	0	0	117
	Percentage	2.4%	8.7%	0.6%	0.9%	0%	0%	2.5%
M	Total target no. of errors	975	1023	1094	1875	124	603	5694
		0	64	4	18	2	0	88
	Percentage	0%	6.3%	0.4%	1.0%	1.6%	0%	1.6%
F	Total target no. of errors	661	684	723	1319	83	414	3884
		12	73	4	15	1	1	106
	Percentage	1.8%	10.7%	0.6%	1.1%	1.2%	0.2%	2.7%
X	Total target no. of errors	825	781	988	1487	112	540	4733
		13	169	8	108	0	1	299
	percentage	1.6%	21.5%	0.8%	7.3%	0%	0.19%	6.3%
Overall	Total target no. of errors	3202	3412	3771	6115	412	2069	18981
		43	386	22	154	3	2	610
	Percentage	1.3%	11.3%	0.6%	2.5%	0.7%	0.1%	3.2%

Table 7 Tests comparing the error rates between tone categories

	T2 versus T4	T4 versus T1	T1 versus T3	(HT3 + FT3) versus T5
<i>p</i> value	<0.0001	<0.0001	<0.001	0.36
<i>z</i>	17.80	3.75	3.29	0.36

tone category and the percentage below this number is the proportion of this number to the total number of errors made by this speaker. For example, J made a total of 18 errors for target T1 and that 18 makes 15.4% of the total 117 tone errors in J's speech sample.

From Table 1 in the previous section, it is seen that each tone category has different numbers of total tone production in the speech. Some tone categories, such as T4, have higher frequency of occurrence than others. In natural native speaker's speech as well as the speech samples collected from these four speakers in this study, the number of tones in each category is not evenly distributed. In other words, some tones simply occur more frequently than others. One would expect that the more often a tone category is pronounced, the more tone errors in this category will occur. So higher frequency of tone errors in one tone category in a sample does not indicate that tone category is more difficult to pronounce. To avoid such possible misrepresentation, in Table 6, the numbers of tone errors are compared to the total target tones in each category, so that we can see whether there are more errors produced for each tone category proportionally.

The occurrences of errors given in the two tables above both show that the four speakers made more mistakes when they pronounced a T2. In three of the four speakers, T4 comes second. T3, whether in the HT3 subcategory, FT3 subcategory or in the T5 subcategory, has very few mistakes. When testing the numbers in Table 6 using statistical *z* tests, the following results, as shown in Table 7, are produced: the error rates of T2 are significantly higher than those of T4, which is significantly higher than those of T1, which is significantly higher than those of T3. Between the error rates found for T3 and T5, there is no significant difference.

The ratio of errors in each tone category and the statistical test results provide support to the discoveries found in previous studies (Miracle 1989; Shen 1989; Leather 1990; Elliot 1991; Guo 1993; Chen 1997; Sun 1998; Zhang 2010, 2013). In these studies, relative difficulty of the four tones to L2 learners is summarized in Table 8. The asterisks used in "the order of difficulties" indicate disfavor following the OT convention used in Zhang's studies, so *T2 > *T3 means that T2 is more difficult than T3.

In most previous studies, T1 has been shown to have a low error rate in production. There are different conclusions regarding the most difficult tone for L2 learners, possibly due to different kinds of evidence found in different production tasks (repeating, responding, reading or speaking). T2 appears to be difficult, ranking as the most or the second most difficult. Different reasons were given to explain the

Table 8 Ranking of difficultness from previous studies

Study	Order of difficultness	Length of subjects' experience of learning Chinese 1
Miracle (1989)	*T2 > *T3 > *T4 > *T1	At least 1 year
Shen (1989)	T4* > *T1 > *T3 > *T2	4 months
Leather (1990)	*T3 = *T2 > *T4 > *T1	Minimal
Elliot (1991)	*T3 > *T2 > *T4 > *T1	1 semester to 400 level
Chen (1997)	*T3 > *T2 > *T4 > *T1	1–2 years
Sun (1998)	T2 being the most difficult	1–3 years
Guo and Tao (2008)	*T3 > *T2 > *T4 > *T1	1 and 2 semesters
Zhang (2010, 2013)	*T2 > *HT3 > *T4 > T5 > *T1 (English L1 speakers)	6–18 months

ranking of error rates in these studies, ranging from articulatory reasons, to production–perception relationship, to L1 prosodic transfer effects, and to pedagogical practices. In Zhang (2010, 2013, 2018), TMS was proposed to be the cause. According to the TMS *Rising \gg *Falling \gg *Level (Ohala 1978, Hyman and VanBik 2004) and *High \gg *Low (Yip 2002), Zhang proposes that in Mandarin, T2 is more difficult than T4, which is more difficult than T1 (Ranking *FT3 \gg *T2 \gg *T4 \gg *T1 \gg *HT3 (Zhang 2013)). The results from these studies have been used to answer the question regarding the “order of tone acquisition,” assuming that the tones that are pronounced with a higher rate of errors are more difficult to acquire and will be acquired later. This assumption has not been verified by longitudinal studies. The results from our study provide valuable insight into this issue by revealing the actual tone production at the very end of the acquisition process, if there is such an order. The numbers in Table 5 show that T2 errors are more than half of the total errors in all four speakers. In fact, the proportion of T2 errors in the total number of errors increases when the speaker’s overall tone error rate decreases as shown in Table 9. This may indicate that as speakers master tonal production better, errors in other categories disappear faster than those in T2, resulting in more T2 errors in the total number of errors. Unfortunately, the current study has only four speakers and this stipulation about developmental change deserves further study.

Next to T2, T4 has the second highest error rate in three out of the four speakers. In many previous studies (Leather 1990; Elliot 1991; Chen 1997; Guo and Tao 2008; Zhang 2010, 2013, 2018), T3 is reportedly difficult, while in our data, T3, including all three of its variants, has a very low error rate. Summarizing the results from

Table 9 Proportion of T2 error and the overall error rate of each speaker

	M	J	F	X
T2 error in all tone errors	72.7%	68.4%	68.9%	56.5%
Overall error rate (number of errors as a percentage of total syllables produced)	1.5%	2.5%	2.7%	6.3%

Tables 5 and 6, the rank of difficulty is $*T2 > *T4 > *T1 > *T3$ (including HT3, FT3 and T5). This is very close to what TMS predicates in Zhang (2013), except that FT3 is not problematic among the four speakers in the current study. Containing three tone targets, FT3 is a contour tone of the most complicated type. The fact that in all four speakers' speech, it is almost error free may be explained in a few ways. First, these four speakers are all very advanced. They have mastered the articulatory mechanisms of tones well. Secondly, in connected speech, FT3 is not abundant and has predictable occurrences, namely in isolated syllables and in prosodic-final positions. The fact that FT3 is almost problem free in these speech samples suggests that the causes for tone errors at this stage of IL are beyond articulatory reasons. In Zhang (2013), the high error rate of T3 found in the experiment ($*T2 > *HT3 > *T4 > T5 > *T1$ (English L1 speakers)) does not fit the predicted outcome based on the proposed TMS ($*FT3 \gg *T2 \gg *T4 \gg *T1 \gg *HT3$). The proposed explanation is the inappropriate pedagogical practice that emphasizes FT3 as the underlying form of T3. The outcome in the present study, in which HT3 behaves the way that the TMS predicts, suggests that the learners' IL may have gone through significant revisions, through which the underlying form (base form) of T3 was reset from FT3 to HT3. Another interesting point to make here regarding the errors in different single tone categories is about T5 or the sandhi form of T3 when it is followed by another T3. There are phonological and phonetic studies on the very nature of this tone particularly around the issue of whether it is distinguishable by native speakers or if it has distinguishable phonetic cues from those of real T2's (Zee 1980; Kratochvil 1987; Xu 1997 to name a few). In our data, T5 has a similar error rate to other variants of T3, which is much lower than that of T2. But we will see in the following discussion that this is related to the distributional properties of T5 and does not really provide many clues with regard to the debate surrounding T5's distinctiveness from T2 in Mandarin Chinese.

After examining the error rates in different single target tone categories, we now move on to the substitute forms of these errors. Previous studies of tone errors in L2 learners of Mandarin reveal that learners often produce tones that are out of the L1 Mandarin tone inventory (including both the citation forms and the legit variants in connected speech). Moreover, in connected speech, L2 speakers produce certain tonal variants (such as the "mid-tone" proposed in Q.H. Chen 2000) that are different from the citation form of the tones but are found in L1 speakers' connected speech. However unlike the L1 speakers, L2 speakers pronounced these variants in illegitimate locations. In the current research, tones in the connected spontaneous speech samples were judged by native speakers against the acceptable form(s) in native speakers' speech. If such variants occur in the places where native speakers would accept as possible forms, then they are not counted as errors. In other words, a tone is counted as an error only if it cannot be accepted by a trained native speaker in these contexts. Therefore, many "mid-tones" were counted as acceptable productions. In the speech samples in this study, the vast majority of the substitute forms do not fall outside of this inventory, and this is not unexpected for this group of high proficiency learners. Table 10 shows numbers of the erroneous substitute forms for the corresponding target tones. For example, in J's speech samples, he mispronounced T1 18

Table 10 Tone errors and their corresponding substitutes

	T1	T2	HT3	T4	FT3	T5	Total error
J	18	80	6	13	0	0	117
	HT3: 3 T4: 14 Mid-tone: 1	HT3: 79 T1: 1	T1: 4 T4: 2	T1: 13			
M	0	64	4	18	0	2	88
		HT3: 61 T1: 3	T1: 4	T1: 16 HT3: 2		HT3:2	
F	12	73	4	15	1	1	106
	HT3: 9 T4: 3	HT3: 66 T1: 7	T1: 3 T4: 1	T1: 12 HT3: 3	T1:1	HT3: 1	
X	13	169	9	108	1	0	299
	HT3: 9 T4: 4	HT3: 159 T1: 10	T1: 9	T1: 105 HT3: 2 Mid-tone: 1	T1: 1		

times, in which 3 of them were pronounced as HT3, 14 of them were pronounced as T4 and the other one was pronounced as a mid-tone.

The substitution patterns among the four speakers are strikingly similar. Overall, HT3 is the most frequent substitute for a mistakenly pronounced target tone (70% in J's speech, 74% in M's speech, 75% in F's speech and 57% in X's speech), followed by T1 and with T4 coming third. The substitute pattern confirms that the effect of TMS (*FT3 >>*T2 >>*T4 >>*T1 >>*HT3 (Zhang 2013)), which predicts that the low-level tone HT3 is the most unmarked member of the tonal system, and T1, being a high-level tone, is the second most unmarked. This substitute ranking generally aligns with the ranking found in Zhang (2013) among English speakers but there are a few minor differences between the two. First, HT3 is not only the most common substitute but also is the majority in the current study. In Zhang (2013), HT3 substitutes constitute only about 15% of the total. Secondly, although T2 and FT3 were found as substitute in Zhang (2013), in the current study, they do not appear in the speech as substitutes for a failed target tone production at all. Thirdly, T4 only appears 24 times as a substitute in our data out of the total 610 tone errors. Such differences between low/mid-level L2 speakers and advanced-level speakers suggest that the effect of TMS becomes more salient as learners' IL progresses and the effects from other factors, such as L1 negative transfer or pedagogical reasons, may have gradually faded away. When target tone information is added to the discussion, we find that: HT3 is the most frequent substitute for a target T1 (except in J's case where T4 is the most frequent substitute) or T2; T1 is the most frequent substitute for a target T4; T1 is the most frequent substitute for a target HT3. Table 11 compares our findings with the findings in Zhang (2013), where the differences are highlighted in gray.

Table 11 Ranking of difficultness in Zhang (2013) and in the current study

	Zhang (2013) (English speakers)	The current study
T1's substitute	HT3 > T4 > T2 > FT3	HT3 > T4
T2's substitute	HT3 > T1 > T4 > FT3	HT3 > T1
HT3's substitute	FT3 > T4 > T1 > T2	T1 > T4
T4's substitute	HT3/T1 > T2 > FT3	T1 > HT3
T5's substitute	HT3 > T1 > T4	HT3 (only a few cases total)

Compared to Zhang (2013)'s findings based on low-mid-level learners, the findings in this study show that the four sampled speakers have mastered the T3 sandhi with only 3 T5 errors. Secondly, in Zhang (2013) data, FT3 often shows up as the substitute of HT3, while in the current study, FT3 is not produced by any of the speakers in place of a target HT3. Instead, T1, the second most unmarked tone, appears in these places. Zhang (2013) attributes the HT3 substitute pattern to pedagogical practice. Along this line of reasoning, then the HT3 substitute pattern in our study suggests the effect of TMS surfaces more visibly among more advanced learners, who seem to have overcome the negative impact of the common pedagogical practice that, according to Zhang, promotes FT3 as the base tone. Thirdly, in Zhang (2013) data, HT3 is the most frequent substitute for all other tones (T1, T2, T4 and T5). This is almost the case in our data, except that we find the most frequent substitute for T4 is T1 by a large margin over the second most frequent substitute tone HT3. Therefore, TMS alone cannot explain the pattern here because it will predict that HT3 is also the most frequent substitute for T4. In the following section, examination of tone errors and substitution patterns in tonal combination contexts will provide a close-up view of the issues.

3.2 TPC

After examining the errors in single tones without taking into considerations its contexts, we now move onto tone error distributions within local contexts, mostly in prosodic words. Table 12 summarizes the tone errors and their position in words. In the table, "initial" means when a tone error occurs in the first syllable of a disyllabic (prosodic) word, "final" means an error occurs in the second syllable of a disyllabic (prosodic) word, "others" include mostly monosyllabic and polysyllabic words. For each speaker, under each tone category, the first row is the total number of tone errors in this tone category and then the cell below the total is broken down into number of tone errors in different positions.

HT3, FT3 and T5 have positional distributions via sandhi rules. So, the error distributions are affected by their allotonic distributions. Also, the errors in these three categories are few, so we will focus on the distributions of T1, T2 and T4 only. In all four speakers' speech, there are more errors found for a target T1 when it is at

Table 12 Tone errors in different positions in a prosodic word

	T1	T2	HT3	T4	FT3	T5	Total
J	Total 18	80	6	13	0	0	117
	Initial: 12 Final: 2 Others: 4	Initial: 61 Final: 11 Others: 8		Initial: 10 Final: 1 Others: 2			
M	0	64	4	18	0	2	88
		Initial: 53 Final: 6 Others: 5		Initial: 8 Final: 9 Others: 1			
F	12	73	4	15	1	1	106
	Initial: 9 Final: 3 Others: 0	Initial: 56 Final: 13 Others: 4		Initial: 7 Final: 6 Others: 2			
X	13	169	8	108	1	0	299
	Initial: 11 Final: 1 Others: 1	Initial: 75 Final: 71 Others: 23		Initial: 40 Final: 54 Others: 14			
Total	43	386	23	154	2	3	610
	Initial: 32 Final: 6 Others: 5	Initial: 245 Final: 101 Others: 40		Initial: 66 Final: 69 Others: 19			

word-initial positions. When T2 is the target pronunciation, more errors are produced when it is word -initial in three of the four speakers individually and overall. T4 is different in this sense from T1 and T2 because overall the number of errors at initial versus final positions is roughly equal. There are individual differences among the four speakers, especially in the case of X, who also made more errors in speech than the other three speakers. In Zhang (2013), the TPC were investigated by comparing error rates (percentage of errors at either word-initial or word-final positions) for each tone category. Table 13 is the results for the English speakers in Zhang’s study.

Zhang (2013) argues that the data in the table above (and the corresponding substitution patterns) demonstrate that “T2 is performed better at word-initial positions,

Table 13 Error rates with regard to positions in a word for English speakers (adapted from Zhang 2013: 111)

	General		English speakers	
	Initial (%)	Final (%)	Initial (%)	Final (%)
T1	18.02	25.42	24.38	30.00
T2	33.33	79.9	36.88	78.13
T4	44.17	22.29	54.69	28.44

Table 14 TPC rankings

*Contour-I \gg Contour-F (Word-final positions are better bearers of contour tones)
*Fall-I \gg *Fall-F (Falling tones are more disfavored at word-initial positions than at final positions)
*Rise-F \gg *Fall-F (Rising tones are more disfavored than falling tones at word-final positions)

and T4 is performed better at word-final positions,” and suggests that “the word-final position is a preferred bearer of a falling contour tone (T4) rather than a rising contour tone (T2)” (Zhang 2013, pp. 119). Using constraint ranking, these propositions can be expressed in Table 14. Again, the asterisks in Table 14, a convention used the OT framework, mean disfavor. So *Fall-I \gg *Fall-F means that a falling tone at word-initial positions is worse than a falling tone that occurs at word-final positions.

Applying the same counting method, the percentages of errors in each tone category at either word-initial or word-final positions found in the current study are summarized in Table 15.

The four speakers in the current study made far fewer errors in tone production overall and especially in T1 and T4. So, one can argue that the comparison between the very small percentages in these two tones is not going to yield many meaningful interpretations. But combining the findings in Tables 12 and 13, one can reasonably argue that the rates of errors in T1 indicate that even for a high-level tone, the initial position is more problematic than the word-final position. The falling tone T4 seems neutral with regard to the positional difference. With regard to T2, J, M and F, all made more mistakes at word-initial positions than word-final positions. X, the least proficient speaker of the four, made more mistakes at word-final positions. Overall, our data presented in Tables 12 and 15 support the TPC claim that the word-initial position is difficult, but it seems to be difficult for all tones rather than for contour tones only, and the rising tones are performed better at word-final positions, not initial positions. If TPC are universal, and their effects exist as detected in previous studies, then it seems such effects are masked by other effects in our speech samples. I will start from the context of error as a possible source of the confounding effects. By combining information of target tone category, substitute tone category, error tone position in a word and the neighboring tone in a word, the error patterns from the speech samples are shown in three Tables 16, 17 and 18.

It is not difficult to notice that more T2 errors show up in the tonal combinations T2-T1 and T2-T4 than in other combinations, and their corresponding substitute forms are HT3-T1 and HT3-T4. Of the four speakers, these two-tone combinations and substitute forms are over 70% of the total T2 errors in the more proficient speakers M (78%), J (70%) and F (70%). X’s speech displays more error tone combinations and substitute forms, but errors in these two still compose more than half of the total.

Both HT3 and T1 are found to be substitutes for T2. However, the detailed distribution of the two substitutes shows that HT3 is not only dominant in the number of errors where it is the substitute, but it also appears in more two-tone combinations

Table 15 Error rates with regard to positions in a word in this study

	J		M		F		X		Overall	
	Initial (%)	Final (%)	Initial (%)	Final (%)	Initial (%)	Final (%)	Initial (%)	Final (%)	Initial (%)	Final (%)
T1	3.1	0.6	0	0	3.0	0.8	2.9	0.2	2.1	0.4
T2	13.0	3.6	10.6	1.7	16.6	5.2	18.5	22.9%	14.3	8.2
T4	1.4	0.1	0.8	1.0	1.0	1.0	5.6	7.1%	2.1	2.4

Table 16 Target T2 errors and substitutions in the contexts of neighboring tones in words

Target → Substitute	J	M	F	X	Total	
<i>Target tone at word-initial positions</i>						
T2-T4	→ HT3-T4	46	25	25	35	131
T2-T1	→ HT3-T1	10	25	26	18	79
T2-T2	→ HT3-T2	1	1	2	7	11
	→ T1-T2	0	0	0	2	2
T2-T3	→ HT3-T3	0	1	0	1	2
T2-Tn	→ HT3-N	4	1	3	12	20
<i>Target tone at word-final positions</i>						
T4-T2	→ T4-T1	1	3	1	4	9
	→ T4-HT3	3	2	0	18	23
T1-T2	→ T1-HT3	7	1	6	38	52
T2-T2	→ T2-HT3	0	0	6	8	14
	→ T2-T1	0	0	0	3	3
Others						
T2	→ HT3	8	5	4	22	39
	→ T1	0	0	0	1	1
Total		80	64	73	169	386

Table 17 Target T1 errors and substitutions in the contexts of neighboring tones in words

Target → Substitute	J	M	F	X	Total	
<i>Target tone at word-initial positions</i>						
T1-T1	→ HT3-T1	0	0	1	0	1
	→ T4-T1	1	0	0	1	2
T1-T2	→ T4-T2	5	0	1	0	6
	→ HT3-T2	2	0	0	1	3
T1-FT3	→ T4-FT3	1	0	2	0	3
T1-T4	→ T4-T4	2	0	0	0	2
	→ HT3-T4	1	0	0	8	9
T1-Tn	→ HT3-N	0	0	5	1	6
<i>Target tone at word-final positions</i>						
T4-T1	→ T4-T4	1	0	3	0	4
T2-T1	→ T2-T4	1	0	0	1	2
Others						
T1	→ T4	4	0	0	1	5
Total		18	0	12	13	43

Table 18 Target T4 errors and substitutions in the contexts of neighboring tones in words

Target → Substitute	J	M	F	X	Total	
<i>Target tone at word-initial positions</i>						
T4-T1	→ T1-T1	0	0	0	2	2
	→ HT3-T1	0	2	3	3	8
T4-T4	→ T1-T4	6	6	4	30	46
	→ HT3-T4	0	0	0	2	2
T4-T2	→ T1-T2	1	0	0	3	4
T4-T3	→ T1-T3	3	0	0	0	3
<i>Target tone at word-final positions</i>						
T2-T4	→ T2-T1	1	2	1	11	15
T1-T4	→ T1-T1	0	0	0	3	3
HT3-T4	→ HT3-T1	0	4	3	12	19
T4-T4	→ T4-T1	0	3	2	28	33
<i>Others</i>						
T4	→ T1	2	1	2	14	19
Total		13	18	15	108	154

as the substitute for T2. For speaker J, M and F, HT3 is the only substitute in all two-tone patterns except in T4-T2 target. X's speech is different from the other three speakers' in that T1 is an alternative substitute tone in T2-T2 target pattern and in monosyllabic words. Even so, in X's samples, the number of HT3 substitutes is much greater than the number of T1 substitutes in the T2-T2 targets and the monosyllabic T2 targets. So, it is reasonable to postulate that HT3 is overall the dominant substitute of T2 in the very last stage of tonal acquisition. T1 substitutes may have existed more in earlier stages, but it has gradually disappeared from the speech.

There are not many errors produced for target T1, and the distribution of number of errors in different two-tone combinations is fairly even. When the target T1 is at word-initial positions, it is not clear whether HT3 or T4 is the more dominant substitution form. When the target T1 is at word-final or monosyllabic positions, the substitutes are all T4.

Among all the errors and the substitutions for target T4, one two-tone combination appears to be the most problematic for the four speakers, namely T4-T4. T4-T4 errors account for 53% of all errors in Table 18. Both the word-initial T4 and the word-final T4 have many errors, with the word-initial T4 being more difficult.⁵ Speakers J, M and F substitute the word-initial T4 with a T1 in all the erroneous productions, while X has two HT3 substitutions. In word-final T4 errors in the T4-T4 combination,

⁵As a reviewer pointed out, in a T4-T4 sequence, native speakers produce the first T4 as HM and not the full falling contour HL, a process called "T4 sandhi." In the current study when the L2 speakers made errors in producing the T4-T4 sequence by substituting the first T4 with T1, they did not articulate an HM but a tone that is judged by transcribers as T1 (HH). It is possible that this error among the L2 learners is related to the perception of the HM in the T4 sandhi.

T1 is the substitution for all four speakers. M and F each have three and two HT3-T4 combinations in their speech for mispronounced T4-T4 sequences. However, it has to be noted that in these five cases, both the initial and the final tones of the combinations were pronounced incorrectly. They are target T2-T4 combinations and were pronounced as HT3-T1. Their existence is not to be confused with those HT3-T4 combinations for target T4-T4 sequences. Both HT3 and T1 are found to be substitutes for T4. However, T1 is the only substitute when the error occurs at word-final or monosyllabic positions. T1 also appears as the substitute in more cases and in more two-tone combinations when the error occurs at word-initial positions.

Tables 16, 17 and 18 list the tone errors and the substitutions in words and their local tone contexts. When positions and neighboring tones are taken into consideration, there are a few generalizations that emerge from the data. First, tone errors occur in some tonal combinations more often than others. T2-T1 and T2-T4 target combinations are the most difficult combinations for T2, while T4-T4 is the most difficult combination for T4. The only substitute form for T2-T1 is HT3-T1 and the only substitute form for T2-T4 is HT3-T4. In the T4-T4 combination, there are practically two substitution forms: T1-T4 and T4-T1. Second, Tx (x stands for any tone) in HT3-Tx combinations (HT3-T1, HT3-T2 and HT3-T4) has the highest accuracy. There is no error produced in HT3-T1 and HT3-T2 combinations. In HT3-T4 combinations, only speaker X produced errors. The five errors for the HT3-T4 combination in M and F's speech are actually for target T2-T4, in which both the initial and final tones were pronounced incorrectly. Please note the two generalizations above are related in that the (almost) error-free HT3-T1 and HT3-T4 combinations are also the only substitute forms for T2-T1 and T2-T4, which are the most difficult combinations, indicating that HT3 is the least marked tone in this context.

Looking at the frequencies of each error and their percentages in the overall target combinations is not enough. In natural connected spontaneous speech, tones and tone combinations occur in different frequencies. For example, T4 has a higher frequency of occurrence than the other tones, which could mean that two-tone combinations with T4 may naturally occur more often, and more errors are expected even if the possibility of making such errors is the same as making errors for other target tone categories. Therefore, it is necessary to measure the number of errors for each two-tone combination against the total occurrence of the target two-tone combination, and the results of such measurement for T2 errors are given in Table 19 below.

The ratios in Table 19 show how often T2 is pronounced incorrectly in a particular two-tone combination as a percentage of the total production of the target combination. It is very clear that the T2-T1 combination has the highest error rate for all four speakers and is 44.13% overall. T2-T1 target combination does not seem to occur at high frequency in the speech samples, but when they occur, our subjects tend to make more mistakes in producing the T2 in this combination. In fact, **T2-T1** has the highest error rate among all possible two-tone combinations in all four speakers. For J, M and F individually and for all speakers overall, the second highest error rate comes in T2-T4 combination. Although T2-T4 errors are many, there are significantly more correct T2-T4 productions in the speech, resulting in a lower error rate than the one found in T2-T1 combination, but it is still higher than other combinations. Because

Table 19 Target T2 errors and substitutions in the contexts of neighboring tones in words measured as a percentage in the total target combination

Target		J	M	F	X	Total
<i>Target tone at word-initial positions</i>						
T2-T4	Total targets	238	237	165	171	811
	no. of errors	46	25	25	35	131
	Percentage	19.3%	10.5%	15.2%	20.5%	16.2%
T2-T1	Total targets	37	60	45	37	179
	no. of errors	10	25	26	18	79
	Percentage	27.0%	41.7%	57.8%	48.6%	44.1%
T2-T2	Total targets	41	74	44	59	218
	no. of errors	1	1	2	9	13
	Percentage	2.4%	1.4%	4.5%	15.3%	6.0%
T2-T3	Total targets	87	80	51	103	321
	no. of errors	0	1	0	1	2
	Percentage	0%	1.3%	0%	1.0%	0.6%
T2-N	Total targets	68	49	32	36	185
	no. of errors	4	1	3	12	20
	Percentage	5.9%	2.0%	9.4%	33.3%	10.8%
<i>Target tone at word-final positions</i>						
T4-T2	Total targets	68	110	62	82	322
	no. of errors	4	5	1	22	32
	Percentage	5.9%	4.5%	1.6%	26.8%	9.9%
T1-T2	Total targets	125	91	97	109	422
	no. of errors	7	1	6	38	52
	Percentage	5.6%	1.1%	6.2%	34.9%	12.3%
T2-T2	Total targets	41	74	44	59	218
	no. of errors	0	0	6	11	17
	Percentage	0%	0%	13.6%	18.6%	7.8%
T3-T2	Total targets	70	86	48	60	264
	no. of errors	0	0	0	0	0
	Percentage	0%	0%	0%	0%	0%

the numbers of errors in other two-tone combinations are very small, with error rates in the low single digits, it is not very meaningful to compare those numbers in the current study. What is clear is that T2-T1 is the most difficult, followed by T2-T4. Data in Table 19 support the generalization made based on Table 16 that T2-T1 and T2-T4 are the most difficult combinations when it comes to T2 targets. The high error rate in these two two-tone combinations contributes significantly to the overall high rate of error of T2 at word-initial positions and the overall high rate of error of T2 in general. This explains the findings regarding TMS and TPC discussed earlier in the paper.

The examination of the data in the current study can be summarized pertaining to the TMS and TPC. With regard to TMS, T2 has the highest error rate (as a proportion of all T2 target production) and T2 errors make up the largest portion of all tonal errors. T4 has the second highest error rate and T4 errors make up for the second largest portion of all tone errors. T1 comes third but the overall error rate for T1 target is very low. HT3 is the least marked in production both in terms of error rate and its high frequency of occurrence as the substitute tone. With regard to TPC, T1 and T2 are both performed better at word-final positions while T4's performance is about the same in the initial position or the final position. Although this finding is not consistent with the linguistic predictions (Zhang 2004) that T2 is expected to be pronounced more accurately at word-initial positions compared with word-final positions, and the findings in previous studies, a closer look at the tonal context reveals that high error rates in a few particular tonal combinations can explain the higher error rate of T2 at initial positions, indicating that tone-position constraints (i.e., initial versus final positions) interact with tonal combination factors. T2-T1 and T2-T4 sequences together account for more than 34% of the total tonal errors and 54% of the total T2 errors regardless of positional difference. The difference between the findings in the current study and previous studies suggests that when learners progress into higher tone production accuracy, the general effects of TPC are masked by specific effects on certain tone combinations, while the effects of TMS are amplified, as the higher error rates found in T2-T1 and T2-T4 significantly contributed to the higher error rate of T2. The number of high error rate combinations may become smaller as the learners become better with tones, while their proportions in total number of errors increase. This is seen in the performance of the three more accurate speakers J, M and F, whose errors are distributed with higher concentration in T2-T1 and T2-T4. The fourth speaker X has more errors in more combinations. Such differences suggest that these two combinations are the most difficult and may be the last ones to master in the tonal acquisition process. In most previous studies, they were not identified because the subjects in those studies have not reached this stage of IL yet, and the numbers of errors in other patterns, sequences and contexts are large and the overall distribution of errors is (more) widely spread.

3.3 OCP

The third type of constraint that concerns us is OCP. The error rates and patterns in Table 17 indicate that T4-T4 is difficult while those in Tables 16 and 18 indicate that most other identical tone combinations, including T2-T2 and T1-T1, are not difficult. Out of the 403 total target T4-T4 articulations, there are 81 mispronounced ones (20.1%), with either the first T4 or the second T4 pronounced incorrectly (in our data, no T4-T4 combination was pronounced incorrectly in both T4s). This error rate is the highest among all two-tone combinations that involve a T4. Moreover, those T4-T4 errors make up 52.6% of the total 154 errors found in any two-tone combination that involves a T4. In contrast, out of the 218 total target T2-T2 articulations,

there are 30 mispronounced ones (13.8%), with either the first T2 or the second T2 pronounced incorrectly (in our data, no T2-T2 combination was pronounced incorrectly in both T2s). This rate of error is way below the 44.1% rate found for T2-T1 combination. Those 30 T2-T2 errors make up 7.8% the total errors found in any T2 two-tone combinations. And there are only 3 mispronounced T1-T1 sequences out of the 235 targets. Zhang (2013) investigates the occurrences of identical tone combinations in L2 Chinese learners and finds that such combinations are in general fewer in production, and more T1-T1 sequences are found than T4-T4 sequences, which are more than T2-T2 sequences. Zhang argues that the low frequency of such Tx-Tx combinations indicates a higher level of difficulty for these sequences and suggests the effects of OCP. Zhang's method involves a statistical comparison between the actual occurrences of the Tx-Tx sequences and the expected frequencies of occurrences. Due to the nature of spontaneous speech, it is impossible to calculate the expected frequencies of Tx-Tx targets and compare them to the actual occurrences in this study. Speakers may have different preferences of words and word combinations, which result in non-random distribution of tones and tone combinations. But we can compare the numbers of errors in each Tx-Tx to the actual total Tx-Tx targets. Table 20 lists the total number of errors for each Tx-Tx combination and their percentage in the overall Tx-Tx targets. For example, the four speakers made 30 tone errors in T2-T2 combinations, which equals to 13.8% of the 218 total T2-T2 targets. Those 30 errors are 8.7% of the 346 total T2 errors made in any two-tone combination with a T2 in it (including T2-Tx and Tx-T2). The rates of Tx-Tx errors out of Tx-Tx targets show that T4-T4 has the highest error rate (20.1%), followed by T2-T2 (13.8%), and T1-T1 error rate is only 1.3%.

Zhang (2013) conducts another test to investigate possible OCP effects. The error rates of Tx in identical tone sequences (ITC) are compared with the same tone's error rates in non-identical tone sequences (NITC) using Chi-square tests. Such error rates are compared separately at word-initial and word-final positions. For example, T2

Table 20 Tx-Tx errors found in total Tx-Tx targets

	T1	T2	T4
Overall number of Tx	3202	3412	6115
Overall proportion of Tx in the speech	16.9%	18.0%	32.2%
Overall number of Tx-Tx combination targets	235	218	403
Overall number of errors found in the Tx-Tx combination targets	3	30	81
Overall error rate of the Tx-Tx combination	1.3%	13.8%	20.1%
Overall proportion of Tx-Tx errors among all Tx two-tone combination errors	7.9% (3 out of 38)	8.7% (30 out of 346)	60.0% (81 out of 135)

error rate in **T2-T2** sequences (where the error is at initial positions) is compared with T2 error rates found in T2-T1, T2-T3, or T2-T4 sequences, where T2 is at initial positions. Then the T2 error rate in **T2-T2** (where the error is at final positions) is compared with T2 error rates found in T1-T2, T3-T2, T4-T2 sequences, where T2 is at final positions. No significant difference was found in the English speakers' data. Table 21 summarizes the test results.

Using the same method to test the data from the four speakers in the current study, we obtained the results shown in Table 22. Statistically significant results are found for T4 at both initial and final positions and T2 only at initial positions. However in

Table 21 OCP test comparing tone error rates found in ITC and NITC contexts from Zhang (2013)

Test items	NITC error rate ^a (%)	ITC error rate	Chi-sq <i>p</i> value	NITC/ITC
T1 initial	24.17	25	0.8805	0.97
T2 Initial	39.17	30	0.1411	1.31
T4 initial	53.75	57.5	0.5595	0.93
T1 final	29.17	32.5	0.5731	0.9
T2 final	77.08	81.25	0.4350	0.95
T4 final	31.25	20	0.0534	1.56

^aThe method of calculation for NITC and ITC error rate in Zhang (2013)

“Tx” refers to the test tones under discussion, “Tx” could be T1, T2, T4

“Ty” means any real stimuli mandarin tone other than Tx. For example, when Tx = T1, then Ty could be T2, T3 and T4

“E” means erroneous tones for target Tx, i.e., any substitute tone for Tx; or the error rates

“N (Tx > Tx/_Tx)” are the number of times that the learners correctly produced a target Tx as a Tx in the context _Tx. This context is labeled as Tone L = Tx in the test below 102

“N (Tx > E/_Tx)” is the number of times speakers incorrectly produced a target Tx as an E in the context _Tx (Tone L = Tx)

When Tone L = Tx, the error rates for target Tx in the two contexts of “_Tx” and “_Ty” are

$E(Tx/_Tx) = N(Tx > E/_Tx) / (N(Tx > E/_Tx) + N(Tx > Tx/_Tx)) \rightarrow$ ITC context

$E(Tx/_Ty) = N(Tx > E/_Ty) / (N(Tx > E/_Ty) + N(Tx > Tx/_Ty)) \rightarrow$ NITC context

When Tone R = Tx, the error rates for target Tx in the two contexts of “Tx_” and “Ty_” are

$E(Tx/Tx_) = N(Tx > E/Tx_) / (N(Tx > E/Tx_) + N(Tx > Tx/Tx_)) \rightarrow$ ITC context

$E(Tx/Ty_) = N(Tx > E/Ty_) / (N(Tx > E/Ty_) + N(Tx > Tx/Ty_)) \rightarrow$ NITC context

The test compares the error rates of Tx at ITC contexts and NITC contexts respectively

Table 22 OCP test comparing tone error rates found in ITC and NITC contexts in this study

Test items	NITC error rate (%)	ITC error rate (%)	Chi-sq <i>p</i> -value	NITC/ITC
T1 initial	0.7	1.2	0.75	0.61
T2 initial	15.5	6.0	3.76 (P < 0.01)	2.60
T4 initial	0.8	4.7	7.26 (P < 0.01)	0.17
T1 final	0.5	0.0	1.11	∞
T2 final	8.3	7.8	0.26	1.07
T4 final	1.9	3.6	2.70P < 0.01)	1.53

the T2 initial position test, the NITC/ITC rate is larger than 1, meaning that there are more errors made in the NITC contexts. This anti-OCP effect is expected because in previous discussion, it is shown that T2-T4 and T2-T1 combination targets have many more errors than T2-T2. The two tests involving T1 both have insignificant results, suggesting that there is not enough evidence for T1-T1 OCP effect. In all, the test results in Table 22 only support the existence of OCP effect in T4-T4 sequence.

4 Discussion and Conclusions

In the previous section, tone error and substitution data were presented and analyzed. Based on the analyses, the first four research questions have been answered.

4.1 Summary of Answers to the Research Questions

The four speakers made more errors in T2, followed by T4, which is followed by T1. T3, including its variants, has the lowest error rate. HT3 is the most dominant substitute form for T1, T2 and T5, while T1 is the most dominant substitute form for T4. The initial position of a two-tone word is more difficult for T1 and T2, but there is not much difference found in T4 error rates with regard to the positions. When tonal errors are analyzed in two-tone combinations, a few sequences stand out as the most difficult. They are T2-T1 and T2-T4 for T2 and T4-T4 for T4. The existence of the high error rate sequences (T2-T1 and T2-T4) could explain the positional differences found for T2 as well as its overall high error rate. The substitute forms for these two-tone combinations also support that HT3 is the least marked tone because the error-free HT3-T1 and HT3-T4 combinations are also the only substitute forms for T2-T1 and T2-T4, which are the most difficult combinations.

The types of errors, the error rates and substitution patterns are very similar among the three more proficient speakers' speech. Overall T2 is the most difficult tone but the proportion of T2 errors in the total number of errors in each speaker's speech negatively correlates to the overall error rate among the four speakers. X, being the least proficient of the four, generally demonstrates the same types of errors, error rates and substitution patterns. However, she made more mistakes in more tone combinations types and the numbers of errors are less concentrated in different sequences.

The findings in the current study pertaining to TMC differ from previous studies where participants were mostly low and intermediate proficiency learners in that T3 does not appear to be difficult in the speech of the four speakers in this study who are advanced/superior level Chinese learners. The current study confirms the previous claim that T2 is among the most difficult tones. Unlike the conclusions made in previous studies, the error study in this research does not support TPC, which predicts that T2 favors word-initial positions and T4 favors word-final positions. A

new discovery of the current study is that among very high proficiency speakers, tone errors occur in higher concentration in only a few combination contexts, based on which I am hypothesizing a common route of IL tonal development: errors are more widely distributed among different tone categories, and in different contexts, but as learners progress, only some combinations remain difficult for the speakers. This route of IL tonal development cannot be explained by TMC, TPC, OCP, L1 transfer or pedagogical reasons alone. The current study provides strong support to TMC (*FT3 >>*T2 >>*T4 >>*T1 >>*HT3, Zhang 2013) except that in the current study, FT3 does not appear to be the most difficult. The effects of TPC in our data seem to be weak. OCP effect is only verified in T4-T4 combination but not in T1-T1. T2-T2 combination displays an anti-OCP effect due to the high error rates found for T2-T1 and T2-T4. In the following section of the paper, I will argue that the high error rate found in T2-T1 and T2-T4 sequences, when studied with the corresponding substitution patterns, points to a special configuration of a coarticulation rule as the source of error among these speakers.

4.2 Rule Configuration as the Explanation

One special characteristic of the four subjects in this study is that they are all very advanced in all aspects of pronunciation. It is reasonable to assume that their L2 Mandarin phonological system is very close to that of native speakers'. An examination of the articulation of the respective tone combinations in native Mandarin speaker's speech will provide some clues to the issues under investigation. In the study of Mandarin tones, besides the canonical sandhi rules, there are a couple of so-called tonal coarticulation rules, which manifest themselves in natural connected speech. Wu (1982, 1985), X. N. Shen (1990, 1992) and Shih (1988, 1991) are among the early studies on the tonal coarticulation phenomenon in Mandarin. Chen M. (2000) summarized the findings based on Shih's study and converted the numeric pitch values into tone category representations, shown in Table 23.

The shaded cells in Table 23 are the combination sequences where tonal coarticulation effects are more salient. The superscript + and- represent up or down shift. Shih (1988) captures these effects into four points, which are stated in formal rules in M. Chen (2000), as shown in Table 24.

Table 23 Tonal coarticulation in Beijing Mandarin (M. Chen 2000: 24)

fu-	-ji			
	T1 HH	T2 MH	T3 ML	T4 HL
T1 HH	HH # HH	HH # MH	HH # ML	HH # HL
T2 MH	MH ⁻ # HH	MH # MH	MH # M ⁺ L	MH ⁻ # HL
T3 ML	ML # HH	ML # M ⁻ H	=T2 + T3	ML # HL
T4 HL	HM # HH	HM # MH	HM # ML	HM # HL

Table 24 Tonal coarticulation rules in Beijing Mandarin (Chen M. 2000: 24)

1. T4 + anyT	HL → HM/___T (all Ts in Mandarin) starts with either H or M)
2. T2 + T3	ML → M ⁺ L/MH___
3. T3 + T2	MH → M ⁻ H/ML___
4. T2 + {T1, T4}	MH → MH ⁻ /___Hx (x may be H or L)

Table 25 Coarticulation rule 4 in Beijing Mandarin (M. Chen 2000: 25)

Base form	MH. HL (T2 + T4)		MH. HH (T2 + T1)	
Step 1 in derivation	M_. HL	Tone absorption	M_. HH	Tone absorption
Step 2 in derivation	MH⁻. HL	Tone interpolation	MH⁻. HH	Tone interpolation

M. Chen (2000) explains that, of the four rules given above, the first three “are all assimilatory in nature,” while the fourth rule looks like a dissimilatory rule but can also be interpreted as a tone absorption followed by a tone interpolation process, sketched out in Table 25. According to Chen, the base form for T2 + T4 is MH. HL, which loses the H in MH during the first step derivation when tone absorption rule is applied. Then in the second step derivation, tone interpolation rule applies, and the output is **MH⁻. HL**. A similar process applies to T2-T1 sequence. Treating the fourth rule as a tone absorption and a tone interpolation process allows a unified account for all four coarticulation rules: they all serve to smooth the transition between tonal targets.⁶

Although the function of the four rules is unified in Chen’s explanation, the processes involved in the fourth rule are different from those found in the first three rules.

The examination of the L1 Mandarin tonal coarticulation rules allows us to contemplate an explanation for the high error rates found for T2-T1 and T2-T4 combinations in the four speakers’ speech in the current study. These two high error rate combinations correspond to the coarticulation forms described by the fourth rule of coarticulation in L1 discussed above. The only substitute form found in incorrectly pronounced T2 in these two combinations is HT3 (ML) while the correct form should be MH⁻, under the coarticulation effect in the speech. Such a correspondence between the errors found in the last stage IL and the coarticulation forms in L1

⁶As one reviewer pointed out, besides the phonological account proposed by Chen, the tonal phenomenon also received phonetic explanations (e.g., Xu 2001), which postulate that the downstep effect is due to the peak delay where the target H is realized in the following syllable. Thus, the pitch height of the first syllable was not as high as if it was pronounced in citation. My standing point is that these two theories are both valid on their own premises, and both capture one key issue: the T2-T4 and T2-T1 combinations are realized differently from other tonal coarticulation combinations. When it comes to L2 learners, such a difference causes their L2 phonological system to form different phonological rules/procedures/configurations. A second question from reviewers is whether in L2 learners the errors were indeed “phonological” and not “phonetic,” or more precisely articulatory. To answer this question, it will take a few well-designed and well-controlled experiments, where articulatory effects are isolated. That would be my next step.

Table 26 Process for T1-T4 error in the four L2 speakers' speech

MH. HL	Base form
M_ HL	Tone Absorption
ML. HL	Tone Dissimilation

should not be viewed as a coincidence. I propose the following process to account for the errors and the corresponding substitutes in these two combinations in IL of the four advanced speakers. The difference between the two processes, one in the IL and one in L1, is highlighted in bold in Table 26 below.

The learners hear the lowering of the H in MH- in native speakers' speech, and then constructed and acquired it as dissimilation rather than interpolation. Because all other three coarticulation rules are assimilatory in nature, in which a tone becomes more like the adjacent tone, in the fourth rule, the lowering of H in MH when it is next to a following H misleads the learners to interpret it as a rule of a completely different kind.

This proposal receives support from the substitute forms in the L2 speech. As shown in Table 26, the attested output of the dissimilation is an ML tone⁷ and not an MM, though both are possible dissimilation outputs. One phonological difference between ML and MM in the Mandarin tonal system is that the change from MH to ML is tonemic while from MH to MM is not. Shen (1992) proposes three diagnostics to distinguish tonal coarticulation from tone sandhi in Mandarin, among which two are related to our discussion here: first, only assimilation is considered coarticulation but tone sandhi may be both assimilatory and dissimilatory, and second, tone sandhi may effect tonemic change while tonal coarticulation involves only allotonic variations. Shen's criteria have received many challenges from scholars including M. Chen (2000), and we have seen above that the fourth rule of tonal coarticulation is not a straightforward assimilation process. In fact, M. Chen (2000) argues that there is no essential difference between the so-called tone sandhi and tonal coarticulation. However, Shen's two points shed light on the IL tonal system, in which a distinction may exist (for other reasons) and therefore explains why the output of the tone dissimilation rule found in mispronounced T2-T1 and T2-T4 sequences is ML and not MM. I argue that the two distinctions may not hold for tonal systems in general or for the Mandarin native speaker's system, but it reflects a distinction between two types of tonal rules in the advanced speakers IL system (Table 27).

If this distinction exists, then we can anticipate that when they hear the sequence "MH-HL," they will process it as a Type 1 rule, because it is dissimilation, and output a tonemic form ML. In fact, ML is the only possible output in the Mandarin system that is both dissimilatory in nature and tonemic. Of course, we do not have to call this distinction as sandhi versus coarticulation, we can call it "Rule Type A" and "Rule Type B." Future experiments will help clarify the nature of the distinction of the two types of processes. In the current study, evidence for this distinction comes from

⁷The two transcribers of tonal data in this study checked many times that the error form that they heard for the target T2 was definitely a low falling, not a MM. The quality of clips does not allow a phonetic study to further confirm the perception, but a future study may continue in this direction.

Table 27 Two types of rules in IL of the four advanced speakers

	Type 1 (Sandhi)	Type 2 (Coarticulation)
Processes allowed	Both assimilation and dissimilation	Only assimilation
Output	Tonemic change	Allotonic variants

the four speakers' acquisition of the other rules. The four subjects have successfully acquired the T3 sandhi rules, the *bu* sandhi rule, the *yi* sandhi rule and some of them also showed competence in using the T2 sandhi rules. All of these sandhi rules are tonemic and include both assimilatory dissimilatory types. They also show little difficulty with the first three coarticulation rules, all which are assimilatory and allotonic. Moreover, they handle the tone-intonation interaction very well and with remarkable fluency. So it is reasonable to suspect the prevalent errors found in T2-T1 and T2-T4 are not due to pure phonetically motivated reasons.

4.3 Constraint Re-Ranking in L2 Acquisition

Lastly, I would like to add some points to the issue of constraint ranking in the OT framework as the explanation of L2 acquisition, the framework employed in Zhang (2010, 2013, 2018). The OT framework is a one-step input–output declarative system, in which markedness constraints and faithfulness constraints interact to select the most optimal candidate as the attested form. Applying the model to SLA, the nature of variability and instability of the IL is captured as the ranking, re-ranking or even the absence of ranking in different conditions (Hancin-Bhatt 2008). Ideally, a complete OT analysis of the L1 system is helpful when studying the rankings of L2 IL, serving as the point of reference for the latter but that does not exist in most cases, including the tonal system of Mandarin. So as Zhang (2013) admits, the rankings/re-rankings analyses only “deal with specific inputs and employ a small amount of the related constraints to illustrate some features of the current inter-language grammars” (Zhang 2013, pp. 186). For example, Zhang (2013) explains the high error rate combinations and their substitute forms using the ranking of a few constraints.

Table 28 only explains why for these L2 learners some (or a lot of) T1-T2 targets were pronounced as T1-T3. We know that the same group also pronounced some (or

Table 28 Tableau for English and Korean speakers' choice of T1-T3 for input T1-T2 (Zhang 2013: 190)

T1-T2	*Rise-F	Id-T
T1-T3		*
T2-T2	*! W	L

a lot of) T1-T2 targets as T1-T2. And the ranking *Rise-F > Id-T is not necessarily working for other tone combinations such as T4-T2 or T3-T2. It is not necessarily true that the existence of correct T1-T2 or T3-T2 output means that Id-T ranks higher in these other contexts. Other higher ranked constraints may be the reason why T4-T2 or T3-T2 are still chosen as the optimal output even if they violate *Rise-F. A complete OT analysis would have to yield all the correct output and none of the incorrect output for all the single tones and tone combinations. Knowing such limitations, I would cautiously propose the following ranking for the large number of errors found in T4-T4 targets. Among all the identical two-tone combinations (T1-T1, T2-T2, and T4-T4), OCP effect is verified only in T4-T4 combination in the current study, which means a specific form of OCP *HL is ranked higher than the faithful constraint Id-T that requires the input and output tones to be the same. The general OCP that punishes all identical two-tone combination is lower, and so are the subtypes OCP (L), OCP (LH) and OCP (H), because T1-T1 and T2-T2 targets surface as T1-T2 and T2-T2 (in the vast majority cases) and the T3 sandhi is acquired by all four speakers. This ranking is different from the constraint ranking/re-ranking proposed in Zhang (2013), where OCP (L) is promoted to the top of the ranking. The current study confirms that subtypes of OCP constraints (in the L2 learners of Chinese whose native language is English) may go through a re-ranking process separate from the general constraint. The question remains whether OCP (HL) constraint is demoted first and then moves up in the final stage of L2 acquisition process.

The proposed “coarticulation rule configuration” explanation for the high error rates found for T2-T4 and T2-T1 combinations in this paper is formulated in the rule-based framework. Constraints (such as OCP) were proposed in pre-OT phonological theories and were used as explanations for many phonological processes. However in OT, at least in the strict versions of OT, constraints replace all rules. It is possible to reformulate the “coarticulation rule configuration” hypothesis using the OT constraint framework. But this will require the ranking of relevant constraints to account for all the tone sandhi phenomena as well as coarticulation effects shown in Table 22, a research that has not been done in L1 Mandarin phonological studies. Yin (2012) attempts to reach one coherent ranking to account for both T3 sandhi and what he calls T4 sandhi, which equals to the first coarticulation rule in Table 24. Nonetheless, even without such a coherent OT analysis, the OT framework predicts that the difference between the L2 IL discovered in this paper for the very advanced learners and the Mandarin L1 system is possibly that in the L2 IL, there is a higher ranked markedness constraint, which favors the substitute output ML over the L1 output MH⁻ in T2-T1 and T2-T4 combinations. Coming up with such a comprehensive OT analysis is worth future studies.

To the teachers of Chinese as a foreign language, this study brings them one suggestion. To help the advanced-level learners, contrast practice involving pairs of “T2-T1/T3-T1” and “T2-T4/T3-T4” will help them reduce the errors. The key is to increase the awareness of the T2 end point in the T2-T1 and T2-T4 combinations. Fossilized errors in high frequency words that were acquired earlier require special attention.

In conclusion, this paper investigates the errors and substitutions forms of the tone production in four very advanced learners' spontaneous connected speech in Chinese. It is found that the overall ranking of difficultness level of the single tones is T2 >> *T4 >> *T1 >> *T3, therefore supporting the TMC (*FT3 >> *T2 >> *T4 >> *T1 >> *HT3) proposed in previous studies. T1 and T2 are performed better at word-final positions while T4 is performed similarly at either word-initial or word-final positions. Such TPC effects are different from the theoretical prediction. A close examination of two-tone combinations reveals that the difference is due to the high error rates found in T2-T1 and T2-T4 combinations. OCP effect is found only in T4-T4 combination and not in T1-T1 combination. An anti-OCP effect is found for T2-T2. The different error rates found in different identical tone combination sequences suggest that the subtype OCP (HL) is ranked higher than the generic one and the other subtypes. The high error rates of T2-T1 and T2-T4 are explained as a rule configuration in the IL where the L1 coarticulation rule "MH → MH-/___Hx" is processed in the learners' phonological system as a tonemic (sandhi) rule. Future study is needed to verify whether the discoveries of this paper are applicable to other very advanced learners of Mandarin. Learners whose native language is not English should be included and longitudinal studies are very much needed to trace the changes. In L1 Chinese phonological study, a complete OT analysis of tone sandhi and coarticulation phenomena will help the L2 researchers pin down exactly the constraint re-rankings that need to take place before the very advanced speakers complete the tonal acquisition.

References

- ACTFL. (2012). *ACTFL performance descriptions for language learners*. ACTFL.
- Bao, Z. (1999). *The structure of tone*. Oxford University Press.
- Chao, Y.-R. (1930). A system of tone-letters. *Le Maître Phonétique*, 45, 24–27.
- Chao, Y.-R. (1968). *A Grammar of Spoken Chinese*. University of California Press.
- Chen, M. (2000). *Tone sandhi: patterns across Chinese dialects*. Cambridge University Press.
- Chen, Q. H. (1997). Toward a sequential approach for tonal error analysis. *Journal of the Chinese Language Teachers Association*, 32, 21–39.
- Chen, Q. H. (2000). *Analysis of Mandarin tonal errors in connected speech by English-speaking American adult learners*. [Doctoral Dissertation]. Brigham Young University.
- Chaudron, C. (2003). Data collection in SLA research. In C. Doughty and M. Long (Eds.) *The handbook of second language acquisition* (pp. 762–828). Blackwell.
- Elliot, C.E. (1991). The relationship between the perception and production of Mandarin tones: An exploratory study. *University of Hawai'i Working Papers in ESL*, 10 (2), 177–204.
- Goldsmith, J. (1976). An overview of Autosegmental Phonology. *Linguistic Analysis*, 2, 23–68.
- Guo, J. -F. (1993). *Hanyu shengdiao yudiao chanyao yu tansuo* [Elucidation and exploration of tone and intonation in Chinese]. Beijing Language Institute.
- Guo, L., & Tao, L. (2008). Tone production in Mandarin Chinese by American students: A case study. *Proceedings of the 20th north american conference on chinese Linguistics* (pp 123–138). Ohio State University Press.
- Hanscin-Bhatt, B. (2008). Second language phonology in optimality theory. In J. Edwards & J. Zampini (Eds.), *Phonology and second language acquisition* (pp. 117–146). John Benjamins.

- Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, 40, 269–279.
- Hyman, L. M., & VanBik, K. (2004). Directional rule application and output problems in Hakha Lai Tone. *Language and Linguistics*, 5(4), 821–861.
- Kratochvil, P. (1987). The Case of the third tone. In *The Wang Li memorial volumes* (pp. 253–276). Joing Publishing Co.
- Leather, J. (1990). Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers. In J. Leather & A. James (Eds.), *New Sounds 90: Proceedings of the Amsterdam Symposium on the Acquisition of Second Language Speech* (pp. 305–341). University of Amsterdam.
- McCarthy, J., & Prince, A. (1993). *Prosodic morphology*. Ms., University of Massachusetts, Amherst and Brandeis University.
- Miracle, C. (1989). Tone production of American students of Chinese: A preliminary acoustic study. *Journal of the Chinese Teachers Association*, 24, 49–65.
- Ohala, J. (1978). Production of tone. In V. A. Fromkin (Ed.), *Tone: A linguistic survey* (pp. 3–39). Academic Press.
- Prince, A., & Smolenski, P. (1993). *Optimality theory: constraint interaction in Generative Grammar*. Ms., Rutgers University and University of Colorado.
- Shen, X. S. (1989). Toward a register approach in teaching Mandarin tones. *Journal of Chinese Language Teachers Association*, 24, 27–47.
- Shen, X. N. (1990). Tonal coarticulation in Mandarin. *Journal of Phonetics*, 18, 281–295.
- Shen, X. N. (1992). On tone sandhi and tonal coarticulation. *Acta Linguistica Hafniensia*, 24, 131–152.
- Shih, C. (1988). *Tone and intonation in Mandarin*. Ms., Cornell University and AT&T Laboratories.
- Shih, C. (1991). Pitch variation across word boundary. Paper presented at the *Third North America Conference on Chinese Linguistics*. Cornell University.
- Sun, S. (1998). *The development of a lexical tone phonology in American adult learners of Standard Mandarin Chinese*. University of Hawai'i Press.
- Van de Weijer, V., & Sloos. (2014). The four tones of Mandarin Chinese: Representation and acquisition. *Linguistics in the Netherlands*, 2014, 180–191.
- Wang, X. C. (2006). Perception of L2 tones: L1 lexical tone experience may not help. *Proceedings of Speech Prosody*. Dresden, Germany.
- Wang, Y., Spence, M., Jongman, A., & Sereno, J. (1999). Training American listeners to perceive mandarin tones. *Journal of the Acoustical Society of America*, 106, 3649–3658.
- White, C. (1981). Tonal pronunciation errors and interference from English intonation. *Journal of the Chinese Language Teachers Association*, 16(2), 27–56.
- Wu, Z. J. (1982). Putonghua yuju zhong de shengdiao bianhua. *Zhongguo Yuwen*, 439–450.
- Wu, Z. J. (1985). Putonghua sanzizu biandiao guilv. *Zhongguo Yuyan Xuebao*, 2, 70–92.
- Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25, 61–83.
- Xu, Y. (2001). Fundamental frequency peak delay in Mandarin. *Phonetica*, 58, 26–52.
- Yang, C. (2011). *The acquisition of Mandarin prosody by American learners of Chinese as a foreign language (CFL)* [Doctoral Dissertation]. Ohio State University.
- Yang, B. (2015). *Perception and production of Mandarin tones by native speakers and L2 learners*. Springer.
- Yang, C. (2016). *The acquisition of L2 Mandarin prosody*. John Benjamins.
- Yin, H. (2012). A unified account of Mandarin tone 3 and tone 4 sandhi. In *Proceedings of the Third international symposium on tonal aspects of languages*. Nanjing University.
- Yip, M. (1980). *The tonal phonology of Chinese*. [Doctoral Dissertation]. MIT.
- Yip, M. (2002). *Tone*. Cambridge University Press.
- Zee, E. (1980). A spectrographic investigation of Mandarin tone sandhi. *UCLA Working Papers in Phonetics*, 49, 98–116.
- Zhang, J. (2004). Contour tone licensing and contour tone representation. *Language and Linguistics*, 5(4), 925–968.

- Zhang, H. (2010). Phonological universals and tone acquisition. *Journal of Chinese Language teachers association*, 45(1), 39–65.
- Zhang, H. (2013). *The second language acquisition of Mandarin Chinese tones by English, Japanese and Korean speakers*. [Doctoral Dissertation]. University of North Carolina-Chapel Hill.
- Zhang, H. (2018). *Second language acquisition of Mandarin Chinese tones—Beyond first language transfer*. Brill.

Categorical Perception of Mandarin Tones by Native and Second Language Speakers



Chunsheng Yang

Abstract Previous studies on categorical perception of tones mainly focus on native speakers and naïve second language (L2) listeners. Attempting to fill in this gap, this study examined the categorical perception of Mandarin tones by both native and L2 speakers along three tone continua in Mandarin Chinese. Both discrimination and identification tasks were employed in the study. The results of the discrimination task showed that the L2 listeners mainly relied on psychoacoustic cues in tone pair discrimination, while native listeners mainly relied on their phonological knowledge. As a result, the non-linguistic tone processing in discrimination tasks would not enable the L2 learners to normalize speech, namely learning to de-emphasize within-category differences and to focus more on between-category differences, hence building a relatively less stable L2 tone system, as well as the difficulty in acquiring tone categories. The results of the identification tasks confirmed the existence of the T2–T3 and T1–T3 and T4–T3 (only to some extent) continua in Mandarin Chinese for both native and L2 listeners, and the potential confusion between Tone 3 and the other tones in Mandarin Chinese seems to explain the difficulty in acquiring this tone for both native and L2 speakers.

Keywords Categorical perception · Mandarin tone · Identification and discrimination · Psychoacoustic cues

1 Introduction

Speech sounds vary across speakers and even within the same speaker. Such variation usually does not pose communication difficulty for native speakers, due to the categorical perception of speech sounds. That is to say, native listeners can filter out within-category differences and learn to pay attention to between-category differences. In this sense, speech perception in one's native language is more discrete than continuous. When it comes to second language (L2) speech perception, it can

C. Yang (✉)
University of Connecticut, Storrs, CT 06269, USA
e-mail: chunsheng.yang@uconn.edu

be discrete or continuous, depending on whether L2 listeners have established categories for the L2 sounds and on the similarity between the native and L2 sounds (Flege 1995). Another factor shaping both native and L2 speech perceptions is whether the perception task taps into listeners' phonological system. In a discrimination task, a listener may not be able to resort to his/her phonological knowledge, due to the limited time of online processing, and mainly rely on the psychoacoustic cues of sounds in discrimination. In the identification task, however, listeners have to compare the continuous acoustical signal with the discrete phonological system in their brain prior to the identification of the sound.

Studies on categorical perception have mainly focused on segments, such as earlier studies on the perception of /b/, /d/, and /g/ that varied along a formant transition continuum by Liberman et al. (1957) and the perception of the voiced–voiceless distinction of the utterance-initial stops along the VOT continuum by Abramson and Lisker (1970). Categorical perception is not what infants are born with, because infants can discriminate the phonetic contrasts of all languages in the first 3 months of their life; however, afterward, L1 exposure starts to shape the infants' speech perception and their perceptual capability starts to decline in L2 speech perception and increase in their native language perception (Kuhl 2004), cited from Peng et al. (2010). Thus, categorical perception can be said to be a normalization process for one to learn to tease apart the unneeded/irrelevant cues in speech perception.

Besides segments, tones or pitch contours may also be categorically perceived. As shown in Wang (1976), there is a linguistic boundary of tones for native Chinese listeners but a psychoacoustic boundary for naïve American English-speaking listeners. Previous studies have investigated the effect of L1 tonal status (i.e., a tone language versus a non-tone language) and tone inventory size on pitch perception. For example, Bent (2005), Halle et al. (2004), Lee et al. (1996), Xu et al. (2006), Peng et al. (2010), Qin et al. (2019), Shen and Froud (2016), Chuang et al. (1972), and Zheng et al. (2012) confirmed the findings of Wang (1976), namely the (quasi-)categorical perception of tones by tone language listeners, and the psychoacoustic perception of tones by non-tone language listeners. In terms of the effect of tone inventory size, Zheng et al. (2012) found that the Cantonese (with six lexical tones) listeners engaged phonological processing in order to discriminate speech stimuli more efficiently than Mandarin (with four lexical tones) listeners, in that Cantonese listeners are required to make finer distinctions in perception of pitch height and slope (contour) than Mandarin listeners in order to discriminate the dense tone system of Cantonese.

While previous studies on tone perception involved both tone and non-tone language listeners, most of these studies, except Shen and Froud (2016) and Qin et al. (2019), only investigated naïve non-tone language listeners, namely the listeners who can not speak the tone language. Hence, there is a need to include more non-tone language listeners who are able to speak the target tone language as an L2 in order to examine the interaction between the native language and L2 interlanguage in the categorical perception of tones. To that end, this study examines the categorical perception of Mandarin tones by native and L2 speakers by focusing on three tone continua, T1–T3, T2–T3, and T4–T3.

2 Research Questions

Mandarin Chinese has four lexical tones: the high level tone (Tone 1), the rising tone (Tone 2), the low-dipping tone (Tone 3), and the falling tone (Tone 4) (Chao 1930). Of the four tones, Tone 3 poses the biggest problem for both L1 and L2 speakers, due to its potential confusion with Tone 2 (Chuang et al. 1972; Shen and Lin 1991; Moore and Jongman 1997; Hao 2012). While Tone 3 is a low-dipping tone in isolation, it phonetically surfaces as a low or low-falling tone when preceding a non-Tone 3, namely the half-third sandhi (Zhang and Lai 2010). Meanwhile, Tone 3 becomes a rising tone (i.e., Tone 2) when it occurs before another Tone 3, namely the third-tone sandhi.

The similarity of Tone 2 and Tone 3 in terms of pitch contour is often claimed to contribute to the learning and perceptual difficulty of Tone 3 (Shen and Lin 1991; Shen et al. 1993; Moore and Jongman 1997). Shen and Lin (1991) and Moore and Jongman (1997) found that the perception of Tone 2 and Tone 3 is cued by the timing of the F0 turning point (i.e., earlier for Tone 2), namely the point in time where the pitch contour changes from falling to rising, and the degree of the initial fall (smaller for Tone 2). It can be said that there is a Tone 2–Tone 3 (i.e., the low-dipping allotone of Tone 3) continuum in both production and perception.

In terms of tone contour, the low Tone 3, the allotone which occurs before a non-Tone 3, and Tone 4 are also similar. Garding et al. (1986) found that in comparable sentences, Tone 3 is characterized with a low pitch level throughout the second half of the vowel and Tone 4 with a gradual fall over the main part of the vocalic segment. In addition, they found that the perception of Tone 4 and Tone 3 in the context has a clear reference, namely the identification of Tone 4 was favoured by an introductory rising or level part, and for Tone 3 an introductory fall seemed to be important. As a result, Tone 3 may form a continuum with Tone 4 to some extent. In addition, since Tone 3 when occurring before a non-Tone 3 differs from Tone 1 mainly in pitch register, and Tone 3 forms a continuum with Tone 1 as well.

In this study, we examine the perception of Mandarin tones by native and L2 speakers along these three tone continua and attempt to answer the following questions:

- (1) Can native and L2 speakers discriminate these within-category and between-category tone pairs?
- (2) Can native and L2 speakers identify the tone stimuli in each tone continua?
- (3) What are the differences in tone discrimination and identification of tone stimuli in the three tone continua between the native and L2 speakers?

It is expected that the inclusion of L2 speakers in the tone perception will shed new light on the categorical perception of tones, and further our understanding of the learning difficulty of Tone 3 for both native and L2 speakers.

3 Methodology

3.1 Subjects

Eighteen intermediate-low to intermediate-high L2 learners of Mandarin Chinese (7 males; 11 females; mean/sd.: 20.5/1.4) and ten Northern Mandarin Chinese speakers (4 males; 6 females; mean/sd.: 32.1/7.4) participated in this study. The L2 learners were recruited in a public university in the USA. Most of the L2 Chinese learners started to learn Mandarin Chinese in high schools and some at an even earlier age, and their length of Mandarin learning ranged from 2 to 10 years (mean/sd.: 7.4/4.8). At the time of the experiment, they were taking the third-year Mandarin Chinese course at the same university. Considering the fact that the length of Mandarin learning varied among the L2 participants, these L2 speakers were further classified into two subgroups according to their scores in a cloze test (credit to Professor Boping Yuan at Cambridge University) in data analysis. However, no significant difference was found in both the discrimination and identification tasks between the two subgroups, very likely because this was a perception study. It is expected that the productions of tones between these two L2 subgroups may reveal differences, which has to be investigated in further studies. As a result, the L2 speakers were treated as one group in data analysis. The ten Northern Mandarin Chinese speakers were all from mainland China and were either studying or working at the university. All participants reported no speech or hearing problems and were paid or received course credits for their participation.

3.2 Stimuli

Since both Tone 3 sandhis occur on the first syllable of a disyllabic phrase, disyllabic phrases were used in the speech stimuli. Specifically, three disyllabic phrases with Tone 1, Tone 2, and Tone 4 on the first syllabic position were constructed in the study. The three disyllabic phrases were ying1lan2 瑛兰 “a girl’s name,” mai2mao1 埋猫 “to bury a cat,” and mai4niu2 卖牛 “to sell a cow.” The speech stimuli of these phrases were produced by a female northern Mandarin speaker, who was in her early 30s at the time of recording and was an experienced L2 Mandarin instructor. Drawing on Peng et al. (2010), the tones in the first syllables of the three phrases were resynthesized by applying the pitch-synchronous overlap and add (PSOLA) method (Mouline and Laroche 1995), cited from Peng et al. (2010) through the program Praat (Boersma and Weenink 2016). Before resynthesizing, the duration of the target syllables was adjusted to 500 ms by using the Praat vocal toolkit (Corretge 2019), and the number of stylized pitch points was manually reduced to 3. Five stimuli were resynthesized for each phrase, as shown in Figs. 1, 2, and 3. To construct the Tone 2–Tone 3 continuum, the duration from the second stylized pitch point to the end of the target syllable was divided into four equal intervals and the five stimuli were resynthesized

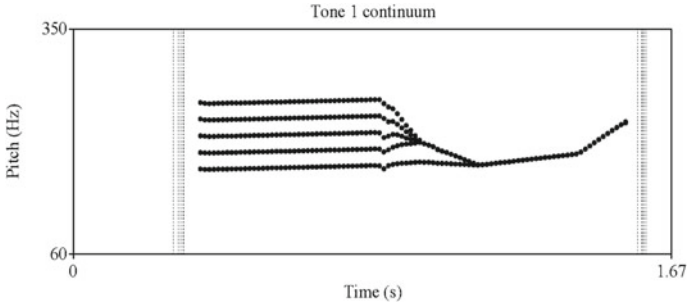


Fig.1 Tone 1–Tone 3 continuum

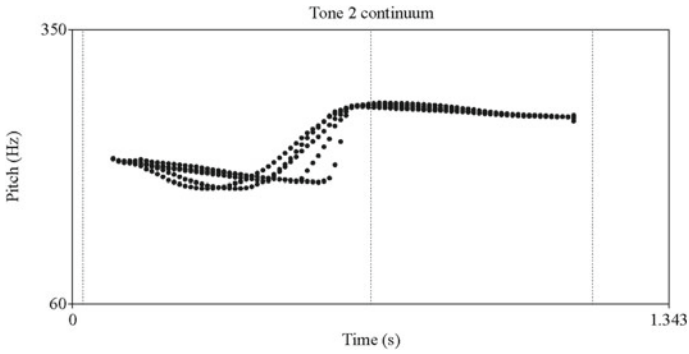


Fig. 2 Tone 2–Tone 3 continuum

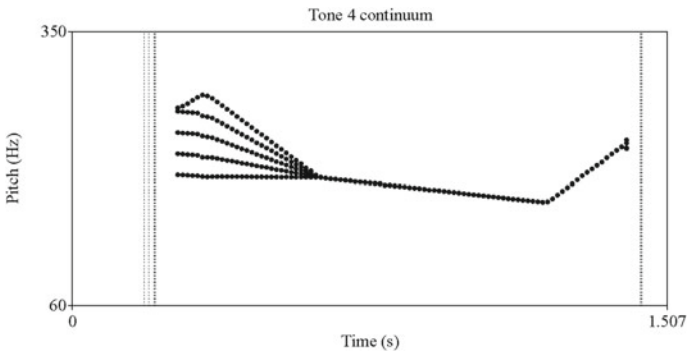


Fig. 3 Tone 4–Tone 3 continuum

by horizontally dragging the second pitch point to different interval points (0.069656 s per step). To construct the Tone 1–Tone 3 continuum, the difference between the 2nd pitch point to the lowest pitch of the phrase was divided into four equal intervals and the five stimuli were resynthesized by dragging the stylized pitch contour to different

interval points (down 22.21 Hz per step). To construct the Tone 4–Tone 3 continuum, the difference between the first and the third pitch point on the target syllable was divided into four intervals and then the five stimuli were resynthesized by dragging the stylized pitch contour to different interval points down (down 22.46 Hz per step).

3.3 Procedures

Both discrimination and identification tasks were used in this study. The two tasks were run with the Multiple Forced Choice (MFC) experiment function on Praat (Boersma and Weenink 2016).

3.3.1 Discrimination Task

The five stimuli in each tone continuum formed 15 pairs. Altogether there were 45 pairs of stimuli. These pairs of stimuli were presented with a 500 ms inter-stimulus interval (ISI), in random order. The two stimuli in each pair were arranged in either of the following orders: (a) 1–1, 1–2, 1–3, 1–4, 1–5, 2–2, 2–3, 2–4, 2–5, 3–3, 3–4, 3–5, 4–4, 4–5, 5–5, or (b) 1–1, 2–1, 3–1, 4–1, 5–1, 2–2, 3–2, 4–2, 5–2, 3–3, 4–3, 5–3, 4–4, 5–4, 5–5. These pairs were repeated six times; hence, there were $45 \times 6 = 270$ stimuli pairs in the discrimination task. Participants were asked to listen to the pairs of stimuli and judge whether the tones on the first syllables in each stimulus were the same or not by clicking the mouse on the computer screen. For every 30 pairs of stimuli, the participants could choose to take a break. The order of the stimulus pair presentation was counterbalanced across participants.

3.3.2 Identification Task

Each of the 15 stimuli in the three tone continua was repeated 6 times in the identification task, hence 90 stimuli. The stimuli were presented in random order. After hearing a stimulus, the participants were asked to decide what is the tone on the first syllable in the stimulus by clicking on “Tone 1,” “Tone 2,” “Tone 3,” or “Tone 4” on the computer screen with the mouse. For every 30 stimuli, the participants could choose to take a break.

Prior to both the discrimination and identification tasks, there was a brief practice session, with the stimuli recorded by another Chinese native speaker. All the setups of the practice sessions were exactly the same as the actual tasks, except for fewer stimuli and fewer repetitions in the practice sessions.

4 Results

4.1 Discrimination Task

For the discrimination task, we used the participants' judgment ("same" or "different") as the response variable to fit a logistic regression model with the group variable as the predictor. This was done on the Tone 2, Tone 4, and Tone 1 continua, respectively.

The confidence intervals for the across-stimuli difference of the Tone 2 continuum in percentages of correct discrimination for the ten tone combinations adjusted using FWER of 0.05 and using Tukey's HSD method are shown in Table 1.

We see that there is a significant difference in the discrimination between L2 and native speakers for all the stimulus pairs in the Tone 2 continuum except for the stimuli pairs 2_4, 3_4, and 4_5, at the 95% level. The estimated proportions along with their respective Wald 95% intervals are shown in Fig. 4.

For Tone 4 and Tone 1 continua, we repeated the same analysis to obtain Tables 2 and 3, respectively.

For the Tone 4 continuum as shown in Table 2, we see that there is a significant difference in the discrimination of all stimuli pairs except the pairs 2_4, 3_5, and 4_5. In Table 3, we see that there is a significant difference in the discrimination of stimuli pairs in the Tone 1 continuum except for pairs 1_4, 1_5, 2_4, 2_5, 3_4, and 3_5.

Figures 5 and 6 show the estimated percentages of discrimination in Tone 4 and Tone 1 continua, respectively, with their 95% Wald intervals.

Table 1 Tukey's HSD-adjusted confidence intervals for the discrimination of stimuli pairs in the Tone 2 continuum

Tone combinations	Logit ⁻¹ (difference)	Logit ⁻¹ (lower adj. limit)	Logit ⁻¹ (upper adj. limit)	Adjusted <i>p</i> -value
1_2	0.1704	0.0615	0.2792	0.0023
1_3	0.1611	0.054	0.2682	0.0034
1_4	-0.2241	-0.3783	-0.0699	0.0047
1_5	-0.5722	-0.7043	-0.4402	0
2_3	0.2222	0.1156	0.3288	1.00E-04
2_4	-0.137	-0.2876	0.0136	0.0742
2_5	-0.4926	-0.631	-0.3542	0
3_4	-0.1259	-0.2691	0.0173	0.0843
3_5	-0.4741	-0.6147	-0.3335	0
4_5	-0.0815	-0.2121	0.0491	0.2199

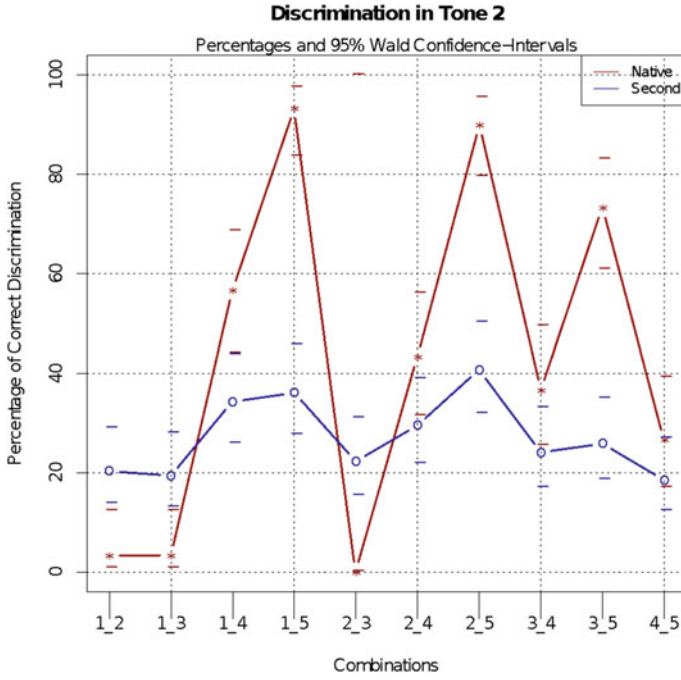


Fig. 4 Percentages of discrimination of stimulus pairs of the Tone 2 continuum by the native and L2 speakers

Table 2 Tukey’s HSD-adjusted confidence intervals for difference in the Tone 4 continuum

Tone combinations	Logit ⁻¹ (difference)	Logit ⁻¹ (lower adj. limit)	Logit ⁻¹ (upper adj. limit)	Adjusted <i>p</i> -value
1_2	0.2593	0.1469	0.3716	0
1_3	0.4759	0.3432	0.6086	0
1_4	0.2148	0.0781	0.3515	0.0023
1_5	-0.1333	-0.2349	-0.0318	0.0104
2_3	0.337	0.2085	0.4655	0
2_4	0.0537	-0.1027	0.2101	0.4987
2_5	-0.1463	-0.2616	-0.031	0.0132
3_4	0.3204	0.1771	0.4636	0
3_5	-0.0963	-0.2361	0.0435	0.1758
4_5	-0.0944	-0.2537	0.0648	0.2433

Table 3 Tukey’s HSD-adjusted confidence intervals for difference in Tone 1

Tone combinations	Logit ⁻¹ (difference)	Logit ⁻¹ (lower adj. limit)	Logit ⁻¹ (upper adj. limit)	Adjusted <i>p</i> -value
1_2	0.5648	0.4377	0.6919	0
1_3	0.4926	0.3542	0.631	0
1_4	0.1093	-0.0167	0.2352	0.0886
1_5	-0.0074	-0.0898	0.0749	0.8593
2_3	0.4185	0.2782	0.5589	0
2_4	0.0389	-0.0909	0.1686	0.5548
2_5	-0.0259	-0.1142	0.0624	0.5629
3_4	0.0648	-0.0802	0.2098	0.3788
3_5	-0.0056	-0.1155	0.1044	0.9206
4_5	0.5389	0.4068	0.671	0

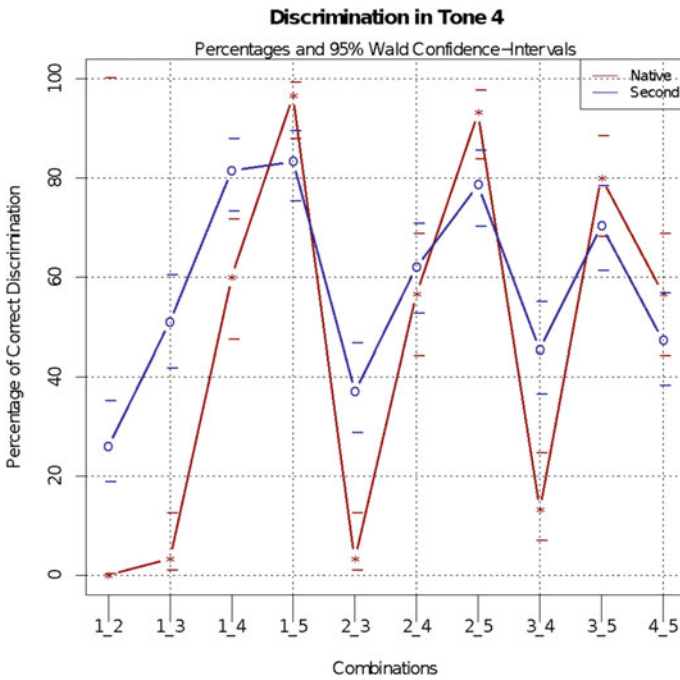


Fig. 5 Percentage of discrimination of stimulus pairs in the Tone 4 continuum by the native and L2 speakers

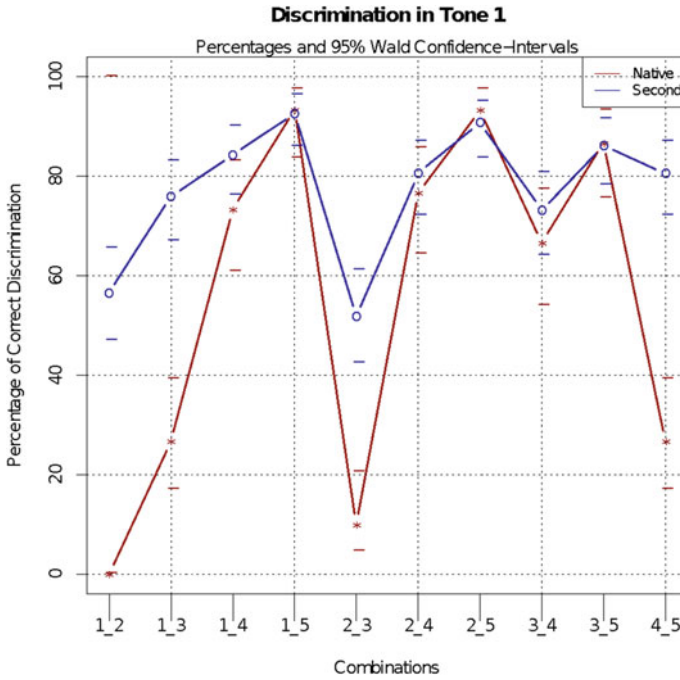


Fig. 6 Percentage of the discrimination of stimulus pairs in the Tone 1 continuum by the native and L2 speakers

4.2 Identification Task

For the identification task, we fitted the multiple logistic regression model with the identification of the target tone (Tone 1, Tone 2, and Tone 4) and Tone 3 as the response variable, and group as the categorical variable. For the Tone 2 continuum, the identification of each stimulus as Tone 2 and Tone 3 is shown in Fig. 7.

The estimated intercepts and coefficients for the identification of each stimulus as Tone 2 and Tone 3, together with the p-values and Cohen’s d, are shown in Table 4.

We see that at the 95% level of significance there is no significant difference in identification for T2 and T3 between the native and L2 speakers. Considering the small sample size in this study, we also calculated the effect sizes, namely Cohen’s d, for the difference in the identification of each tone (last column of Table 4). While overall the effect sizes are small, the identification of Stimulus 2 and Stimulus 4 as Tone 2, and the identification of Stimuli 4 and 5 as Tone 3 in the Tone 2 continuum, by the native and the L2 groups, differ by 0.3 standard deviations or more, although not statistically significant. While at Stimulus 4, it may not be that significant, in that the identification by either group is around the chance level, showing that this may be a boundary between Tone 2 and Tone 3 identifications. For Stimulus 5, the native

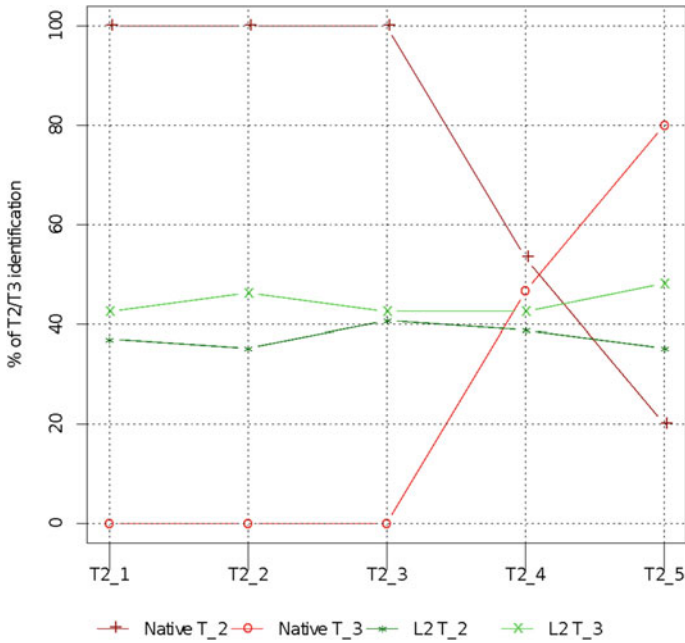


Fig. 7 Identification of the stimuli as Tone 2 and Tone 3 in the Tone 2 continuum

Table 4 Identification of the stimuli in the Tone 2 continuum

Tone 2 continua	(Intercept)	Coefficient	<i>p</i> -values	Cohen's <i>d</i>
linmai2_1 Tone 2	0.597864875	11.53023075	0.441633917	0.14682794
linmai2_1 Tone 3	0.737612912	-2.988511206	0.504686126	-0.0117466
linmai2_2 Tone 2	0.642052811	9.559252652	0.374875947	0.31896653
linmai2_2 Tone 3	0.916290918	-2.913694069	0.513405234	-0.0336083
linmai2_3 Tone 2	0.893911229	10.54967926	0.424981892	0.18916464
linmai2_3 Tone 3	0.938369257	-3.180999106	0.507049695	-0.0176719
linmai2_4 Tone 2	0.742152209	8.792311071	0.382457136	0.29903378
linmai2_4 Tone 3	0.833264046	8.56766021	0.385375609	0.29139241
linmai2_5 Tone 2	0.74738112	7.878055571	0.398026484	0.25845864
linmai2_5 Tone 3	1.06119144	8.950705348	0.384504465	0.2936715

speakers identified more Tone 3 than the L2 speakers, implying that the last stimulus in the Tone 2 continua has become Tone 3 for native speakers. As shown in Fig. 7, there is more confusion in the identification of the stimuli in the Tone 2 continuum for the L2 speakers. For the native speakers, there is clear categorical perception and the boundary for the identification of Tone 2 and Tone 3 is around Stimulus 4.

Table 5 Identification of the stimuli in the Tone 4 continuum

Tone 4 continua	(Intercept)	Coefficient	<i>p</i> -values	Cohen's <i>d</i>
linmai4_1 Tone 3	-0.606157753	-1.118019729	0.503729582	-0.0093488
linmai4_1 Tone 4	1.213159627	9.867864756	0.416012575	0.21210496
linmai4_2 Tone 3	-1.098613394	-0.78763349	0.502476079	-0.0062067
linmai4_2 Tone 4	0.818319155	10.24223286	0.411994289	0.2224179
linmai4_3 Tone 3	-0.079740622	8.768778398	0.390231271	0.27871631
linmai4_3 Tone 4	0.802717662	9.272457341	0.384091794	0.29475167
linmai4_4 Tone 3	-0.051298506	8.980522107	0.371906617	0.32680783
linmai4_4 Tone 4	-0.287388133	9.909655236	0.359188678	0.36062827
linmai4_5 Tone 3	-0.095309281	1.481598118	0.010179531	2.31966398
linmai4_5 Tone 4	-0.606135422	1.522421806	0.013896039	2.20020888

The same procedures were repeated for the Tone 4 and Tone 1 continua. Table 5 shows the estimated intercepts and coefficients for the identification of each stimulus as Tone 4 and Tone 3 in the Tone 4 continuum, together with the *p*-values and Cohen's *d*.

We can see that at the 95% level of significance there is only significant difference in identification (with large effect sizes) for Stimulus 5 in the Tone 4 continuum by the native and L2 speakers, showing that the continuity of Tone 4 and Tone 3 and that the last stimulus tended to be identified as Tone 3 by the native speakers. While not statistically significant, the identifications of Stimulus 4 as both Tone 4 and Tone 3 by the two groups have small effect sizes as well, implying that the trend of Tone 4 continuum being identified as Tone 3 starting from Stimulus 4. Also as shown in Fig. 8, the boundary of Tone 4 and Tone 3 identification by the native speakers is between Stimulus 3 and Stimulus 4.

Table 6 shows the estimated intercepts and coefficients for the identification of each stimulus as Tone 1 and Tone 3 in the Tone 1 continuum, together with the *p*-values and Cohen's *d*.

Again, we see that at the 95% level of significance there is no significant difference in identification for Tone 1 and Tone 3 between native and L2 speakers, although the identification of Stimulus 4 as Tone 3 and the identifications of the Stimulus 5 as both Tone 1 and Tone 3 have an effect size of 0.3 or above, suggesting that toward the end of the Tone 1 continuum the stimulus tended to be identified as Tone 3 by the native speakers, but not by the L2 speakers as shown in Fig. 9.

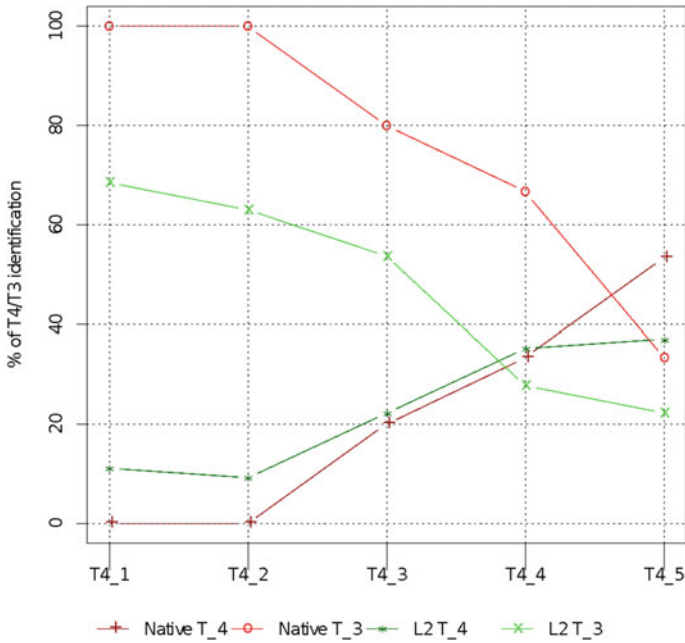


Fig. 8 Identification of the stimuli as Tone 4 and Tone 3 in the Tone 4 continuum

Table 6 Identification of the stimuli in the Tone 1 continuum

Tone 1 continua	(Intercept)	Coefficient	p-values	Cohen's d
linying1_1 Tone 1	1.098786328	10.15077606	0.420523417	0.20055463
linying1_1 Tone 3	-0.693243652	-0.747735242	0.50257846	-0.0064633
linying1_2 Tone 1	0.931578564	10.05361241	0.410326998	0.2267039
linying1_2 Tone 3	-0.485479412	-1.227005693	0.50431537	-0.0108172
linying1_3 Tone 1	0.794796772	9.689710927	0.393143155	0.27113616
linying1_3 Tone 3	-0.441866517	8.286702201	0.408335786	0.23182805
linying1_4 Tone 1	0.588013333	7.851167629	0.408720981	0.23083632
linying1_4 Tone 3	-0.222923345	10.53413263	0.378378685	0.30974172
linying1_5 Tone 1	-0.788142652	9.045817211	0.360665891	0.35667946
linying1_5 Tone 3	0.000115765	9.64383834	0.351856652	0.38031271

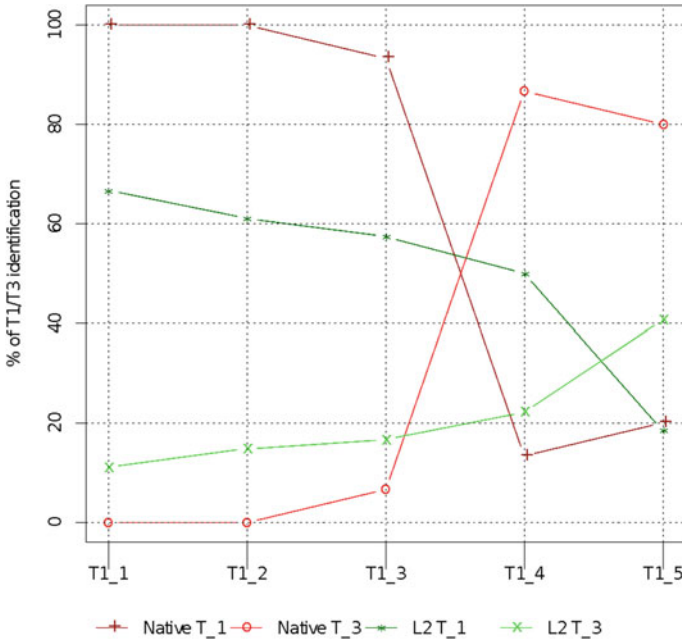


Fig. 9 Identification of the stimuli as Tone 1 and Tone 3 in the Tone 1 continuum

5 Summary and Discussion

5.1 Results of the Discrimination Task

The discrimination of tone pairs in the three tone continua displayed different patterns, especially between the Tone 2 continuum and the Tone 1 and Tone 4 continua. As regards the Tone 2 continuum, both the native and L2 speakers could not differentiate the one- and two-step stimulus pairs (e.g., 3–4 or 4–5). More importantly, the discrimination of all tone stimulus pairs by the L2 learners is below the chance level (50%), even the tone stimulus pairs 1–4 and 1–5, confirming the confusion of Tone 2 and Tone 3 as found in previous studies. By contrast, the native listeners perceived the tone pairs in the Tone 2 continuum more categorically, hence tending to ignore the difference in the one-step and even some two-step pairs.

Similar to the Tone 2 continuum, the two groups could not discriminate the one- and two-step stimulus pairs in the Tone 4 continuum, but could discriminate some two-step and all three- and four-step pairs very well. Meanwhile, the discriminations of all stimulus pairs in the Tone 4 continuum by the L2 speakers were better than those in the Tone 2 continuum, above the chance level for all stimulus pairs except for some one-step pairs. Thus, it can be seen that the individual stimuli in the Tone 4

continuum were more distinctive than those in the Tone 2 continuum, especially for the L2 speakers.

As for the Tone 1 continuum, both the native and L2 listeners could discriminate most of the pairs of two steps or above. Furthermore, the discrimination of stimulus pairs 1–2, 1–3, 2–3, and 4–5 is significantly different between the native and L2 listeners. Interestingly, the native speakers barely discriminated these stimulus pairs, whereas the L2 listeners differentiated them pretty well (all above the chance level and even up to 80%). This result seems to suggest that the stimuli in the Tone 1 continuum were more discrete than continuous, even between the neighboring two stimuli, for the L2 listeners, as opposed to the native listeners, due to their focus on the psychoacoustic cues in their discrimination.

In summary, it seems that the stimuli in the Tone 1 continuum are the most discrete for the L2 listeners, followed by Tone 4 continuum, and the stimuli in the Tone 2 continuum are the most continuous and pose the greatest problem for L2 listeners and even native listeners. Meanwhile, the L2 listener tended to pay more attention to the psychoacoustic cues (the difference in the two neighboring stimuli of one tone continuum), whereas the native listeners tended to ignore such psychoacoustic difference and focus more on the categorical difference. Both L2 and native listeners could discriminate the stimulus pairs of three- and more-step away from each other in the Tone 4 and Tone 1 continua, suggesting that both Tone 4 and Tone 1 and the resynthesized “Tone 3” are more distinct. However, the L2 listeners could not discriminate the stimulus pairs in the Tone 2 continuum, though. Shen and Lin (1991) found that the cue for the Tone 2 and Tone 3 perception is the timing of the F0 turning point, namely early turning point for Tone 2 and delayed turning point for Tone 3. As shown in Fig. 2, all stimuli in the Tone 2 continuum have somewhat a falling F0 contour but differ in the timing of the F0 turning point. The L1 listeners’ linguistic experience seems to have provided them with such perceptual capability to discriminate the stimuli in the Tone 2 continuum, hence their better discrimination in the tone pairs in the Tone 2 continuum. Considering the fact that the T2–T3 continua only involve pitch contour changes and T1–T3 and T4–T3 continua mainly involve pitch height changes, the findings in the discrimination tasks also support previous findings that L2 English listeners had more difficulty in pitch contour perception than pitch height perception (Gandour 1983; Qin et al. 2019).

5.2 Results of the Identification Task

The results of the identification task are very different from those of the discrimination tasks. On the one hand, there was only statistical significance in the identification of the last stimulus in the Tone 4 continuum between the native and L2 listeners. On the other hand, the identification of the last one or two stimuli in all tone continua between the native and L2 listeners had a small effect size, indicating the small difference in the stimulus identification in all three tone continua across groups. Another interesting pattern in the identification of the three tone continua is that the L2 listeners’

identifications of all the tone stimuli in the Tone 2 continuum are below the chance level but native speakers clearly had a categorical perception with the identification boundary of Tone 2 and Tone 3 around stimulus 4, whereas the identification rates of the stimuli in the Tone 1 continua followed similar trend between the native and L2 groups, but the identification boundaries for either group are slightly different. As for the Tone 4 continuum, the native and L2 listeners performed in the identification task similarly, except on the last stimulus, indicating that the differences between Tone 4 and Tone 3 are more discrete than continuous.

In summary, it can be said that L2 listeners had the most difficulty in the identification of stimuli in the Tone 2 continuum and the least difficulty in the identification of the stimuli in the Tone 1 continuum, as compared to native speakers. However, for both L2 and native listener groups, the differences between the stimuli in the Tone 4 continuum are not that categorical indicating that the chance of identification of the misproduced “low” Tone 4 as Tone 3 is relatively low.

5.3 *General Discussion*

To compare the results of the discrimination and identification tasks, we can find that findings regarding the Tone 2 continuum are consistent, in that L2 listeners consistently had difficulty in both discriminating two stimuli and identifying certain stimuli (all below the chance level), whereas the native listeners showed categorical perception in both identification and discrimination of the stimuli in this tone continuum. As for the stimuli in the Tone 1 continuum, the L2 listeners seemed to be able to discriminate all stimulus pairs, while the native speakers could discriminate most of the three- or four-step stimulus pairs. Thus, as compared to the native listeners, L2 listeners’ issue with tone discrimination is not that they could not detect the F0 differences, but their F0 differences were not linked to tone categories. The identifications of the stimuli in the Tone 1 continuum were the most similar between the native and L2 listeners, although the native speakers had the tendency to identify Stimulus 5 as Tone 3. Of the three tone continua, the Tone 4 identification is the least categorical, according with the results in the Tone 4 discrimination. The L2 speakers’ discrimination and identification patterns of the stimuli in the three tone continua show the difference in the discrimination and identification tasks; namely, the discrimination task taps into phonetic or psychoacoustic cues, while the identification task taps into phonological cues. By contrast, native speakers rely on both phonetic and phonological cues in their discrimination and identification tasks.

The native speakers’ identification of the last stimulus in the Tone 1 continuum suggests that the most important cue for Tone 3 perception for the native listeners is the low-pitch contour, whereas the L2 listeners need to be more sensitive to this important cue for Tone 3 perception. However, it is interesting to note that the low-pitch contour of the last stimulus in the Tone 4 continuum does not lead to the overall identification of Tone 3 for both native and L2 listeners. The only difference between the last stimuli in the Tone 1 continuum and the Tone 4 continuum lies in the slight

falling F0 in the last stimulus in the Tone 4 continuum. Therefore, it seems that the most important cue for Tone 3 identification is the low F0 and the falling F0 does not matter that much.

Previous studies on tone perception have shown that naïve non-tone listeners perceived tones psychoacoustically, different from tone listeners' categorical/linguistic perception (Wang 1976; Bent 2005; Halle et al. 2004; Xu et al. 2006; and Zheng et al. 2012). The current study involved L2 learners who have learned Mandarin for two years or longer. The results of the discrimination and identification tasks showed that the learning experience of a tonal L2 influences their tone perception. Although the L2 experience does not seem to affect the discrimination task drastically; namely, the L2 learners mainly process the discrimination of tone pairs psychoacoustically, such as the discrimination of most stimulus pairs in the Tone 1 continuum, and their experience does seem to affect the identification task; this difference is due to the fact that these two tasks tap into different levels: Discrimination tasks mainly tap into the phonetic knowledge/auditory information, while the identification tasks tap into the phonological/linguistic knowledge. However, the L1 listeners' results in the discrimination task showed that their discrimination of tone pairs in different tone continua taps into their phonological or linguistic knowledge as well. The L2 listeners' psychoacoustic processing in the discrimination task showed that they have not internalized the L2 Mandarin phonological system and establish different tone categories in their brain. The L2 speakers' performance in the discrimination task may be related to their proficiency level. Note that the L2 speakers in this study are at the intermediate-low to intermediate-high level and have not achieved the advanced or higher proficiency. Future studies should recruit more advanced L2 speakers to investigate whether and how their categorical perception of Mandarin tones differs from Chinese native speakers and whether their increased and abundant experience with Chinese helps them perceive tones more categorically and native-like.

This study involves three pairs of tones: Tone 2–Tone 3, Tone 1–Tone 3, and Tone 4–Tone 3. Each pair of tones shares similar overall pitch contours, while displaying difference in either the timing of the F0 turning point (T2 vs T3) or the initial F0 (Tone 1 and Tone 4 having rising or level initial F0, whereas Tone 2 and Tone 3 have initial F0 fall). Assuming that the identification of tones merely relies on F0 contours, it would be expected that there would be categorical perception for all three tone continua, at least for L1 listeners. The perceptual results in this study, however, show that, in addition to the overall F0 contour, the initial F0 (i.e., high or low, falling or rising) also influences tone perception. L1 speakers' experience with L1 provides them with the significant cues in tone perception; the inadequate linguistic experience, however, poses a big problem for the L2 speakers/listeners, especially those who are only exposed to the target language in formal classroom setting (like the L2 speakers in this study).

5.4 Pedagogical Implications

The possible confusion of Tone 3 and the other tones in Mandarin Chinese further emphasizes the importance of Tone 3 teaching. As this study and many other studies have shown, Tone 3 should be taught as a low tone when it precedes a non-T3 tone; Tone 3 sandhi, namely Tone 3 becomes a rising tone when it precedes another T3, should also be taught. When Tone 3 occurs at phrase—or sentence—final positions, it often surfaces as atone, unless it is stressed or focused as in the response to what you want to buy, 买马, in which case Tone 3 is surfaces as low-dipping tone. Worth noting is that when Tone 3 is taught as a low tone, it had better been taught in the context, instead of in isolation, in order to avoid any possible complications. That is to say, the production of Tone 3 should be practiced in three contexts, respectively, namely before a non-T3, before a Tone 3, and at phrase-final position. The findings from the discrimination and identification tasks suggest that the identification task is more important than the discrimination tasks in helping L2 learners tease apart the important cues (i.e., the low F0) from the not so important cues (i.e., the falling F0) in Tone 3 (as well as other tones) and establish the categories of different tones. For the L2 listeners/speakers, they need to internalize the L2 tone phonology through abundant linguistic exposure, (re-)learn to neglect the within-category information in L2 speech, and focus on the between-category cues in tone perception. Furthermore, as the chapter by Yingjie Li and Goun Lee in this volume shows, the high variability tone training will be very useful for Tone 3 pronunciation. It is expected that such training will not only train learners to produce the two or three variants of Tone 3 correctly, but also raise their awareness of the contexts in which different variants occur.

6 Conclusion

The results of the identification tasks in this study showed that there exist the T2–T3, T1–T3, and T4–T3 (only to some extent) continua in Mandarin Chinese, especially for L2 listeners. This means that Tone 1, Tone 2, and Tone 4, if produced differently than the correct forms (e.g., when the overall F0 contour for Tone 1 was lowered or the turning point in Tone 2 was delayed), would be likely to be perceived as Tone 3. Considering that the average pitch range of Chinese is 1.5 times that of English (White 1981), the chance of L2 speakers' Tone 1 and even Tone 4 being perceived as Tone 3 may be high, although the context may help avoid such confusion to some extent. More importantly, the possible confusion between Tone 3 and the other tones in Mandarin Chinese seems to help explain the difficulty in acquiring this tone for both L1 and L2 learners. The results of the discrimination tasks showed that the L2 listeners mainly relied on psychoacoustic cues in tone pair discrimination, while L1 listeners mainly relied on their phonological knowledge. As a result, the non-linguistic tone processing in discrimination tasks would not enable the L2 learners to normalize

speech, namely learning to de-emphasize the within-category differences and to focus more on the between-category differences, hence building a relatively less stable L2 tone system, as well as the difficulty in acquiring tone categories. However, the reason why the L2 listeners did not rely on their phonological knowledge of tones does not seem to be merely due to the perceptual task; whether they have such linguistic knowledge of tones in their phonology is another important factor. For example, it seems that the L2 listeners did not acquire the low status of Tone 3 when it occurs before a non-Tone 3, hence their difficulty in the identification task. The findings of this study have important research and pedagogical implications.

References

- Abramson, A. S., & Lisker, L. (1970). Discriminability along the voicing continuum: Cross language tests. In *The Proceedings of the 6th International Congress of Phonetic Sciences* (pp.569–573), Prague.
- Bent, T. (2005). The perception and production of non-native prosodic categories. Unpublished Ph.D. dissertation. Northwestern University.
- Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer [Computer program]. Version 6.0.14, retrieved February 2016 from <https://www.praat.org/>.
- Chao, Y.-R. (1930). A system of tone-letters. *Le Maître Phonétique*, 45, 24–27.
- Chuang, C.-K., Hiki, S., Sone, T., & Nimura, T. (1972). The acoustical features and perceptual cues of the four tones of standard colloquial Chinese. In *Proceedings of the 7th International Congress of Acoustics* (Vol. 3, pp. 297–300). Budapest: Academiai Kiado.
- Corrette, R. (2019). Praat Vocal Toolkit. <https://www.praatvocaltoolkit.com>.
- Flege, J. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 229–273). Timonium, MD: York Press.
- Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics*, 11(2), 149–175.
- Garding, E., Kratochvil, P., Svantesson, J. O., & Zhang, J. (1986). Tone 4 and Tone 3 discrimination in modern Standard Chinese. *Language and Speech*, 29, 281–293.
- Halle, P. A., Chang, Y.-C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, 32, 395–421.
- Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, 40, 269–279.
- Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, 5, 831–843.
- Lee, Y.-S., Vakoč, D. A., & Wurm, L. H. (1996). Tone perception of Cantonese and Mandarin: A cross-linguistic comparison. *Journal of Psycholinguistic Research*, 25, 527–542.
- Lieberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5), 358–368.
- Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America*, 102, 1864–1877.
- Moulines, E., & Laroche, J. (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16(2), 175–205.
- Peng, G., Zhang, C., Zheng, H.-Y., Minett, J. W., & Wang, W.S.-Y. (2012). The effect of intertalker variations on acoustic-perceptual mapping in Cantonese and Mandarin tone systems. *Journal of Speech, Language, and Hearing Research*, 55(2), 579–595.

- Peng, G., Zheng, H.-Y., Gong, T., Yang, R.-X., Kong, J.-P., & Wang, W.S.-Y. (2010). The influence of language experience on categorical perception of pitch contours. *Journal of Phonetics*, 38, 616–624.
- Qin, Z., Tremblay, A., & Zhang, J. (2019). Influence of within-category tonal information in the recognition of Mandarin-Chinese words by native and non-native listeners: An eye-tracking study. *Journal of Phonetics*, 73, 144–157. <https://doi.org/10.1016/j.wocn.2019.01.002>.
- Shen, G., & Froud, K. (2016). Categorical perception of lexical tones by English learners of Mandarin Chinese. *The Journal of the Acoustical Society of America*, 140(6). <https://doi.org/10.1121/1.4971765>.
- Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech*, 34, 145–156.
- Shen, X., Lin, M., & Yan, J. (1993). F0 turning point as an F0 cue to tonal contrast: A case study of Mandarin tones 2 and 3. *Journal of the Acoustical Society of America*, 93, 2241–2243.
- Wang, W. S.-Y. (1976). Language change. *Annals of the New York Academy of Science*, 208, 61–72.
- White, C. (1981). Tonal pronunciation errors and interference from English intonation. *Journal of Chinese Language Teachers Association*, 16(2), 27–56.
- Xu, Y., Gandour, J. T., & Francis, A. L. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *Journal of Acoustical Society of America*, 120(2), 1063–1074.
- Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, 27(1), 153–201.
- Zheng, H.-Y., Minett, J. W., Peng, G. & Wang, W. S.-Y. (2012). The impact of tone systems on the categorical perception of lexical tones: An event-related potentials study. *Language and Cognitive Processes*, 27(2), 184–209.

What if What You Think is the Opposite of What I Say? Evidence from Putonghua/Lanzhou Bidialectal Speakers on the Online Processing of Two Tonal Systems



Yadong Xu and Kevin Russell

Abstract A challenge in word recognition of Chinese bidialectal speakers is that their two linguistic systems may contain conflicting information. For instance, the pitch contours of Putonghua tone 2 and tone 4 are systematically reversed in Lanzhou Mandarin. That is, the word 轴 ‘axis’ is pronounced with a rising contour in Putonghua but a falling contour in Lanzhou, conversely, the word 咒 ‘to curse’ is pronounced with a falling contour in Putonghua but a rising contour in Lanzhou. This study uses the visual world paradigm to investigate whether this conflict causes interference for Putonghua/Lanzhou bidialectal listeners in word recognition of both dialects. Our behavioral and eyetracking results show that the bidialectal listeners experience greater competition from the opposite-tone competitor for words with the reversed tones (tone 2 and 4) in Lanzhou stimuli, compared to words with non-reversed tones (tone 1 and 3). But the evidence of interference is much weaker in their recognition of Putonghua. Our findings confirm the interference of conflicting tonal information in bidialectal listener’s word recognition. The imbalanced interference between two dialects suggests that proficiency plays an important role, specifically, it is their dominant dialect (Putonghua) that affects the processing of the non-dominant dialect (Lanzhou).

Keywords Bilingual word recognition · Eyetracking · Mandarin dialects · Interfering effect of tones

1 Introduction

The most common local dialect of Mandarin spoken in the city of Lanzhou (Gansu Province) uses different pitch contours for its tones, with the contours of tone 2 and 4 essentially reversed compared to Putonghua/Beijing Mandarin, as illustrated in

Y. Xu (✉) · K. Russell (✉)
University of Manitoba, Winnipeg, MB, Canada
e-mail: yadong.xu@umanitoba.ca

K. Russell
e-mail: kevin.russell@umanitoba.ca

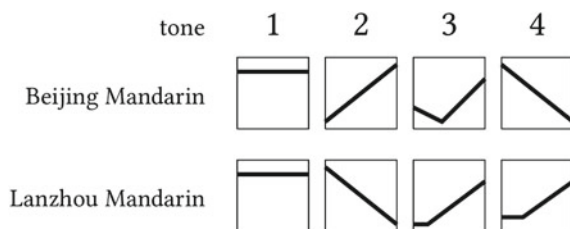


Fig. 1 Pitch contours for tones 1–4 in Putonghua/Beijing Mandarin and Lanzhou Mandarin. Throughout this paper, we will refer to tones 2 and 4 as *disagreeing tones* (i.e., tones whose pitch contours move in the opposite direction in the two dialects) and refer to tones 1 and 3 as *agreeing tones* (i.e., tones whose pitch contours move in roughly the same direction in both dialects)

Fig. 1. For example, the word 轴 *zhou2* ‘axis’ is pronounced with a rising contour in Beijing and a falling contour in Lanzhou, while the word 咒 *zhou4* ‘to curse’ is pronounced with a falling contour in Beijing and a rising contour in Lanzhou.

A very large number of people in Lanzhou grow up speaking the local dialect in the community, but have learned to become proficient second-dialect speakers of Putonghua in the education system. What are the consequences for word recognition when a speaker has learned two dialects¹ that have diametrically opposed tonal properties for roughly half of the vocabulary? In this paper, we use eyetracking in the visual world paradigm to investigate the potential for a bidialectal Lanzhou/Putonghua speaker to experience cross-dialect interference while recognizing words that have ‘disagreeing’ tonal specifications in the two dialects.

1.1 Mandarin Dialects: Lanzhou Mandarin

Putonghua is based on the Beijing dialect of Mandarin, but there are at least seven other sub-groups within the Mandarin family, including Northeast, Ji Lu, Jiao Liao, Zhongyuan, Lan Yin, Jianghuai, and Southwest (Wurm et al. 1987). Several of the dialects use different pitch contours for the four tones, and some dialects in the Lan Yin subgroup have fewer than four contrastive tones.

The Lan Yin subgroup includes the local varieties of Mandarin spoken in Lanzhou (population 3.6 million), the capital city of the Northwestern Province of Gansu. Previous research on Lanzhou Mandarin tones disagrees on details. For example, Norman (1988) and Zhang (2009) agree it to be a four-tone system but differ in the nuance of pitch contours, see Table 1. In the survey of tonal systems in Gansu

¹We use the term “dialect” in this paper to refer to any particular variety of language used by a group of people rather than its restrictive sense associating the language variety to its geographical regions. For native Mandarin speakers, Beijing Mandarin is perceived more ‘accented’ than Putonghua, but our interest primarily concerns about the tonal properties (pitch contours), which are identical between Putonghua and Beijing Mandarin. Hence, we will treat Beijing Mandarin and Putonghua interchangeably in this paper.

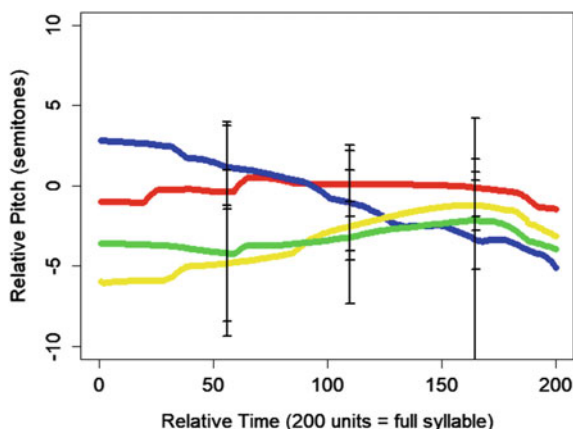
Table 1 Reported tone values of Lanzhou Mandarin

	Norman (1988)	Zhang (2009)
Tone 1	31	31
Tone 2	53	53
Tone 3	33	442
Tone 4	24	13

Province by Xu (2015: 67–68), the counties of Yongdeng, Gaolan, and Yuzhong are classified as three-tone dialect regions; while the districts of Honggu and Xigu (only in the Majiashan area) are classified as two-tone dialect regions. At least for Yuzhong, tone reduction seems to be a recent innovation—Zhang (1990) reports it having four tones and the later fieldwork by Zhang (2009) reports it as having only three tones. To our knowledge, there has been no research at all into the ramifications for the dialects’ tonal systems of the fact that so many of their speakers are bidialectal in Beijing-based Putonghua.

The tonal system that the first author of this paper hears most often in her native Chengguan district of Lanzhou is more similar to the three-tone system that Norman (1988) gives for Yinchuan, another member of the Lan Yin subgroup. Its contours are illustrated in Fig. 2 which shows the normalized pitch tracks of the word productions of one of the participants in this experiment (using Stanford’s 2008 tonetic method). Note that tone 3 (yellow) and tone 4 (green) are extremely similar, though perhaps not completely neutralized. Since all of the Lanzhou participants in this experiment come from the same central district of Chengguan, we will refer to this tonal system as simply ‘Lanzhou Mandarin’.

Fig. 2 Normalized pitch tracks for Lanzhou tones: tone 1 (red), tone 2 (blue), tone 3 (yellow), and tone 4 (green)



1.2 *The Visual World Paradigm (VWP), Tone, and Bilingualism*

One of the most commonly used experimental methodologies for studying word recognition is tracking the participants' eyegaze in the *visual world paradigm* (VWP), first introduced by Allopena et al. (1998). In Allopena et al.'s experiment, participants heard auditory instructions to click on and drag various pictured objects around a computer screen. In a more streamlined form of the VWP, a participant simply sees an image or a printed word in each corner of the computer screen, then hears an auditory word corresponding to one of the four corners, and has to click with their mouse in the appropriate corner as quickly as possible. Such studies *can* produce some useful results in the usual behavioral measures of response time and accuracy, although the response times are much slower and more variable than in, for example, a lexical decision experiment, since participants can't just press a button as soon as they've made a conscious decision, but must spend extra time moving the mouse pointer to the appropriate corner of the screen.

What researchers are more interested in VWP experiments is the data from the eyetracker that shows where on the computer screen the participants are looking at *while* they are still in the middle of recognizing the auditory word. Consistently across such experiments, the proportion of participants who are looking at the corner corresponding to the target word increases steadily during the first second or so after the onset of the auditory stimulus, until nearly 100% of participants are looking at the correct corner. Obviously, any participant can only be looking at one corner at a time, but the proportions of eyegazes toward the various images on the screen, averaged across all participants, corresponds closely to what most current theoretical models of word recognition predict as the activation levels of the various candidate words that are competing for selection within the word recognition process of any single listener.

Even more interesting than looking at how the target word gains an increasing share of gazes is looking at the share for competing words that aren't the target word but which the listeners may have been seriously considering as candidates. In Allopena et al. (1998), for example, one screen that participants saw included a picture of a *beaker* (the *target word* that the participants hear), a picture of a *beetle* (a *cohort competitor* that overlaps in the initial phonemes), a picture of a *speaker* (a *rhyme competitor* that shares the rhyme), and a picture of a *carriage* (an unrelated *distractor*). The graph of eyegaze proportion toward each shows *beaker* and *beetle* increasing neck-and-neck while the participants hear [bi], until about a quarter second after it becomes clear that the auditory stimulus contains [bik] and not [bit]—it takes about a quarter second after someone has decided to change the direction of their gaze for their eyes to finish the saccade—after which *beetle*'s share of gazes slowly dwindles down to 0. The rhyme competitor *speaker* enjoys a weak growth spurt once listeners notice that the end of the auditory stimulus [bikə] is compatible with the end of [spikə], but not enough to overcome what is by then the overwhelming lead of *beaker*.

So the visual world paradigm allows researchers to track not just which word wins the word recognition competition, but also to explore which other words can act as serious competitors.

Researchers studying Chinese word recognition have used the VWP to show that listeners take tonal information into account as soon as they can. For example, Malins and Joanisse (2010) showed that tonally divergent *hua4* ‘painting’ starts losing ground in its competition against the target word *hua1* ‘flower’ just as early as segmentally divergent *hui1* ‘gray’ does (Schirmer et al. (2005) reach a similar conclusion about the timing of tonal and segmental information using an event-related potentials methodology. The fact that participants can in principle use tone information to eliminate competitors as early as possible will be relevant for our experiment.).

In the study of bilingualism, VWP has allowed researchers to show how words in one of the listener’s languages can interfere with word recognition in their other language. For example, Weber and Cutler (2004) showed that words from their L1 act as serious competitors while the listener is trying to recognize words in their L2, e.g., Dutch *dekse1* ‘lid’ acts as a competitor for a Dutch/English bilingual trying to recognize English *desk*. Ju and Luce (2004) showed that if the phonetic realizations are anomalous enough, even words of an L2 can be competitors in L1 word recognition, e.g., when a Spanish/English bilingual listening to Spanish hears [p^hlaja] with strong English-like aspiration, the English word *pliers* can act as a competitor to the Spanish target *playa* ‘beach’. Phonological features which are relevant only in words of the listener’s L1 can still affect the recognition of those words’ L2 translation equivalents. For example, Shook and Marian (2016) found that Mandarin/English bilinguals performing a translation task would look toward the translation 树 on the screen more quickly if they heard the English word *tree* spoken with a falling pitch (matching *shu4*) than if it was spoken with a different pitch contour. Wang et al. (2017) found that the English word *rain* acted as a strong competitor for Mandarin/English bilinguals trying to recognize the English word *feather*, since both are translated as *yu3* in Mandarin, but neither English *fish* (translated to Mandarin as *yu2*) nor *wheat* (translated as *gu3*) acted as competitors—once again showing that tonal information is just as effective at eliminating competitors as segmental information is, even if, as here, that tonal information is completely irrelevant to the L2 words.

1.3 This Study

Our goal is to extend these findings to Chinese speakers who speak two dialects that have conflicting tone information—Lanzhou Mandarin and Putonghua. From earlier research, we expect that listeners *should* be able to use tonal information quickly to eliminate competitors during word recognition. We expect that they would also be able to use tonal information that exists in words of their first dialect to eliminate competitors while recognizing words in their second dialect. But for

Lanzhou/Putonghua bidialectal listeners, that tonal information is conflicting in the two dialects.

We should be able to use this conflict to test what kind of interference there is between the two dialects in word recognition. When a Beijing Mandarin listener hears *zhóu* with a rising pitch and has to recognize it as 轴 ‘axis’ *zhou2*, they will be able to use tonal information to quickly and efficiently rule out the word 咒 ‘to curse’ *zhou4* as a serious competitor. But a Lanzhou/Putonghua bidialectal listener may experience interference between their two dialects. Hearing Beijing Mandarin *zhóu* may activate the rising-pitch Lanzhou word 咒 ‘to curse’ *zhou4* as a serious competitor. If there is also interference from ‘L2’ to ‘L1’, then hearing Lanzhou Mandarin *zhóu* may also activate the rising-pitch Beijing word 轴 ‘axis’ *zhou2* as a serious competitor. In the visual word paradigm, the effects of this competition should be detectable in the relative proportion of eyegaze toward the target relative to its opposite-tone competitor, and perhaps also in slower reaction times and lower accuracy.

The eyetracking portion of the experiment should essentially be considered as a pilot study. We used the Eye Tribe, a passive eyetracker mounted just below the computer screen, which sold for \$100 (US) between 2013 and 2016, when the company was bought out and the product was discontinued. Our first practical goal for the eyetracking pilot study was to test whether a \$100 eyetracker would be good enough to be used in psycholinguistic studies using the visual world paradigm. One weakness in eyetracking studies is that the researchers often choose the time window to analyze only after looking at the data, which raises concerns that they may not be basing their choice on principled grounds but might instead be ‘cherry-picking’ the time window that best supports their hypotheses in this particular set of data, thus, increasing the chance of Type I error. So our second practical goal for the pilot study was to determine what would be an appropriate time window for us to plan ahead of time to use in a future larger study.

Our specific hypotheses for this study are:

Hypothesis 1 If Putonghua acts like the ‘L2’ of Lanzhou Mandarin speakers, then their knowledge of Lanzhou Mandarin may interfere with their comprehension of a speaker of Beijing Mandarin. In particular, we would predict that Lanzhou speakers listening to Beijing Mandarin words will be slower and less accurate on words with disagreeing tones (2 and 4) than on words with agreeing tones (1 and 3). In the eyetracking data, we predict that the target word will be more slowly and weakly activated (as reflected in eyegaze toward the target character) if it has a disagreeing tone than if it has an agreeing tone.

Hypothesis 2 There may be bidirectional interference between the dialects. Based on the literature on bilingual word recognition, this is less likely to be the case than a unidirectional influence of ‘L1’ on ‘L2’. But, if it *is* the case, we would predict the same pattern when Lanzhou participants listen to Lanzhou Mandarin as we predicted in Hypothesis 1 for listening to Beijing Mandarin: Words with disagreeing tones should be responded to more slowly, less accurately, and with less gaze toward the target in the eye-tracking data.

2 Method

2.1 Participants

Twenty adult Mandarin Chinese native speakers living in Manitoba were recruited for this experiment—ten speakers of the Lanzhou dialect and ten speakers of the Beijing (or a similar) dialect. All were born in mainland China and did not leave for Canada at least until their adulthood. The Lanzhou speakers were all from Chengguan district, Lanzhou City, a district which is reported to have reversed tones 2 and 4, but not to have neutralized the contrast between tones 3 and 4. The speakers in the ‘Beijing’ group were from Beijing or from neighboring northern cities whose dialects fall in the Northeast or Ji Lu subgroup, closely resembling the standard dialect in tonal properties. The participants were compensated with fifteen Canadian dollars for their participation.

In order to minimize as much as possible the influence of written Chinese (and thus possibly Putonghua) on the Lanzhou participants on the day of the experiment, participants received the informed consent form in Chinese several days before their appointment and were asked to read it and ask any questions they had about it before the day of their appointment.

During their session, participants filled out a questionnaire on their language background, rating their proficiency in both dialects as well as how often they used each. The proficiency question asked how familiar they were with Putonghua and with Lanzhou dialect. The proficiency scale was gradient from 0 to 10 with a 1-point interval. The lowest score 0 means they have ‘no knowledge of’ that dialect, and the highest score 10 means that they are ‘perfectly fluent in’ that dialect. The frequency question asked how often they had used each dialect in the previous month, with the score of 0 representing ‘never’, 0.5 ‘occasionally’, and 1 ‘daily’. As seen in the summary of Table 2, both groups were highly proficient in Putonghua and used it daily. Unfortunately, not exactly as expected, the Lanzhou group rated their proficiency in Putonghua as higher than their proficiency in Lanzhou Mandarin and reported using Putonghua more frequently. In interpreting the results that follow, it would be useful to keep in mind that the speakers in our Lanzhou dialect are indeed

Table 2 Mean proficiency and frequency of participants

Participants	In Putonghua	In Lanzhou dialect
Proficiency of Lanzhou group	9.7	7.5
Proficiency of Beijing group	10	0.5
Frequency of Lanzhou group	0.95	0.25
Frequency of Beijing group	1	0

bidialectal, but for most of them, it appears more accurate to say that Putonghua is their dominant dialect and Lanzhou Mandarin is their non-dominant dialect.

2.2 Materials

Each target or distractor character used in the eyetracking experiment was drawn from a pool of monosyllabic words meeting the following criteria:

- It represented an unambiguous (segmental) syllable (thus excluding, for example, 行 = xing2/hang2);
- It represented an unambiguous tone (excluding, for example, 空 = kong1/kong4);
- It had at least moderately high frequency—our least frequent stimulus ranks 3195th in the list of character frequencies of Da (2005).

There were 94 pairs of words with the same segmental syllable but with *disagreeing* tones 2 and 4, e.g., 竹 *zhu2*, 住 *zhu4*. There were 119 pairs of words with the same syllable and *agreeing* tones 1 and 3, e.g., 科 *ke1*, 可 *ke3*. There were another 340 characters, representing 206 other syllables (i.e., distinct syllables that do not repeat any syllables used in the 119-pair set) that were used only as distractors, e.g., 工 *gong1*, 共 *gong4*.

In our experiment, we use written Chinese characters as visual stimuli instead of pictures. This broadens our set of potential stimuli beyond highly imageable nouns to any relatively common one-syllable word of Chinese. Several studies in the visual world paradigm have used written words on the computer screen and found comparable results to studies that use pictures, e.g., McQueen and Viebahn (2007) and Mitterer and Russell (2013) on Dutch, Shook and Marian (2016) on Chinese.

During the experiment, each participant experienced a different randomized set of targets, competitors, and distractors. To minimize priming effects, the same syllable was never used more than once as a target in the 128 trials of the experiment, and no distractor ever used the same syllable as any of the targets. We also managed to make sure that 82% of the distractors didn't share a syllable with any other distractor. During the randomization process for each participant, 64 target words were first drawn from the pool of 'disagreeing' pairs, 32 with tone 2 and 32 with tone 4. Then, 64 target words were drawn from the pool of 'agreeing' pairs—32 with tone 1 and 32 with tone 3—subject to constraint that none of the syllables had already been chosen as a 'disagreeing' target. All remaining unchosen syllables from the 'agreeing' and 'disagreeing' pools were added to the pool of distractors, from which 320 distractors were chosen. Since the distractor pool contained only 294 unique syllables at this point, it was necessary to repeat 26 syllables (but never the same tone and character) as distractors in different trials of the experiment. For example, if 工 *gong1* was used as a distractor early in the experiment, 共 *gong4* might be used a distractor later in the experiment.

For each participant, half of the target syllables were used in the initial Lanzhou block of the experiment, and half in the second Beijing block.

Half of the target syllables were used in a *competitive* condition, where the opposite-tone competitor was one of the three other characters appearing on the screen. Half were used in a *non-competitive* condition, where the target character appeared with three unrelated distractors. The non-competitive trials were essentially filler trials intended to keep the participants from concluding that two of the characters on the screen always shared the same syllable and that the right answer was always one of those two. Nevertheless, the non-competitive trials can tell us whether participants have trouble processing disagreeing tones even without a competitor on the screen, in the same way that many eyetracking studies have found an effect of higher cohort competition (more competitor words that share the same initial phonemes), regardless of whether one of those competitors is actually shown on the screen.

The number of tokens used in each condition is shown in Table 3.

The auditory stimuli that the participants heard were recorded by two female speakers, who were not participants in the experiment. The first author, who is from the Chengguan district of Lanzhou, read the Lanzhou stimuli. The Beijing stimuli were read by a speaker from Hebei Province, near Beijing. Each word was read in the carrier phrase *wo xianzai yao nian ___ zhe ge zi* (我现在要念 ___ 这个字) ‘I now am going to read ___ this word’. The recordings were made using Praat in the sociolinguistics laboratory at the University of Manitoba and saved as WAV files at a sample rate of 44.1 K and a 16-bit resolution (see Appendix for the complete list of stimuli). Later, the 432 potential target words were segmented out of the carrier phrase and saved as individual WAV files.

Table 3 Number of target characters used in each condition

Lanzhou auditory stimuli	<i>Competitive</i> trial	Agreeing tone	Tone 1	8
			Tone 3	8
		Disagreeing tone	Tone 2	8
			Tone 4	8
	<i>Non-competitive</i> trial	Agreeing	Tone 1	8
			Tone 3	8
		Disagreeing	Tone 2	8
			Tone 4	8
Beijing auditory stimuli	<i>Competitive</i> trial	Agreeing	Tone 1	8
			Tone 3	8
		Disagreeing	Tone 2	8
			Tone 4	8
	<i>Non-competitive</i> trial	Agreeing	Tone 1	8
			Tone 3	8
		Disagreeing	Tone 2	8
			Tone 4	8

2.3 Procedure

The experiment was run using the experiment software OpenSesame (Mathôt et al. 2012), using PyGaze to interface with the eyetracker (Dalmaijer et al. 2014). The participants were tested individually in the linguistics laboratory of the University of Manitoba. When the participant arrived for their session, the experimenter briefed them about the three tasks they would perform: picture naming, eyetracking, and word-reading. These instructions were given in Putonghua to the Beijing group and in Lanzhou Mandarin to the Lanzhou group, in order to minimize the influence of being exposed to Putonghua just before the experiment. Since many of the Lanzhou participants may strongly associate written Chinese with Putonghua, which might prime them to be more influenced by Beijing Mandarin tonal patterns, we also tried to minimize the participants' exposure to written Chinese during the experiment for as long as possible. For this reason, the picture naming task was done first; in the eyetracking task, which necessarily involved written characters, all participants listened to the block of Lanzhou stimuli before the block of Beijing stimuli.

2.3.1 Picture Naming Task

Since it is possible to live in Lanzhou using only Putonghua, a brief picture naming task confirmed that each Lanzhou participant really was familiar with the Lanzhou dialect being studied. For symmetry, Beijing participants performed it too. Participants saw five pictures (i.e., 'keyboard', 'corn', 'horse-racing', 'Forbidden City', and 'beef noodles'), in a random order, in the center of the computer screen, and were asked to name the object in the picture in their hometown dialect. Each participant in the Lanzhou group pronounced the words as would be expected for a speaker of the Lanzhou dialect.

2.3.2 Eye-Tracking Task

For the eye-tracking task, participants wore closed-back headphones and sat about 60 cm away from an ASUS laptop with a 14-inch screen. An Eye Tribe eyetracker was mounted on a short table-top tripod immediately in front of the laptop about 50 cm from the participants' eyes. Since our preliminary tests suggested that the Eye Tribe was easily confused by head movement, participants placed their chins on a chin-rest attached to a table-top tripod that was adjusted to a comfortable height. Before any trials began, the Eye Tribe's nine-point calibration procedure was conducted. The sampling rate of the eyetracker was 30 Hz.

The first three trials were practice trials to familiarize the participants with the task.

In each trial, participants looked at the fixation dot in the center of an otherwise blank screen. Then, four characters appeared in the corners of the screen, as illustrated

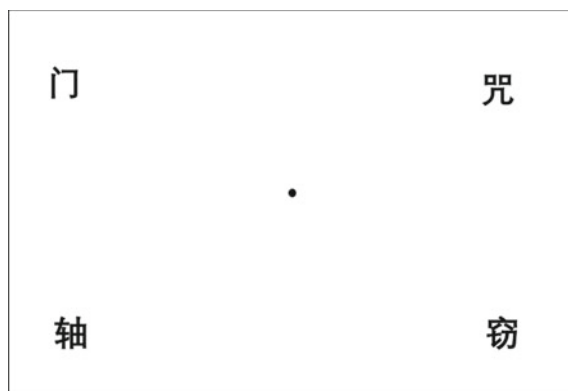


Fig. 3 Example of a possible computer screen in one trial. The four characters (clockwise from the top-left) are *men2*, *zhou4*, *qie4*, and *zhou2*. In this example, *zhou4* would be the target character that the participant should click on, *zhou2* is the competitor, and *men2* and *qie4* are the unrelated distractors. In non-competitive trials, the competitor *zhou2* would be replaced with a third unrelated distractor

in Fig. 3. One second later, the sound recording of the target word began playing over the headphones, and participants used the mouse to click on the character that matched the word they heard as quickly as possible. Once they clicked, the experiment advanced to the next trial. If participants did not click on any character within four seconds, the trial timed out and the experiment advanced to the next trial.

For the first block of 64 trials, participants were told (truthfully) that they would hear words spoken in the Lanzhou Mandarin dialect. There was a ten-minute break after the first block. In the second block of 64 trials, participants were told (truthfully) that they would hear words spoken in Putonghua.

2.3.3 Word-Reading Task

As the final task of the experiment, participants read out loud, using their preferred pronunciation, a list of words in randomized order, which included representative words in each tone. These recordings were used for acoustic analysis of the participants' tone systems, and the results of which are not reported here.

3 Results

The results include the standard behavioral measurements of response time and accuracy for the participants' mouse clicks, as well as analysis of data from the eyetracker that recorded every 30 ms which location on the computer screen the participant was looking at. All statistical analyses were performed in version 4.0.0

of R (R Core Team 2020). Mixed-effects models were constructed using the lme4 package (Bates et al. 2015), and the lmerTest package provided estimated p -values using the Satterthwaite approximation for degrees of freedom (Kuznetsova et al. 2017).

3.1 Behavioral Results

3.1.1 Accuracy

We eliminated trials where the response was faster than 600 ms after word-onset (0.8% of the data) because those clicks mostly happened before the auditory stimulus even began. Timed-out trials were counted as incorrect (1.5% of the data).

Figure 4 shows violin plots of participants’ accuracy for each of the four tones, both when they were listening to their own dialect and when they were listening to the other dialect. Table 4 summarizes participants’ mean accuracy by the target’s tone condition (agreeing vs. disagreeing). Unsurprisingly, Beijing listeners performed at or near ceiling on Beijing stimuli, but made considerably more errors on Lanzhou stimuli, especially on disagreeing tones 2 and 4. Lanzhou listeners performed *almost* as well on Beijing stimuli as Beijing listeners do; in contrast, their performance on Lanzhou stimuli showed much more individual variation, with a lower average accuracy than their performance on Beijing stimuli.

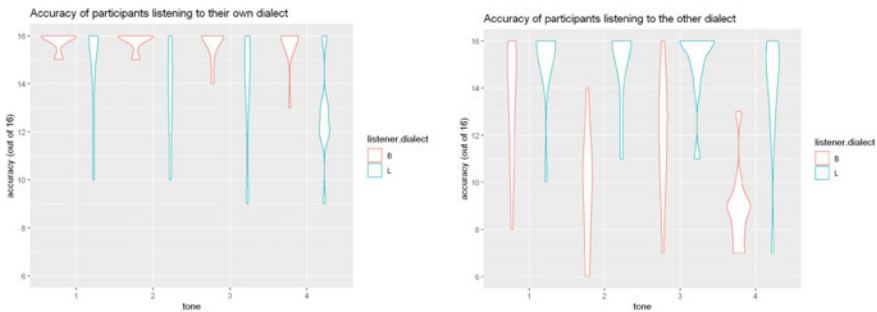


Fig. 4 Participants’ accuracy

Table 4 Mean accuracy (%) for the mouse click

	Lanzhou stimuli		Beijing stimuli	
	Agreeing	Disagreeing	Agreeing	Disagreeing
Lanzhou listeners	90.25	84.81	98.71	94.48
Beijing listeners	81.59	59.37	99.37	98.75

Table 5 Coefficients of a logistic mixed-effects model predicting accuracy of Lanzhou listeners

	Estimate	Std. error	z value	Pr (> t)
(Intercept)	6.3236	0.5663	11.166	<2e-16***
Lanzhou stimulus	-1.6207	0.2629	-6.166	7.02e-10***
Disagreement	-0.7993	0.2390	-3.345	0.000823***
Competitive	-3.0421	0.3997	-7.610	2.74e-14***

0 ***, 0.001 **, 0.01 *

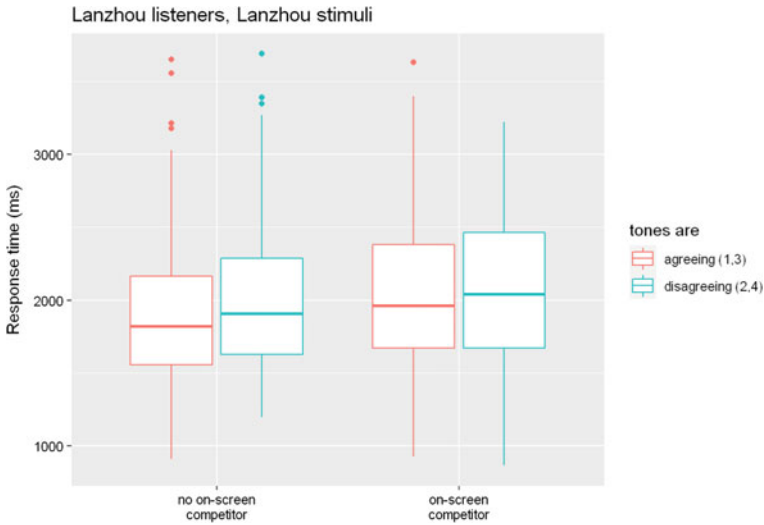


Fig. 5 Boxplot of Lanzhou listeners’ response times to Lanzhou stimuli

These impressions of the Lanzhou listeners’ accuracy are borne out statistically. Table 5 shows the coefficients of a logistic mixed-effects model predicting the accuracy of Lanzhou listeners’ mouse clicks with participant and target word as random effects. The odds of a Lanzhou listener making an error are about 21 times greater if the opposite-tone competitor is on-screen (i.e., increased log-odds of 3.04), about 5 times greater if they are responding to a Lanzhou stimulus rather than a Beijing stimulus, and about 2.2 times greater if the tones are disagreeing than if the tones are agreeing. All three effects are significant.

3.1.2 Response Time

Our main concern is whether Lanzhou listeners experience interference (i.e., longer response times) when listening to disagreeing tones.

Table 6 Coefficients of a mixed-effects model predicting response time for Lanzhou listeners hearing Lanzhou stimuli

	Estimate	Std. error	df	t value	Pr (> t)
(Intercept)	1920.09	76.14	11.77	25.218	1.31e-11***
Competitive	139.38	41.10	535.98	3.391	0.000748***
Disagreement	56.82	40.74	535.97	1.395	0.163684

0 ***, 0.001 **, 0.01 *

Table 7 Coefficients of a mixed-effects model predicting RT, including an interaction between tone disagreement and the listener's self-rated Lanzhou dialect proficiency

	Estimate	Std. error	df	t value	Pr (> t)
(Intercept)	2005.038	121.364	9.638	16.521	2.16e-08***
Competitive	140.393	40.957	534.783	3.428	0.000655***
Disagreement	-50.485	63.432	535.658	-0.796	0.426454
Proficiency	-55.339	60.655	9.147	-0.912	0.384993
Disagreement: proficiency	68.854	31.305	535.019	2.199	0.028270*

The Lanzhou participants all rated themselves between 6 and 10 out of 10. The 'proficiency' variable in this model is their rating minus 6, so that the model intercept corresponds to the lowest-proficiency Lanzhou participants.

0 ***, 0.001 **, 0.01 *

The boxplot in Fig. 5 shows the RTs of Lanzhou listeners hearing Lanzhou stimuli, in trials where they correctly clicked on the target. The listeners appear to be slower if the trial has the opposite-tone competitor on-screen, and it appears that they may also be slower with disagreeing tones.

Unfortunately, the appearance of an effect for tone disagreement is not borne out statistically. Table 6 shows the coefficients of a mixed-effects model predicting RT with participant as a random effect. The presence of an on-screen competitor slows down responses by 139 ms, but the effect of tone disagreement is non-significant.²

However, during post-hoc exploration of the data, we did find a significant interaction between tone disagreement and the listener's self-rating of their proficiency in the Lanzhou dialect on the questionnaire, as shown in the model coefficients in Table 7. The lowest-proficiency Lanzhou listeners still show no significant effect of tone disagreement on RT (at least in their accurate responses), but the more proficient the listener is, the more their responses are slowed down by a tone disagreement. Since this is a post-hoc finding, it is more likely to be a Type I error than our planned comparisons, but it suggests that a replication of this study conducted in Lanzhou with more proficient dialect speakers could well-find evidence for our original hypotheses.

Figure 6 shows the corresponding boxplot for the RTs of the correct responses of Lanzhou listeners hearing Beijing stimuli. Table 8 shows the coefficients of a model

²Both the Akaike information criterion and the Bayesian information criterion prefer a model without tone disagreement as a predictor.

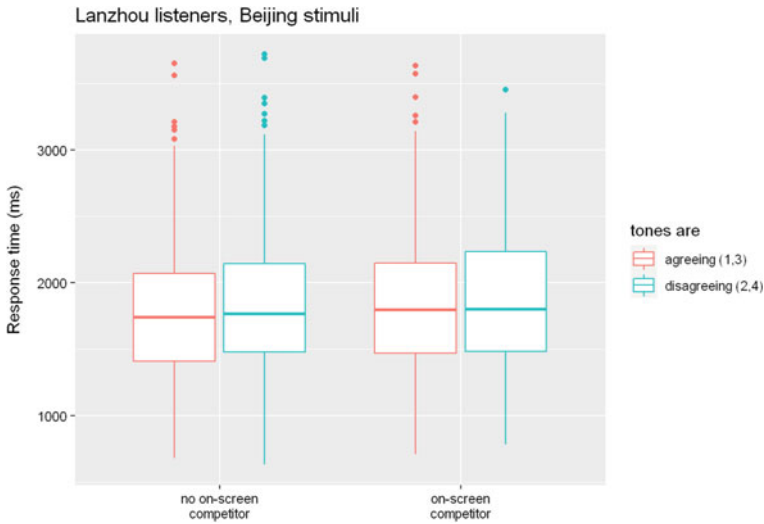


Fig. 6 Boxplot of response times of Lanzhou listeners hearing Beijing stimuli

Table 8 Coefficients of model predicting RT for Lanzhou listeners hearing Beijing stimuli

	Estimate	Std. error	df	t value	Pr (> t)
(Intercept)	1781.04	68.11	10.87	26.150	3.67e-11***
Disagreement	55.14	30.21	298.34	1.825	0.0690
Competitive	72.60	29.58	1128.40	2.454	0.0143*

0 ***, 0.001 **, 0.01 *

for Beijing stimuli RTs, with the same predictors as the Lanzhou stimulus model. The presence of an on-screen competitor is again significant, but it slows listeners down only half as much as it does for Lanzhou stimuli. The estimated effect size for tone disagreement is about the same as it was for Lanzhou stimuli, now with a marginal *p*-value of 0.069 (This model is not improved by adding either proficiency in either the Lanzhou dialect or Putonghua as predictors.).

3.2 Eyetracking Results

Figure 7 graphs all of the gaze locations recorded by the eyetracker during the trials of participant 1, who shows the same pattern as other participants. As we can see, almost all gazes were toward the fixation dot in the center or toward the four corners of the computer screen, where the characters were located. In a small minority of measurements, the eyetracker has caught the participant’s gaze in mid-saccade between one of these five locations. Therefore, we can answer the first

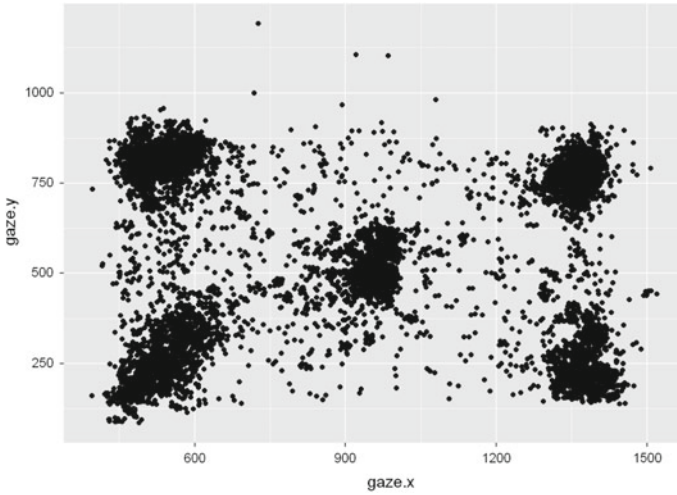


Fig. 7 All gaze locations for participant 1 during the experiment

practical question we posed for our eye-tracking pilot study: Yes, the \$100 Eye Tribe eyetracker is good enough to detect which corner of the computer screen the participant is looking at.

The pixel locations recorded by the eyetracker were recoded to designate which corner of the screen (if any) the participant was looking at. A gaze location was considered to be directed toward one of the four corners if both of its coordinates were at least 200 pixels away from that participant’s median gaze location. For example, if the median coordinates of a participant’s gaze were (1000, 500), then a gaze was considered to be toward the top-left corner of the screen if its x -coordinate was less than 800 and its y -coordinate was greater than 700. This was also recoded as a gaze toward the target character, the competitor character, or one of the distractor characters, depending on which corners those characters were in on that trial.

One Lanzhou participant needed to be excluded from the eye-tracking analysis due to equipment failure.³

Unlike most other eyetracking studies, we will not exclude trials where the participant chose the ‘wrong’ character with the mouse. In a typical eyetracking study, an incorrect behavioral response can be taken as evidence that the participant was not paying attention during that trial. In this study, we *expect* perfectly attentive participants to choose the wrong character much of the time. The whole point of eyetracking is to examine which candidates the listener is seriously considering *before* they have completely made up their mind about which word they are hearing, so it is largely

³The equipment failure was not due to the eyetracker, but to something on the laptop that caused random long delays between some trials, which allowed the eyetracker to drift out of calibration in fewer trials than for other participants. Honestly, we suspect we forgot to turn dropbox off before that session.

irrelevant which word they end up consciously deciding they heard and whether that choice corresponds to what the recorded speaker intended to say.

Figure 8 shows for Lanzhou listeners how the proportion of their gazes toward the target, the competitor, and each of the two distractors evolves in the time immediately following the onset of the auditory word. In interpreting Fig. 8, keep in mind that there is about a 250 ms lag between the point when listeners change their confidence about the identity of the word they are hearing and the point when that changed confidence becomes reflected in their eyegaze. The vertical dotted line in each panel indicates the median behavioral response time for that condition. While it takes someone about a quarter second to physically move their eyeballs after deciding they want to change the direction of their gaze, it takes even longer for them to physically move a computer mouse and click its button after deciding which character they want to choose. By the point of the median mouse RT, participants will have already made their word recognition decisions in a large majority of the trials. So, to the right of the dotted lines in Fig. 8, the gaze curves are based on a small and shrinking minority of exceptionally slow decisions.

When listening to Beijing stimuli (top two panels, Fig. 8), our bidialectal participants' eyegaze is mostly what we would expect for a monodialectal Beijing speaker. Their gazes toward the target character peak early and strongly, for both agreeing and disagreeing tones (though perhaps slightly later and less strongly for the disagreeing tones). The listeners spend barely any more time considering the opposite-tone

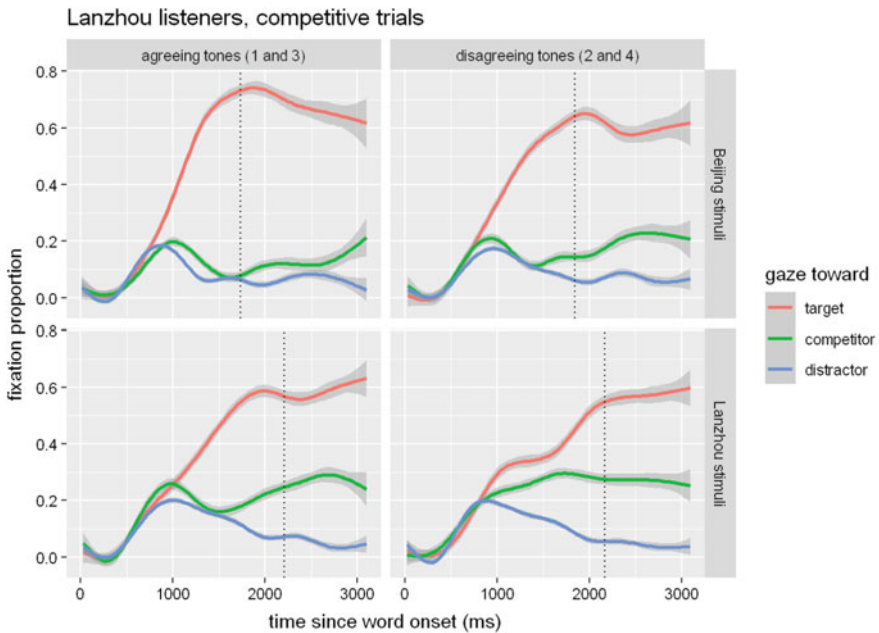


Fig. 8 Gazes for Lanzhou listeners in competitive trials. The dotted line indicates the median response time for mouse clicks in each condition

competitor than they spend considering the unrelated distractors. The only exception to this is in the small minority of very slow responses, especially to disagreeing tones, where listeners spend some time considering the opposite-tone distractor, perhaps sometimes second-guessing their earlier word identifications.

The situation is different when Lanzhou participants are listening to Lanzhou stimuli (bottom two panels, Fig. 8). Even when participants listen to *agreeing* tones, the opposite-toned competitor offers some serious competition to the target word for about the first second after word-onset, before its activation begins to die away as expected (though it makes a modest comeback in the small and shrinking minority of very slow responses). Crucially for us, when listening to a disagreeing tone, the opposite-tone competitor once again offers strong competition for about the first second, but in this condition, its activation *never* dies away; instead, the target pulls ahead of the competitor only slowly and indecisively.

Figure 9 shows the corresponding evolution in eyegaze for Beijing listeners. When Beijing participants listen to stimuli in their own Beijing dialect, the target quickly and decisively takes the lead over its opposite-tone competitor, and, unsurprisingly, this doesn't depend on whether the contours of the tones are reversed in somebody else's dialect. Their eyegaze behavior is also unsurprising when they're listening to words in the unfamiliar Lanzhou dialect. With agreeing tones 1 and 3, where the pitch contours are comparable in Beijing and Lanzhou, the Beijing listeners are able to eventually activate the target word weakly, though the unfamiliarity of

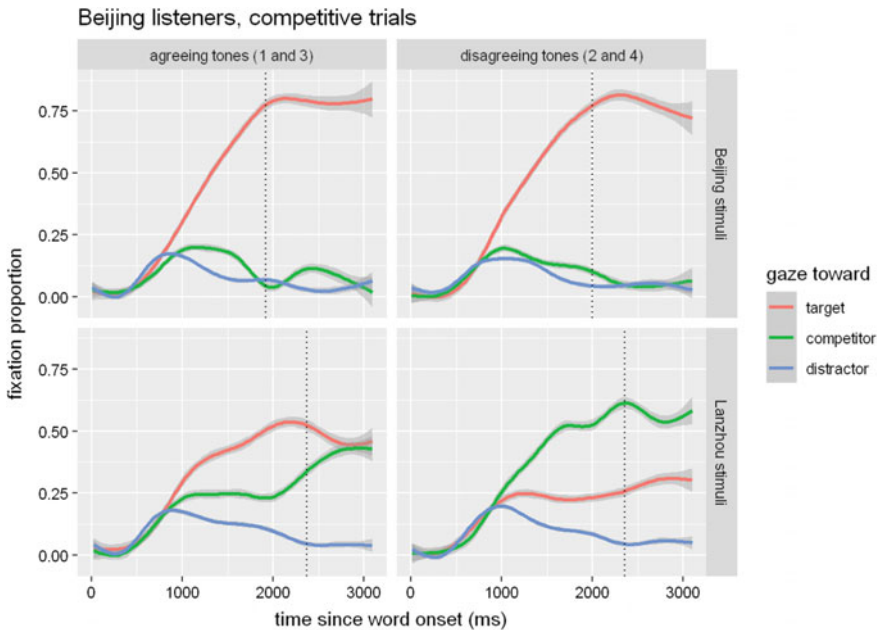


Fig. 9 Gazes for Beijing listeners in competitive trials. The dotted line indicates the median response time for mouse clicks in each condition

the dialect makes them unwilling to completely rule out considering the opposite-tone competitor as a candidate. The situation is the same for the disagreeing tones, where the pitch contours are reversed between Beijing and Lanzhou, except that the candidate that the listeners eventually settle here on is the ‘incorrect’ one.

These informal observations about the time-courses of the word recognition competition can be confirmed statistically, although it requires us to ‘cherry-pick’ an appropriate time window for post-hoc analysis.⁴ (Recall that one of the main purposes of the eye-tracking pilot study was to determine what an appropriate planned time window would be for future studies.) One common time window used in eyetracking studies is between 500 and 1000 ms that would be inappropriate here. In a typical eyetracking study on monolingual word recognition in European languages, gaze toward the target has usually reached or is approaching its peak by 1000 ms after word-onset. In our data, the target is barely beginning to pull away from the competitor and distractors by 1000 ms and its share of the gazes is rarely more than 30% in any condition. Instead, we will choose one of the longest reasonable time windows possible. Based on the above results, because the median RT for mouse clicks across all conditions was 1908 ms, and because the above graphs show that the winning candidate has reached its activation peak in all conditions by or shortly after 1900 ms, it is appropriate to use 1900 ms as the end of our window. Otherwise, a window later than this runs the risk of including many gaze locations that reflect the poorly understood processes, consequently affecting eyegaze after word recognition has already happened, rather than telling us anything about word recognition.

For eyetracking data, we computed two additional variables for each trial:

- (1) The *target proportion*, i.e., what proportion of the eyegazes recorded during that trial between 500 and 1900 ms are directed toward the target character.
- (2) The *target dominance*, i.e., the difference between the proportion of gazes directed toward the target and the proportion directed toward the competitor within the time window (This is only relevant for competitive trials, where the opposite-tone competitor actually appears on-screen.).

Table 9 gives the coefficients of the best mixed-effects model predicting the proportion of gazes toward the target for Lanzhou listeners responding to Lanzhou stimuli, with participant as a random effect. The coefficients tell us that listeners spend 5% less of the window looking toward the target when an opposite-tone competitor with an ‘agreeing’ tone (1 or 3) is on the screen, and that listeners also spend 6% less of the window looking toward the target when it has a ‘disagreeing’ tone (2

⁴The kind of statistical analysis we will be doing was standard in most eyetracking studies a decade ago. Many more recent eyetracking studies use growth curve analysis (GCA) to analyze their data, as outlined in Mirman et al. (2008) and Mirman (2014). GCA is valuable in removing the need for researchers to cherry-pick a particular narrow time window for statistical analysis, e.g., between 617 and 784 ms. But the researcher still needs to choose a wide time window that is usually shorter than the entire trial; otherwise, the GCA model will waste its parameters trying to fit late and uninteresting wiggles, like those toward the right-hand side of Figs. 8 and 9, producing coefficients that are impossible to interpret in a meaningful way. Our analysis will simply use that entire wide window anyway. Given our relatively modest amount of data, a GCA analysis offers no additional insights that would justify the space it would require to explain its complexities here.

or 4) than when it has an agreeing tone, even when the opposite-tone competitor is absent from the screen. Adding an interaction between on-screen competition and tone disagreement does not result in a better model (the approximated p -value of the interaction is 0.47 and both the Akaike information criterion and the Bayesian information criterion prefer the model without the interaction.). Thus, beyond the sum of 5 and 6%, there seems to be no additional disadvantage for trials that use *both* a ‘disagreeing’ tone target *and* an on-screen opposite-tone competitor.

Similarly, the best mixed-effects model for target dominance in the competitive trials has an intercept of 0.155 and a coefficient for tone disagreement of -0.096 ($t = 3.543$, $df \approx 11.56$, $p \approx 0.008$). In other words, when listeners hear a word with an agreeing tone, they spend about 16% more of the window looking at the target character than at its opposite-tone competitor. But when they hear a word with a disagreeing tone, they spend only 6% more of the window looking at the target than at its competitor.

For predicting target proportion when Lanzhou participants listen to Beijing stimuli, there is no significant effect of on-screen competitor or of tone disagreement.⁵

4 Discussion

The eyetracking component of this experiment served as a pilot study to determine whether an inexpensive Eye Tribe eyetracker could produce results that were usable in answering research questions about interference in word recognition in bidialectal speakers. The simple answer to this practical question is: Yes.

The results of our experiment allow us to begin to address the two research questions we began with.

Table 9 Coefficients of a mixed-effects model for target proportion, Lanzhou listeners hearing Lanzhou stimuli

	Estimate	Std. error	t value	Pr ($> t $)
(Intercept)	0.40121	0.02611	15.365	5.83e-09***
Disagreeing	-0.05951	0.01482	-4.015	6.75e-05***
Competing	-0.04971	0.01482	-3.354	0.000851***

0 ***, 0.001 **, 0.01 *

⁵That is, the Bayesian information criterion prefers a model with just the random effect of participant over all alternatives. However, the Akaike information criterion just barely prefers ($\Delta AIC = .76$) a model including both on-screen competition (though non-significant) and tone disagreement ($p \approx 0.0295$), and it prefers a little more strongly ($\Delta AIC = 1.57$) a model with competition, disagreement, *and* their interaction (all non-significant). By itself, this is unconvincing evidence that these participants’ knowledge of the Lanzhou dialect affects their Putonghua word recognition, although it does point weakly in the same direction as the marginally significant effect of tone disagreement on RTs to Beijing stimuli and the non-significance of any interaction to counteract the additive effect of tone disagreement and Beijing stimulus on accuracy.

Question 1: Does knowing both dialects interfere with how our Lanzhou participants recognize words spoken in Putonghua/the Beijing dialect, especially those that have disagreeing tones?

Essentially, no, or not very much.

Behaviorally, there is no obvious effect of disagreeing tone on their response time to Beijing Mandarin stimuli. The Lanzhou listeners have faster response times to Beijing words than to Lanzhou words, regardless of tone.⁶ Some Lanzhou participants are somewhat less accurate than the typical Beijing listener, again regardless of tone. But all of our Lanzhou participants are faster and at least as accurate when listening to Beijing words as they are when listening to Lanzhou words.⁷ The eyetracking data also produced no firm evidence of any effect of tone disagreement when listening to Beijing words.

The plausible conclusion from this is that none of our participants were actually dominant in the Lanzhou dialect. A couple could be considered as almost balanced bilinguals (equally competent in both dialects), but most are dominant in Putonghua and act as if the Lanzhou dialect were their ‘L2’, recall an imbalance reflected in their own self-ratings of their proficiency in the two dialects. Despite that, our Lanzhou participants are genuinely bidialectal. Disagreeing tones may confuse them and slow down their word recognition, but, unlike the Beijing listeners, they usually *are* able to eventually choose the correct target.

Question 2: Does knowing both dialects interfere with how our Lanzhou participants recognize words with disagreeing tones spoken in the *Lanzhou* dialect?

The answer is yes, although the implication of that answer is not, as we expected it would be, that there must be bidirectional interference between the dialects.

Behaviorally, our Lanzhou listeners make more mistakes when the tone is disagreeing, although they make even more mistakes still when the opposite-tone competitor of an agreeing tone is on the screen. We found no RT differences for disagreeing tones in our planned analysis, although a post-hoc analysis suggests that the most proficient Lanzhou speakers *may* be affected by the disagreement in tone contour between the dialects. The Lanzhou participants are still mostly giving right answers on the disagreeing tones, unlike the Beijing listeners, but whatever word recognition processes they are using for listening to Lanzhou seem to have almost as much trouble with agreeing tones as with disagreeing tones. The eyetracking results

⁶This is not just the result of a gradual increase in speed across all trials, making the second Beijing block faster on average than the first Lanzhou block. Response times were relatively constant throughout the first Lanzhou block, and relatively constant throughout the second Beijing block, with an abrupt discontinuity between the two blocks.

⁷Our Lanzhou participants were also faster overall than our Beijing participants, even when listening to Beijing stimuli. This may just be an irrelevant difference due to a small sample size—perhaps the Lanzhou participants in our sample happened to be somewhat more biased toward speed in the speed-accuracy trade-off than the Beijing participants were (which, if true, could also explain why their accuracy is also somewhat lower than that of the Beijing participants). Or it might be that the Beijing listeners got so confused by the unfamiliar stimuli in the initial Lanzhou block that they remained uncharacteristically cautious throughout the second block of Beijing stimuli.

strongly support interference from Putonghua on Lanzhou word recognition. Activation levels of the target word are weaker and slower to grow. Our listeners are worse at identifying the target word when an opposite-tone competitor is on the screen, and, crucially, they are worse at activating words that have disagreeing tones, even in the absence of an on-screen competitor.

One possible explanation is that our Lanzhou participants are just plain worse at using *all* tone information in their second dialect. The more proficient the speaker is in the Lanzhou dialect, the better they seem to be able to use unambiguous tone information that does not conflict with Putonghua, but they still have a disadvantage on words with disagreeing tone information.

Although this is not exactly either of the two situations we hypothesized, our results are consistent with a model where phonological information from the bidialectal listener's dominant dialect (here Putonghua) can interfere with word recognition in their non-dominant dialect (here, Lanzhou Mandarin). We did not find strong evidence for bidirectional interference, with mostly non-significant or marginally significant effects of tone disagreement when listening to Beijing Mandarin. A bidialectal listener's word recognition in their dominant dialect seems not to be affected very strongly by phonological properties of their non-dominant dialect.

We were hoping to find participants whose dominant dialect was Lanzhou Mandarin. Given that we could only recruit people who were living in Winnipeg, it isn't really surprising that they acted like Putonghua was their dominant dialect. Many of our Lanzhou participants were university students, meaning that they have been strongly influenced by Putonghua over several years of education (It would seem that these participants offer an existence proof that it *is* possible for children who speak minority Mandarin dialects to learn Putonghua at school without being at a permanent psycholinguistic disadvantage, but we must keep in mind that they are a biased sample drawn from those who actually succeeded in the education system.). Some of our participants had grown up speaking Lanzhou Mandarin, but had lived for years without speaking it in the predominantly Putonghua- and Cantonese-speaking Chinese community in Winnipeg. And some of our participants had always had Lanzhou Mandarin as their second dialect. For example, one participant grew up in Lanzhou in a family that exclusively spoke Putonghua and reported that he began learning Lanzhou Mandarin only later so that he could fit in better with other students at school.

We expect that repeating the study in Lanzhou will allow us to recruit participants with a wider range of proficiencies in Lanzhou Mandarin, and we continue to expect that the most proficient of those Lanzhou Mandarin speakers will behave in accordance with our original hypotheses.

The debriefing sessions with the participants after the experiment have left us with two other worrying indications that the dialectal situation on the ground may be more complicated than we expected.

First, speaking of 'Lanzhou Mandarin' is an oversimplification. As noted in the introduction, there are a number of different dialects within Gansu Province, and even within the city limits of Lanzhou. Many of these dialects have reversed tones (2 and 4) relative to Putonghua. Some of them have neutralized the contrast between tone

3 and (rising) tone 4. Some have neutralized even more contrasts to produce a two-tone system. One of our Lanzhou participants may have been a speaker of a two-tone dialect. Another participant, during the final word-reading task of the experiment, not only used consistently reversed contours for tone 2 and tone 4 relative to Beijing Mandarin, but also reversed the contours of tone 1 and tone 3 a majority of the time.

Second, speaking of ‘Beijing Mandarin’ is also an oversimplification. Our Beijing participants had greater than expected accuracy when listening Lanzhou Mandarin, often much greater than expected, even on disagreeing tones with an on-screen competitor. Part of this may be due to higher levels of noise/randomness in making a decision on an unfamiliar dialect, which would lower the accuracy on agreeing tones and raise the accuracy on disagreeing tones. There may also be something in the acoustics of Lanzhou tone 2 and tone 4 that make them sound less like Beijing tone 4 and tone 2 than we’d expect, i.e., calling the contours of tones 2 and 4 ‘reversed may not be completely accurate. But the cause may also lie partly in the Beijing listeners and not the Lanzhou stimuli. During the debriefing, one of our ‘Beijing’ participants remarked that the Lanzhou stimuli reminded him of how people spoke in his home village a very short distance outside the city of Beijing. Some other ‘Beijing’ participants, despite growing up in a region dominated by Beijing Mandarin, reported that they had had considerable exposure to non-Putonghua dialects, for example, from a parent or grandparents who had immigrated to Beijing from other provinces. It would be wise for future studies not to assume that Beijing Mandarin is essentially identical to Putonghua, or even that Beijing is a monolithic dialect region and that everybody growing up in its vicinity has had comparable language experiences.

5 Conclusion

We studied whether bidialectal speakers of both Lanzhou Mandarin and Putonghua would experience interference during word recognition resulting from the fact that tone 2 and tone 4 have reversed pitch contours between the two dialects. Although our Lanzhou participants were not dominant in the dialect that we expected them to be dominant in, evidence from eyetracking and behavioral accuracy, as well as a post-hoc effect on response time among the more proficient Lanzhou speakers, all point to the conclusion that their dominant dialect (Putonghua) interferes with word recognition in their non-dominant dialect (Lanzhou).

Appendix: List of Stimuli Used in Eyetracking Trials

1. List of targets and competitors

	Tone 1	Tone 2	Tone 3	Tone 4
1. /pi/	批 ‘batch’	皮 ‘skin’	痞 ‘ruffian’	屁 ‘fart’
2. /bi/	逼 ‘push’	鼻 ‘nose’	笔 ‘pen’	币 ‘coin’
3. /ti/	梯 ‘ladder’	提 ‘lift, carry’	体 ‘body’	替 ‘replace’
4. /di/	低 ‘low’	敌 ‘enemy’	底 ‘bottom’	地 ‘ground’
5. /tu/	秃 ‘bald’	屠 ‘killing’	土 ‘soil’	兔 ‘rabbit’
6. /du/	督 ‘supervise’	毒 ‘poison’	赌 ‘gamble’	渡 ‘ferry’
7. /chi/	痴 ‘obsession’	迟 ‘late’	尺 ‘rule’	翅 ‘wing’
8. /ci/	疵 ‘defect’	词 ‘word’	此 ‘here, this’	次 ‘secondary’
9. /fu/	敷 ‘cover’	扶 ‘hold’	府 ‘mansion’	富 ‘rich’
10. /wu/	诬 ‘slander’	吴 ‘a surname’	五 ‘five’	雾 ‘fog’
11. /ba/	八 ‘eight’	拔 ‘pull’	靶 ‘target’	罢 ‘finish’
12. /da/	搭 ‘match’	答 ‘answer’	打 ‘beat’	大 ‘big’
13. /chang/	昌 ‘prosperous’	尝 ‘taste’	厂 ‘factory’	畅 ‘smooth’
14. /zha/	渣 ‘crumble’	闸 ‘gate’	眨 ‘blink’	诈 ‘fraud’
15. /ke/	科 ‘science’	壳 ‘shell’	可 ‘can, may’	刻 ‘moment’
16. /ge/	哥 ‘brother’	隔 ‘separation’	葛 ‘a surname’	各 ‘each’
17. /mi/	眯 ‘squint’	迷 ‘lost’	米 ‘rice’	密 ‘dense’
18. /ma/	妈 ‘mother’	麻 ‘numb’	马 ‘horse’	骂 ‘scold’
19. /yi/	医 ‘medicine’	姨 ‘aunt’	乙 ‘second’	义 ‘righteous’
20. /zhou/	舟 ‘canoe’	轴 ‘axis’	肘 ‘elbow’	咒 ‘curse’
21. /cai/	猜 ‘guess’	财 ‘wealth’	彩 ‘colorful’	菜 ‘vegetables’
22. /duo/	多 ‘many’	夺 ‘deprive’	躲 ‘dodge’	剁 ‘chop’
23. /fan/	帆 ‘sail’	烦 ‘annoyed’	反 ‘inside out’	饭 ‘food’
24. /fang/	方 ‘square’	房 ‘house’	纺 ‘fabric’	放 ‘place’
25. /fei/	飞 ‘fly’	肥 ‘fat’	匪 ‘gang’	废 ‘waste’
26. /fen/	芬 ‘fragrance’	坟 ‘tomb’	粉 ‘pink’	奋 ‘strive’
27. /guo/	锅 ‘pot, wok’	国 ‘nation’	果 ‘fruit’	过 ‘pass’
28. /han/	憨 ‘silly’	韩 ‘a surname’	喊 ‘shout, yell’	汗 ‘sweat’
29. /hu/	忽 ‘sudden’	胡 ‘mustache’	虎 ‘tiger’	户 ‘household’
30. /huan/	欢 ‘merry’	环 ‘ring’	缓 ‘postponed’	幻 ‘unreal’
31. /hui/	灰 ‘grey’	回 ‘return’	毁 ‘destroy’	汇 ‘merge’
32. /ji/	击 ‘strike’	急 ‘hurry’	挤 ‘squeeze’	记 ‘mark’
33. /mao/	猫 ‘cat’	毛 ‘hair’	柳 ‘stud’	贸 ‘trade, commerce’
34. /miao/	喵 ‘meow’	苗 ‘seedling’	秒 ‘second’	庙 ‘temple’
35. /piao/	飘 ‘flutter’	嫖 ‘debouch’	瞟 ‘glance’	票 ‘ticket’
36. /pin/	拼 ‘put together’	贫 ‘poor’	品 ‘product’	聘 ‘employ’
37. /pu/	扑 ‘throw oneself at’	菩 ‘bodhi’	普 ‘ubiquitous’	瀑 ‘falls’
38. /qi/	期 ‘period’	棋 ‘chess’	企 ‘company’	气 ‘air’
39. /qu/	区 ‘district’	渠 ‘canal’	取 ‘pick’	趣 ‘fun’
40. /shen/	身 ‘body’	神 ‘god, divine’	审 ‘interrogate’	肾 ‘kidney’

(continued)

(continued)

	Tone 1	Tone 2	Tone 3	Tone 4
41. /xi/	西 'west'	习 'exercise'	洗 'wash'	细 'thin, careful'
42. /shi/	诗 'poem'	石 'stone'	史 'history'	事 'matter'
43. /she/	赊 'owe'	蛇 'snake'	舍 'abandon'	射 'shoot'
44. /wa/	洼 'puddle'	娃 'child'	瓦 'brick'	袜 'socks'
45. /ni/	妮 'little girl'	霓 'neon'	你 'you'	腻 'sick of'
46. /zhu/	猪 'pig'	竹 'bamboo'	煮 'boil'	住 'dwell'
47. /bao/	包 'bag'		宝 'treasure'	报 'report'
48. /san/	三 'three'		伞 'umbrella'	散 'dismiss'
49. /chao/	抄 'copy'	潮 'tide, humid'	炒 'stir-fry'	
50. /che/	车 'vehicle'		扯 'rip, tear'	撤 'retreat'
51. /cun/	村 'village'	存 'save, deposit'		寸 'inch'
52. /cuo/	搓 'rub against'	矮 'dwarf'		错 'mistake'
53. /dai/	呆 'dull'		歹 'evil, vicious'	带 'bring'
54. /sha/	沙 'sand'		傻 'stupid'	厦 'tall building'
55. /dian/	颠 'bumpy'		点 'point'	电 'electricity'
56. /dao/	刀 'knife'		岛 'island'	到 'arrive'
57. /dang/	裆 'crotch'		党 'party'	荡 'swing'
58. /dan/	丹 'red'		胆 'liver, guts'	淡 'light, dim'
59. /tou/	偷 'steal'	头 'head'		透 'transparent'
60. /gai/	该 'ought to'		改 'correct'	盖 'cap'
61. /gang/	钢 'steel'		港 'harbor'	杠 'bar'
62. /gao/	高 'high'		搞 'do, make'	告 'inform, accuse'
63. /gou/	钩 'hook'		狗 'dog'	构 'build'
64. /gu/	姑 'aunt'		古 'ancient'	固 'fix'
65. /gua/	瓜 'melon'		寡 'widower'	挂 'hang'
66. /ai/		癌 'cancer'		爱 'love'
67. /ban/	班 'class'		版 'format, layout'	
68. /biao/	标 'mark, label'		表 'table, chart'	
69. /bing/	冰 'ice'		饼 'flat bread'	
70. /can/	餐 'meal'		惨 'unfortunate'	
71. /chou/	抽 'draw'		丑 'ugly'	
72. /dong/	东 'east'		懂 'understand'	
73. /duan/	端 'hold, end'		短 'short'	
74. /e/		俄 'Russia'		呃 'uh mm'
75. /guan/	关 'turn off'		馆 'store, gallery'	
76. /guang/	光 'light, just'		广 'widespread'	
77. /gui/	归 'return'		鬼 'ghost'	
78. /hai/	嗨 'hi'	孩 'child'	海 'ocean'	害 'harm'
79. /he/		河 'river'		贺 'congrats'
80. /hen/		痕 'trace'		恨 'resent'

(continued)

(continued)

	Tone 1	Tone 2	Tone 3	Tone 4
81. /hua/		滑 'slide'		话 'word'
82. /huo/		活 'live, alive'		或 'or'
83. /huang/	荒 'barren'		谎 'lie'	
84. /jia/	家 'home'		甲 'armor, first'	
85. /jian/	坚 'firm, strong'		减 'reduce, cut'	
86. /jie/	接 'receive'		姐 'older sister'	
87. /jin/	今 'present'		紧 'tight'	
88. /jiu/	究 'investigate'		酒 'alcohol'	
89. /ju/		局 'bureau'		巨 'giant'
90. /kai/	开 'open'		凯 'triumphant'	
91. /ku/	哭 'cry'		苦 'bitter'	
92. /kuan/	宽 'broad'		款 'section'	
93. /kuang/		狂 'wild'		况 'condition'
94. /lai/		来 'come'		赖 'bilk, rely'
95. /lan/		蓝 'blue'		烂 'rot'
96. /lang/		郎 'man, male'		浪 'wave'
97. /lei/		雷 'thunder'		类 'category'
98. /li/		黎 'dawn'		力 'force, power'
99. /lian/		连 'consecutive'		恋 'crush, love'
100. /liang/		良 'good'		亮 'bright'
101. /liao/		疗 'heal'		料 'ingredient'
102. /ling/		灵 'spirit'		另 'other'
103. /liu/		留 'stay'		六 'six'
104. /lou/		楼 'building'		漏 'leak'
105. /lu/		卢 'a surname'		录 'record'
106. /luo/		罗 'a surname'		洛 'a river'
107. /man/		瞒 'withhold'		慢 'slow'
108. /mei/		梅 'plum flower'		妹 'younger sister'
109. /meng/		盟 'vow, alliance'		梦 'dream'
110. /ming/		名 'first name'		命 'command, life'
111. /na/		拿 'take'		纳 'accept'
112. /nian/		年 'year'		念 'read'
113. /niu/	妞 'little girl'		纽 'button'	
114. /ou/	欧 'europe'		偶 'by chance'	
115. /pa/		爬 'climb'		怕 'fear'
116. /pai/		排 'row, platoon'		派 'send'
117. /pan/		盘 'plate'		判 'judge'

(continued)

(continued)

	Tone 1	Tone 2	Tone 3	Tone 4
118. /pei/		陪 'accompany'		配 'deserve'
119. /qiao/	敲 'knock'		巧 'skillful'	
120. /qian/		钱 'money'		歉 'apology'
121. /qin/	亲 'dear'		寝 'sleep'	
122. /ren/		人 'person'		认 'recognize'
123. /shan/	山 'mountain'		闪 'flash'	
124. /shang/	伤 'injure'		赏 'appreciate'	
125. /shou/	收 'receive'		手 'hand'	
126. /si/	丝 'silk'		死 'die'	
127. /suo/	缩 'shrink'		所 'place, office'	
128. /sui/		随 'let'		岁 'year'
129. /ta/	他 'he, him'		塔 'tower'	
130. /tai/		台 'platform'		太 'too'
131. /tao/		逃 'flee'		套 'trap, harness'
132. /tiao/		条 'strap'		跳 'jump'
133. /tan/	滩 'beach'		坦 'honest'	
134. /tie/	贴 'stick, post'		铁 'iron'	
135. /ting/	厅 'hall'		挺 'very, stand'	
136. /tong/		童 'child'		痛 'hurtful'
137. /tui/	推 'push'		腿 'leg'	
138. /wang/		亡 'perish'	望 'look out'	
139. /wei/	威 'awe'		伟 'great'	
140. /wen/	温 'warm, review'		稳 'steady'	
141. /xian/	先 'before'		险 'risk, danger'	
142. /xiang/	乡 'county'		想 'ponder'	
143. /xiao/	消 'offset'		晓 'dawn'	
144. /xie/	些 'a little'		写 'write'	
145. /xu/	须 'must, beard'		许 'promise'	
146. /xuan/	宣 'announce'		选 'select'	
147. /xun/		寻 'search'		训 'train'
148. /ya/	押 'deposit'		雅 'graceful'	
149. /yan/	烟 'smoke'		演 'acting'	
150. /yang/		阳 'sun, yang'		样 'look'
151. /ye/		爷 'grandpa'		业 'career'
152. /yu/		鱼 'fish'		育 'nurture'
153. /yuan/		园 'garden'		愿 'wish'
154. /yun/		云 'cloud'		孕 'pregnant'

(continued)

(continued)

	Tone 1	Tone 2	Tone 3	Tone 4
155. /yao/	邀 ‘invite’		舀 ‘scoop’	
156. /yin/	音 ‘tone, pitch’		引 ‘lead, bring’	
157. /yong/	拥 ‘own, possess’		永 ‘eternal’	
158. /you/	优 ‘superior’		友 ‘friend’	
159. /zao/	遭 ‘suffer’		早 ‘morning, early’	
160. /zhan/	詹 ‘a surname’		展 ‘stretch’	
161. /zhang/	张 ‘a surname’		掌 ‘palm’	
162. /zhao/	招 ‘incur’		找 ‘seek’	
163. /zhen/	真 ‘true’		诊 ‘diagnose’	
164. /zhi/	知 ‘knowledge’		指 ‘finger, point’	
165. /zong/	宗 ‘ancestor’		总 ‘sum’	
166. /zu/	租 ‘rent’		组 ‘group’	
167. /zuo/		昨 ‘yesterday’		坐 ‘sit’

2. List of characters that served only as distractors

安 ‘peace’	暗 ‘dark’	奥 ‘abstruse’	本 ‘basis, origin’
笨 ‘dumb’	编 ‘weave’	辩 ‘argue’	波 ‘wave’
博 ‘knowledgeable’	补 ‘patch, repair’	布 ‘cloth’	测 ‘measure’
插 ‘insert’	茶 ‘tea’	城 ‘city’	充 ‘fill’
虫 ‘worm’	穿 ‘wear’	船 ‘boat, ship’	聪 ‘smart’
从 ‘follow’	堆 ‘pile’	对 ‘pair’	敦 ‘sincere’
顿 ‘pause’	罚 ‘punish’	法 ‘law’	风 ‘wind’
敢 ‘dare’	根 ‘root’	工 ‘work, industry’	共 ‘public’
乖 ‘docile’	怪 ‘strange’	豪 ‘deluxe’	好 ‘good’
黑 ‘black’	后 ‘after, back’	婚 ‘marriage’	魂 ‘soul’
讲 ‘speak’	酱 ‘sauce’	交 ‘deliver’	叫 ‘scream, yell’
绝 ‘severe, extreme’	康 ‘healthy, good’	抗 ‘resist’	考 ‘exam’
靠 ‘rely’	恐 ‘terror’	控 ‘accuse’	口 ‘mouth’
扣 ‘buckle, deduct’	快 ‘quick’	拉 ‘pull’	辣 ‘spicy’
劳 ‘labor’	老 ‘aging, old’	列 ‘array’	林 ‘forest, woods’
旅 ‘tour, travel’	律 ‘law’	轮 ‘shift, turn’	买 ‘buy’

(continued)

(continued)

迈 ‘step out’	门 ‘door’	民 ‘folk, citizen’	敏 ‘sensitive, allergy’
某 ‘someone’	母 ‘female’	木 ‘wood’	奶 ‘grandma, breast’
男 ‘male’	脑 ‘brain’	闹 ‘noisy, stir up’	努 ‘exert’
怒 ‘angry’	偏 ‘leaning’	骗 ‘deceive’	平 ‘flat’
枪 ‘gun’	墙 ‘wall’	且 ‘and, in addition’	窃 ‘sneaky’
侵 ‘invade’	琴 ‘instrument’	求 ‘beg’	全 ‘complete’
缺 ‘missing’	确 ‘indeed’	群 ‘crowd’	然 ‘so, then, correct’
染 ‘dye’	扰 ‘disturb’	绕 ‘coil, detour’	惹 ‘annoy’
热 ‘hot’	扔 ‘throw’	仍 ‘still’	容 ‘look’
弱 ‘weak’	烧 ‘burn, boil’	绍 ‘a place name’	水 ‘water’
睡 ‘sleep’	松 ‘pine, loose’	送 ‘send, farewell’	苏 ‘a surname’
诉 ‘appeal’	糖 ‘sugar’	躺 ‘lie on back’	天 ‘sky’
田 ‘filed’	托 ‘ask, hold’	玩 ‘play’	晚 ‘evening, late’
我 ‘I’	卧 ‘lie, crouch’	瞎 ‘blind’	下 ‘down’
心 ‘heart’	信 ‘letter’	凶 ‘unfortune’	熊 ‘bear’
休 ‘recess’	秀 ‘elegance, show’	月 ‘moon’	杂 ‘mixed’
灾 ‘disaster’	再 ‘again’	责 ‘blame’	者 ‘person’
这 ‘this’	钟 ‘bell’	众 ‘crowd, many’	装 ‘outfit, fake’
撞 ‘bump, crash’	资 ‘capital, income’	字 ‘letter, word’	走 ‘walk’
奏 ‘play’	嘴 ‘mouth’	罪 ‘sin’	尊 ‘respect’
杯 ‘cup’	北 ‘north’	初 ‘beginning’	厨 ‘kitchen, cook’
触 ‘touch, tactile’	窗 ‘window’	闯 ‘break-in’	春 ‘spring’
蠢 ‘dumb’	灯 ‘light, lamp’	等 ‘wait, equal’	丁 ‘a surname’
顶 ‘peak’	东 ‘east’	懂 ‘understand’	经 ‘pass’
景 ‘view’	轻 ‘light, gentle’	情 ‘emotion’	请 ‘please’
庆 ‘celebrate’	叔 ‘uncle’	熟 ‘familiar’	鼠 ‘mouse’
树 ‘tree’	摔 ‘fall’	甩 ‘swing, throw’	双 ‘double, pair’
爽 ‘feel good’	星 ‘star’	醒 ‘awake’	婴 ‘infant, baby’
赢 ‘win’	影 ‘film, shadow’	硬 ‘hard’	争 ‘argue, fight’

(continued)

(continued)

整 ‘whole’	白 ‘white’	败 ‘defeated’	末 ‘ending’
朋 ‘friend’	碰 ‘touch, bump’	婆 ‘old woman’	迫 ‘involuntary’
绳 ‘rope’	圣 ‘saint’	闲 ‘free, available’	陷 ‘trap’
学 ‘study, learn’	血 ‘blood’	油 ‘oil’	右 ‘right’
肮 ‘filthy’	昂 ‘chin-up’	崩 ‘collapse, burst’	宾 ‘guest’
擦 ‘wipe’	仓 ‘barn, storage’	层 ‘layer, floor’	拆 ‘break apart’
柴 ‘firewood’	缠 ‘tangle’	产 ‘produce’	忤 ‘confess’
陈 ‘a surname’	衬 ‘contrast’	吹 ‘blow’	垂 ‘droop’
凑 ‘collect’	粗 ‘thick’	促 ‘urge’	崔 ‘a surname’
脆 ‘crispy’	德 ‘virtue’	雕 ‘sculpture’	掉 ‘drop’
爹 ‘dad’	碟 ‘plate’	丢 ‘throw’	兜 ‘pocket, cover’
抖 ‘shake’	豆 ‘bean’	恩 ‘mercy’	而 ‘but’
耳 ‘ear’	尬 ‘embarrass’	滚 ‘roll, gone’	棍 ‘stick’
航 ‘ship, navigate’	亨 ‘go smoothly’	恒 ‘eternal’	轰 ‘boom’
红 ‘red’	怀 ‘chest, mind’	坏 ‘bad, evil’	捐 ‘donate’
倦 ‘tiring’	咖 ‘brown, coffee’	刊 ‘journal’	砍 ‘chop’
看 ‘look, view’	肯 ‘willing’	坑 ‘pit, hole’	夸 ‘compliment’
垮 ‘collapse’	跨 ‘step over’	亏 ‘owe, deficit’	奎 ‘a surname’
溃 ‘be routed’	昆 ‘offspring’	困 ‘drowsy’	扩 ‘expand’
冷 ‘cold’	俩 ‘two’	龙 ‘dragon’	侣 ‘couple, pair’
虑 ‘consider’	略 ‘omit, rough’	乱 ‘messy’	忙 ‘busy’
棉 ‘cotton’	面 ‘flour’	灭 ‘extinguish’	谬 ‘absurd’
内 ‘inner, inside’	能 ‘capable’	娘 ‘mother’	您 ‘you (honorific)’
柠 ‘lime, citrus’	农 ‘farm’	虐 ‘abuse’	暖 ‘warm’
挪 ‘move’	诺 ‘promise’	噢 ‘ah’	旁 ‘beside, next’
抛 ‘throw’	盆 ‘basin, tub’	恰 ‘just’	穷 ‘poor’
嚷 ‘yell’	让 ‘yield’	日 ‘sun’	柔 ‘soft, gentle’
肉 ‘meat, flesh’	如 ‘like’	辱 ‘shame, disgrace’	入 ‘enter’
软 ‘soft, flexible’	瑞 ‘lucky’	润 ‘moist, profit’	洒 ‘sprinkle, spill’
萨 ‘a surname’	赛 ‘competition’	桑 ‘mulberry’	嗓 ‘throat’
骚 ‘upset’	瑟 ‘an instrument’	森 ‘full of trees’	晒 ‘shine upon’

(continued)

(continued)

耍 ‘play, fool’	顺 ‘fluent, smooth’	搜 ‘search’	酸 ‘sour’
算 ‘calculate’	孙 ‘grandchildren’	损 ‘decrease, harm’	特 ‘special’
疼 ‘ache, hurt’	团 ‘round, unite’	吞 ‘swallow, stutter’	外 ‘outside’
翁 ‘old man’	赞 ‘praise’	葬 ‘bury’	贼 ‘thief’
怎 ‘how’	增 ‘increase’	赠 ‘give’	摘 ‘pick’
宅 ‘house, mansion’	债 ‘debt’	抓 ‘grab, hold’	爪 ‘claw, paw’
专 ‘specific’	坠 ‘fall down’	准 ‘permit’	桌 ‘table’
耐 ‘endure’	谋 ‘plot, conceive’	军 ‘military’	

References

Alloppenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38, 419–439.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistics Software*, 67(1), 1–48.

Cai, Z. G., Pickering, M. J., Yan, H., & Branigan, H. P. (2011). Lexical and syntactic representations in closely related languages: Evidence from Cantonese-Mandarin bilinguals. *Journal of Memory and Language*, 65, 431–445.

Da, J. (2005). *Character frequency list of Modern Chinese*. <https://lingua.mtsu.edu/chinese-computing/statistics/char/list.php?Which=MO>

Dalmajjer, E., Mathôt, S., & Van der Stigchel, S. (2014). PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-013-0422-2>.

Ju, M., & Luce, P. (2004). Falling on sensitive ears: Constraints on bilingual lexical activation. *Psychological Science*, 15(5), 314–318.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>.

Malins, J. G., & Joanisse, M. F. (2010). The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study. *Journal of Memory and Language*, 62(4), 407–420.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, 44(2), 314–324.

McQueen, J. M., & Viebahn, M. C. (2007). Tracking recognition of spoken words by tracking looks to printed words. *Quarterly Journal of Experimental Psychology*, 60(5), 661–671. <https://doi.org/10.1080/17470210601183890>.

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Chapman and Hall/CRC.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, 59, 475–494.

Mitterer, H., & Russell, K. (2013). How phonological reductions sometimes help the listener. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 39(3), 977–984. <https://doi.org/10.1037/a0029196>.

- Norman, J. (1988). *Chinese*. Cambridge: Cambridge University Press.
- R Core Team (2020). R: A language and environment for statistical computing (version 4.0.0) [computer software]. R Foundation for Statistical Computing, Vienna, Austria. Available from <https://www.R-project.org/>.
- Schirmer, A., Tang, S., Penney, T. B., Gunter, T. C., & Chen, H. (2005). Brain responses to segmentally and tonally induced semantic violations in Cantonese. *Journal of Cognitive Neuroscience*, 17(1), 1–12.
- Shook, A., & Marian, V. (2016). The influence of native-language tones on lexical access in the second language. *The Journal of the Acoustical Society of America*, 139, 3102–3109.
- Stanford, J. (2008). A sociotoneic analysis of Sui dialect contact. *Language Variation and Change*, 20(3), 48–81.
- Wang, X., Wang, J., & Malins, J. G. (2017). Do you hear ‘feather’ when listening to ‘rain’? Lexical tone activation during unconscious translation: Evidence from Mandarin-English bilinguals. *Cognition*, 169, 15–24.
- Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, 50(1), 1–25.
- Wurm, S. A., Li, Baumann, T., & Lee, M. W. (1987). *Language Atlas of China*. Longman, ISBN 978-962-359-085-3.
- Xu, D. (2015). The role of geography in the northwest China linguistic area. In D. Xu & J. Fu (Eds.), *New studies of space and quantification in languages of China* (pp. 57–73). Springer.
- Zhang, A. S. (1990). *Yinchuan fangyan yangping shangsheng herbingshi tanxi* [The exploration to the history of level and falling-rising merger in Yinchuan dialect]. *Guyuan shizhuan xuebao*, 2 [Journal of Guyuan Normal College].
- Zhang, Y. L. (2009). *Lanyin guanhua yuyin yanjiu* [The phonetic study of Lan-Yin Mandarin]. Beijing: Beijing Language and Culture University Press.

The Effect of Perceptual Training on Teaching Mandarin Chinese Tones



Yingjie Li and Goun Lee

Abstract Computer-assisted perceptual training is effective in learning Mandarin tones both in monosyllabic and disyllabic words, but disyllabic tone training is nearly twice as effective as monosyllabic tone training (Li et al., *A sound approach to language matters in honor of Ocke-Schwen Bohn*. Aarhus University Library/Royal Danish Library—AU Library Scholarly Publishing Service, pp. 303–319, 2019). The present study analyzed the tone identification performance of Mandarin learners on both monosyllabic and disyllabic stimuli to provide explicit, meaningful information for teaching and learning tones. The results showed that all learners identified T4 significantly better than T1, T2, and T3 in the monosyllabic stimuli. In the disyllabic stimuli, T3 in the first syllable position was the most difficult to perceive, followed by T2, T1, and T4; in the second syllable position, T2 was the most problematic, followed by T1, T3, and T4. In addition, tone errors in both types of stimuli were analyzed to find the easier and the more difficult tone pairs after training. Finally, the learners were found to be generally proficient at perceiving tones in the final but not initial syllable, in compatible but not in conflicting tonal context, and in same tonal sequence but not in different tonal sequence. These findings provide evidence for successful tone learning through the increased utilization of disyllabic stimuli in Chinese language teaching.

Keywords Computer-assisted learning · Perceptual training · Mandarin Chinese · Tone learning

Y. Li (✉)

Department of Asian Languages and Civilizations, University of Colorado Boulder, Boulder, CO, USA

e-mail: yingjie.li@colorado.edu

G. Lee

College of Liberal Arts, Dankook University, 152 Jukjeon-ro, Sugi-gu, Yongin-si, Gyeonggi-do, South Korea

1 Introduction

Mandarin Chinese is a tonal language: every Chinese character has a tone. Tone is a key component of the lexicon in the language. There are four phonemic tones in Mandarin that native speakers use to distinguish word meaning. Thus, perceiving and producing tones correctly is of critical importance for all Mandarin language learners in order to communicate successfully in the target language. For those learners whose native language is non-tonal, the ability to categorically acquire lexical tones is challenging, since these phonemic tones are not in their lexicon, (Shen and Froud 2016).

Computer-assisted short-term perceptual training has been shown effective in assisting learners to acquire new phonetic contrasts that do not exist in their native phonological language system (Logan et al. 1991; Lively et al. 1993; Wang et al. 1999; Kingston 2003; Francis et al. 2008; Herd et al. 2013; Li et al. 2019). In these studies, after a short period of perceptual training, the learners' perception of a target sound that was not in their native language system was significantly improved. The target languages investigated in these studies included English, Chinese, German, Cantonese, and Spanish. Wang et al. (1999) did the first suprasegmental perceptual training using tonal contrasts in Mandarin to train native English-speaking Mandarin learners (hereafter "English-speaking learners") to identify isolated tones in monosyllabic Chinese words. The beginning learners who received perceptual training all improved significantly in tonal perception of the four phonemic tones in monosyllabic words when compared to those in the control group who did not receive training. Unfortunately, that study did not address the issue of whether monosyllabic tone training would help the learners identify tones in disyllabic words, nor did it conduct perceptual training by using disyllabic words. To fill this gap, Li et al. (2019) extended the perceptual training by using both monosyllabic and disyllabic tones to train seventeen beginning-level, English-speaking learners in two groups to identify tones in both monosyllabic and disyllabic words. They found that, in general, both types of training were successful in assisting the learners' tone identification. More importantly, the learners showed greater improvement after the disyllabic training than the monosyllabic training in identifying tones in both monosyllabic and disyllabic words. The current study aims to analyze the tone identification performance of these seventeen learners in response to both monosyllabic and disyllabic stimuli. The learners' tonal confusion in both types of stimuli was examined in hopes that the findings would benefit Chinese language teachers and learners.

2 Literature Review

2.1 Tones in Mandarin Chinese

Tone is a suprasegmental feature of the Chinese language. As shown in Table 1, the four phonemic tones of Mandarin can be perceptually distributed on a five-point pitch scale that provides direct visual representation of the pitch contours. The four tones are usually indicated by four diacritic marks in *pinyin*, as illustrated in the example column for the syllable //u/ in Table 1. In terms of tonal classification, T1 is a level tone due to its relatively consistent high pitch value at 55, while T2, T3 and T4 are contour tones that contain a rising and/or falling pitch within a syllable with pitch values of 35, 214, and 51, respectively (Chao and Pian 1955). The tones in Table 1 are all in canonical form, which means that the tones in such isolated environment are quite stable in pronunciation, and easier to perceive.

Mandarin tones often undergo alternation when produced in connected speech. This tonal coarticulation is quite common in disyllabic words. For example, the T3 Sandhi rule stipulates that when T3 is followed by another T3, the first T3 changes to a rising T2. In other non-final positions, when preceding any tone other than T3, T3 is pronounced as a low tone with a pitch value of 21—without the final rise that occurs when the tone is produced at the end of a prosodic phrase or in isolation. Also, T4 changes to a high-mid tone with a pitch value of 53 in connected speech (Lin 2007). Xu (1994, 1997) systematically examined Chinese native speakers' perception and production of coarticulated tones. After an investigation of tonal variation in naturally produced tri-syllable Chinese words, he proposed the concept of “compatible” versus “conflicting” tonal contexts, in which the pitch value of one tone is affected by adjacent tone. In compatible contexts, such as T1 (55) + T4 (51), adjacent tones share identical or similar pitch values at the syllable boundary, while in conflicting contexts, such as T1(55) + T2 (35), adjacent tones differ substantially in their pitch values (Xu 1994). Xu discovered that Chinese native speakers use information from the coarticulation of adjacent tones to help identify the target tones correctly. He concluded that there was greater carryover than anticipatory tonal coarticulation in tri-syllabic words and phrases. This carryover effect was confirmed by

Table 1 Descriptions of the four Mandarin phonemic tones, pitch values, and examples

Tone	Description	Pitch value	Example
1	High level	55	lū “sound of grumbling or chattering”
2	High rising	35	lú “stove”
3	Dipping/falling-rising	21(4)	lǔ “to brine”
4	High falling	51	lù “road”

Xu (1997). However, Xu's findings conflicted with those of Shen (1990), who also studied Mandarin tri-syllables and found symmetrical bi-directional effects, which suggests that the carryover effect among adjacent tones is similar to the anticipatory effect. These studies of tones demonstrate that tonal coarticulation differs depending on the tonal environment. Learning only monosyllabic tones as has been examined in many previous studies can merely provide a partial picture of tone learning in Mandarin Chinese. Disyllabic tone perception must also be examined in order to simulate the variability of natural speech more accurately.

2.2 English-Speaking Learners' Perception of Tones in Monosyllabic and Disyllabic Words

Many studies have examined English-speaking learners' perception and production of Mandarin tones in monosyllabic words. These learners were found to have difficulty perceiving and producing tones in general, especially T2 and T3 in monosyllabic words, mainly because the phonemic tone feature is not in part of their native language system (Miracle 1989; Shen 1989; Shen and Lin 1991; Sun 1998; Jongman et al. 2006; Lee et al. 2010a, b; Chang 2011; Hao 2012).

Only a few studies have investigated English-speaking learners' perceptions of disyllabic words (Sun 1998; He 2010; Hao 2012; He and Wayland 2013). He (2010), He and Wayland (2013) and Sun (1998) investigated the relationship between linguistic experience/proficiency levels and tonal perceptions of both monosyllabic and disyllabic words in Mandarin by English-speaking learners, and their findings align with each other. They found that across learning experience and proficiency levels, English-speaking learners did significantly better at identifying tones in monosyllabic words than in disyllabic words. Moreover, the higher the proficiency level or the longer the English-speaking learners had studied Mandarin, the more accurate their tone identification was. Sun (1998) found that identification of T2 and T3 across proficiency levels was significantly worse than T1 and T4 across all four groups of different proficiency levels in both monosyllabic and disyllabic words. Similarly, He (2010) found that across both monosyllabic and disyllabic tonal contexts, for inexperienced learners, T3 was the most difficult to identify, followed by T1, T2 and T4, while T2 was the most difficult of the four tones for experienced learners to identify. Sun (1998) analyzed tones in the initial and final syllable positions and found that the accuracy rate of tone identification in the final position was higher than in the initial position in all disyllabic words. T1 and T4 were identified with higher accuracy at both the initial- and final- position than T2 and T3 in disyllabic words. This finding resonates with that of He (2010) and He and Wayland (2013), which examined the tonal coarticulation of disyllabic words in compatible and conflicting tonal contexts (proposed by Xu 1994) through the use of tone identification tasks for English-speaking learners. They found that the learners' tonal perception of disyllabic words was significantly better in compatible than in conflicting contexts, and

that of the four tones, T3 was still the most difficult to identify in both tonal environments. Bi-directional T2–T3 confusion was also observed by Hao (2012) in American learners' identification tasks in both monosyllabic and disyllabic Mandarin words. Hao attributed the major difficulty to American learners' lack of association between the pitch of a tone and its corresponding tonal category. From these studies, one can see that building the English-speaking learners' tonal categories is vital to achieving native-like pronunciation of the target language.

2.3 High Variability Phonetic Training

High variability phonetic training has been proven to be an effective method for improving learner perception and the production of segmental properties in the target language. It was first proposed by Logan et al. (1991) and included three important aspects: stimuli were presented in a variety of phonetic environments; natural speech tokens were used instead of synthesized ones; and multiple speakers were used. These characteristics converged to enable listeners to form robust phonetic categories by increasing stimulus variability (Logan et al. 1991; Lively et al. 1993). Logan et al. (1991) found that native Japanese-speakers learned to perceive the English segment contrast /l/ and /r/ after a short three-week high variability training. Wang et al. (1999) successfully used this type of perceptual training to train their American learners to identify tones in monosyllabic Mandarin words. The trainees performed significantly better on all tests (pretest, posttest and two generalization tests) than participants in the control group, who had not received any training but only the same in-classroom instruction as the trainee group. This study demonstrates that perceptual training is highly effective at the suprasegmental level, such as for Mandarin tones.

Li et al. (2019) were the first to extend this high variability suprasegmental perceptual training from isolated tones in monosyllabic words to connected tones in disyllabic words. Adopting all the monosyllabic training stimuli from Wang et al. (1999), Li et al. had seventeen beginning-level English-speaking learners randomly divided into two groups: a monosyllabic training group and a disyllabic training group. For a short two-week period, monosyllabic training group trained only in monosyllabic stimuli while disyllabic training group trained only in disyllabic stimuli. After the pretest, training, and posttest, the tone identification performance for both monosyllabic words and disyllabic words of the two groups of learners was statistically compared. The results showed that overall, for beginning learners, the accuracy rate of tone identification increased significantly from pretest (72%) to posttest (80%). The improvement of the disyllabic training group, however, nearly doubled (11%) in overall tone identification compared to those in the monosyllabic training group (6%). This suggests that disyllabic training facilitated more effective tone learning for learners identifying tones in both monosyllabic and disyllabic words than it did for those in the monosyllabic training group. These findings indicate that when teaching the tones in Mandarin, it is more helpful for teachers to use disyllabic rather than mainly monosyllabic words because the disyllabic word exposure provides learners

with more tonal variability, and, crucially, encourages learners to develop more robust tonal categories.

The current study presents an analysis of the tone identification performance and improvement in the four phonemic tones in Mandarin by English-speaking learners based on the established high-variability perceptual training experiment by Li et al. (2019). The learners' tone identification errors in both monosyllabic and disyllabic words were identified and analyzed. Also, their tonal perception of disyllabic stimuli was examined through linguistic factors such as syllable position, tonal context, and tonal sequence. To understand the learners' tonal behavior before and after perceptual training, the following research questions were posed:

1. Which tones are improved in monosyllabic and disyllabic words after high-variability phonetic perceptual training?
2. What are the problematic tones, as well as the easily confused tone pairs in monosyllabic and disyllabic words for English-speaking learners in high-variability phonetic perceptual training?
3. During high-variability phonetic perceptual training, is the tone perception of disyllabic Mandarin words by English-speaking learners affected by linguistic factors? Specifically, is the tone perception of learners affected by syllable position, tonal context, or tonal sequence?

3 Methodology

Three phases were included in the perceptual tone training experiment: a pretest, a training session (either monosyllabic or disyllabic training), and a posttest. All Mandarin learners participated in identical pretests and posttests, with a forced-choice identification (ID) task. For the pretest and the posttest, both monosyllabic stimuli and disyllabic stimuli were used. Both monosyllabic training and disyllabic training consisted of four perceptual sessions. The monosyllabic training group was trained exclusively with monosyllabic stimuli while the disyllabic training group was trained exclusively with disyllabic stimuli. In all the sessions and for both groups, immediate feedback was given after each response. The two groups were compared across the pretest and posttest to observe any improvement after the training and their tonal errors were analyzed to find patterns in the tonal behavior before and after training.

3.1 Participants

Seventeen English-speaking learners of Mandarin participated in a two-week training program. All were beginning-level college-student learners with fewer than two semesters (no more than 7 months) of Mandarin study. Nine participated in the monosyllabic training group, and eight in the disyllabic group. The participants were

randomly assigned to one of the two training groups. None of the seventeen learners had a history of hearing, speech, or language difficulties.

3.2 *Stimuli*

To ensure speaker variability, all the stimuli were recorded by six (three male and three female) native Mandarin speakers. Two types of stimuli, monosyllabic and disyllabic, were used throughout the pretest, training, and posttest. All the monosyllabic stimuli were adopted from Wang et al. (1999). These included all possible permissible combinations of various initials and finals and different syllabic structures in Mandarin (i.e. Vowel, ConsonantV, CVNasal, VN, CGlideV, and CGVN). To ensure the monosyllabic and disyllabic training were comparable, each disyllabic stimulus was composed of two randomly combined syllables from the monosyllabic stimuli. In this way, each individual syllable used for the disyllabic stimuli was identical to those used in the monosyllabic stimuli. In addition, the disyllabic stimuli were essentially very low-frequency disyllabic words and non-words in Mandarin. This was meant to ensure there was little or no learning effect to influence the learners' tonal performance in this training. For example, the monosyllabic stimuli—*mǎ* (horse) and—*shāng* (injury) were combined to form a two-syllable word that served as a disyllabic stimulus,—*mǎ shāng*. To preserve the characteristics of the disyllabic words in connected speech, all six speakers were instructed to produce the stimuli as naturally as possible, and to avoid producing any disyllabic stimuli as two separate, individual syllables. In total, 288 monosyllabic stimuli and 144 disyllabic stimuli were used in the perceptual training.

3.3 *Procedure*

The experiment consisted of three phases: pretest, training, and posttest. All the tests and training were conducted in a university phonetics and psycholinguistics laboratory. The stimuli were all presented over headphones using Paradigm software (Tagliaferri 2008) and the learners' responses were all recorded in Paradigm. Seventeen English-speaking learners participated in the two-week training program, with training on six days of each week. There were three sessions held in the first week (Pretest, Training 1, and Training 2) and the three other sessions were held in the second week (Training 3, Training 4, and Posttest). The pretest and posttest were each 60 min long and each training session was 30 min long.

3.3.1 Pretest

The pretest consisted of two parts, monosyllabic word identification and disyllabic word identification. All stimuli were produced by a male Mandarin native speaker (Speaker 1). For both parts, learners indicated which tones they heard. No feedback was provided. The pretest lasted about 60 min with approximately 30 min for each part.

For the monosyllabic pretest, the learners heard a monosyllabic stimulus and were instructed to give their tone identification response by pushing the corresponding button that represented one of the four tones (1 = Tone 1, 2 = Tone 2, 3 = Tone 3, and 4 = Tone 4). All tonal diacritics and numbers were labeled on the buttons on the keyboard. There were 96 monosyllabic stimuli in the pretest, all of which were the same as those used in the Wang et al. (1999) study. There were 24 monosyllabic words for each of the four phonemic Mandarin tones. All monosyllabic stimuli were presented with a 3 s inter-trial interval (ITI). The learners' accuracy during the identification task was recorded in Paradigm (Tagliaferri 2008). For the disyllabic pretest, the learners heard a disyllabic stimulus and they were asked to indicate their tone identification response by pushing, in order, the two corresponding buttons that represented the tone of the first syllable, followed by the tone of the second syllable (1 = Tone 1, 2 = Tone 2, 3 = Tone 3, and 4 = Tone 4). All tonal diacritics and numbers were labeled on the buttons on the keyboard. There were 48 disyllabic stimuli in the pretest, each of which was composed of two randomly combined syllables from the monosyllabic stimuli. Thus, every individual syllable used for the disyllabic stimuli could be found in those used as monosyllabic stimuli. The purpose of keeping both types of stimuli identical is to ensure the valid comparison between the two training groups. There were three disyllabic words for each of the 16 combinations (4 tones \times 4 tones = 16 pairs). To directly compare identification of the disyllabic and monosyllabic stimuli, accuracy for each syllable of the disyllabic stimuli was tabulated. Thus, if T1 + T4 was presented and the response was T2 + T4, the first syllable was recorded as incorrect and the second syllable was recorded as correct. Also, due to a productive third tone sandhi rule in Mandarin, for one of the sixteen pairs (Tone3 + Tone3), the first Tone 3 syllable is systematically produced as a Tone 2 when followed by a Tone 3 syllable. For these stimuli, the correct identification was Tone2 + Tone 3. The ITI was 3 s as well. All disyllabic tonal diacritics and numbers were labeled on the keyboard, and no feedback was given. Learners' accuracy in the identification task was also recorded in Paradigm (Tagliaferri 2008).

3.3.2 Training Sessions

Both the monosyllabic and disyllabic training consisted of four perceptual training sessions lasting 30 min each. The learners participated in a forced-choice identification task and immediate feedback was given after each response for all training sessions to help them focus their attention on the critical acoustic cues of the four tones.

3.3.3 Posttest

The posttest was identical to the pretest for both monosyllabic stimuli and disyllabic stimuli. The learners indicated which tones they heard by pushing the corresponding button for the four tones (1 = Tone 1, 2 = Tone 2, 3 = Tone 3, and 4 = Tone 4). They received no feedback. The posttest lasted about 60 min, approximately 30 min for each part.

3.3.4 Data Analysis

The statistical design of the present study had one dependent variable: the tone identification accuracy rate which included both the monosyllabic and the disyllabic stimuli tonal accuracy rate. There were four independent variables: the tests (pretest and posttest), the training groups (monosyllabic and disyllabic), the stimuli (monosyllabic and disyllabic), and the tones (T1, T2, T3, and T4). Analysis of the independent variables was conducted to determine if there were significant differences between the two training groups in the identification of the two types of stimuli from pretest to posttest.

A repeated measures ANOVA and Paired Sample t-test were used in the study to compare the accuracy of the learners' responses on the tests. All statistical analyses were performed using SPSS software. All p -values and the F -values were adjusted using the Greenhouse–Geisser correction (Greenhouse and Geisser 1959), and the post-hoc pairwise comparisons and paired t-tests were adjusted using the Bonferroni correction ($p < 0.05$). All significant results were reported.

4 Results

Research Question 1. Which tones are improved in monosyllabic and disyllabic words after high-variability phonetic perceptual training?

4.1 Tones in the Monosyllabic Stimuli by the Two Training Groups

Identification of the four individual tones in the monosyllabic stimuli in the pretest and posttest by English-speaking learners is presented in Fig. 1 (monosyllabic training group) and Fig. 2 (disyllabic training group).

A three-way repeated measures ANOVA with accuracy as a dependent variable was conducted. Test (pretest and posttest) and Tone (T1, T2, T3, T4) were used as between-subjects independent variables and Training Group (monosyllabic and disyllabic) was used as a between-subjects independent variable. The results revealed the main effect of Test [$F(1,15) = 12.653, p = 0.003$], suggesting that across groups, learners were significantly better at identifying all four tones in the monosyllabic stimuli on the posttest (90%) than on the pretest (84%) after training. The main effect of Tone [$F(3,45) = 8.221, p < 0.001$] was also found, indicating that there was a significant difference among the four tones in the monosyllabic stimuli. A post hoc pairwise comparison with Bonferroni correction revealed that, in the monosyllabic stimuli, T4 (96%) was significantly better than T1 (86%) ($p = 0.029$), T2 (84%) ($p = 0.005$), and T3 (84%) ($p < 0.001$). There were no significant differences among T1, T2 and T3 ($p > 0.999$).

No main effect of Training Group [$F(1,15) = 1.022, p = 0.328$] was found, nor were there any two-way or three-way interactions ($p > 0.1$).

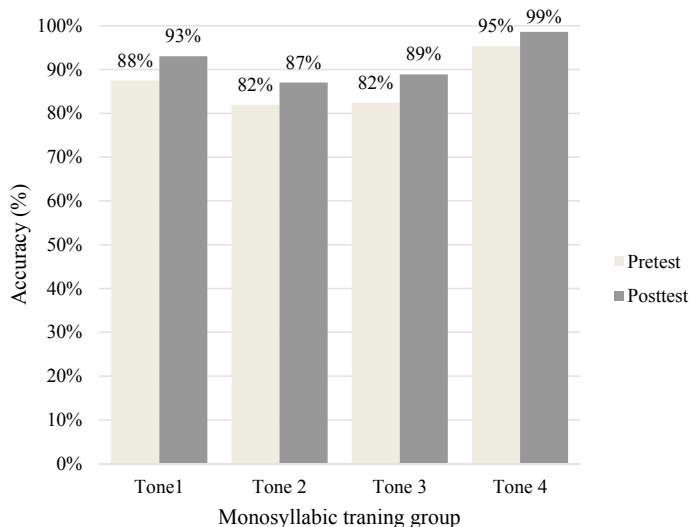


Fig. 1 Average accuracy of the monosyllabic stimuli by the English-speaking learners in the monosyllabic training group for the pretest and posttest

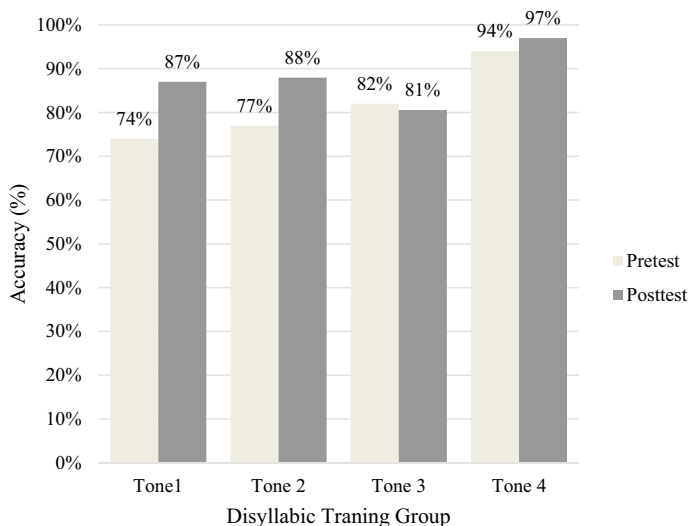


Fig. 2 Average accuracy of the monosyllabic stimuli by the English-speaking learners in the disyllabic training group for the pretest and posttest

Overall, both training groups made significant increases from the pretest to the posttest, demonstrating that both types of training can help improve tone identification in general. The difference between the two groups is that the disyllabic training group learners did significantly better on T4 than the other three tones in the monosyllabic stimuli. Moreover, the disyllabic training group learners' tonal perception of T1 improved significantly after training. However, there was no significant difference in improvement for individual tones after the monosyllabic training. It is worth noticing that the T4 accuracy rates in the monosyllabic stimuli across both training groups were the highest among the four tones on the pretests (95 and 94%), and after training, the accuracy rates were near ceiling effect at 99 and 97%. The fact that T4 had started in such a high position may account for the lack of significant improvement in T4 after training.

4.2 Tones in the Disyllabic Stimuli by the Two Training Groups

Because each disyllabic stimulus has two syllables with two tones, for example, *má hù* is comprised of *má* (σ_1) and *hù* (σ_2), the results below were analyzed to determine the learners' tonal performance for each syllable position (σ_1 , σ_2).

4.2.1 Individual Tones in the First Syllable Position (σ_1)

Figures 3 and 4 display the tone identification of the first syllable position in the disyllabic stimuli by the English-speaking learners in the two training groups.

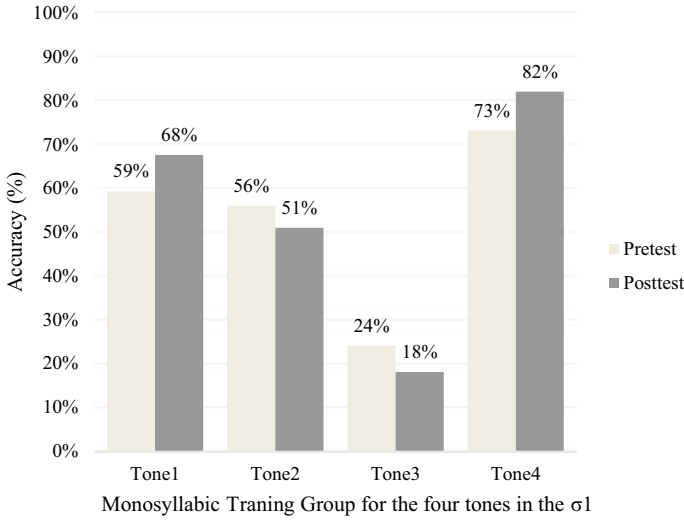


Fig. 3 Average accuracy in the first syllable position (σ_1) of the disyllabic stimuli by the English-speaking learners in the monosyllabic training group

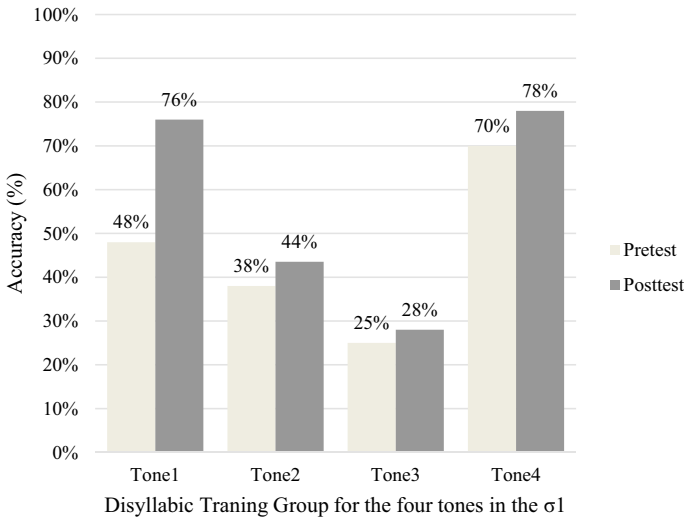


Fig. 4 Average accuracy in the first syllable position (σ_1) of the disyllabic stimuli by the English-speaking learners in the disyllabic training group

Results of a three-way repeated measures ANOVA, with Test (pretest and posttest) and $\sigma 1_Tone$ (Tone1, Tone2, Tone3, Tone4) as the within-subjects factors, and Training Group (monosyllabic and disyllabic) as the between-subjects factor, yielded a main effect of Test [$F(1,15) = 6.531, p = 0.022$], indicating that the learners across both training groups did significantly better on the posttest (56%) than the pretest (49%) in tone identification of the first syllable position in the disyllabic stimuli. It also yielded a main effect of $\sigma 1_Tone$ [$F(3,45) = 30.913, p < 0.001$]. Pairwise comparisons with the Bonferroni correction showed that in the tests, across tones in the first syllable position, the accuracy rates of T1 (62%), T2 (47%), and T4 (76%) were significantly higher than that of T3 (24%) (with $p \leq 0.001$). T4 identification was also significantly better than T2 ($p = 0.001$), and T1 was marginally better than T2 ($p = 0.105$). There was no significant difference between T1 and T4 ($p = 0.124$). In other words, T3 was the most difficult to identify among all four tones in the first syllable position ($\sigma 1$) by learners in both training groups. The Test \times $\sigma 1_Tone$ interaction was also notable [$F(3,45) = 3.309, p = 0.028$], suggesting that there was significant improvement in the tones in the first syllable position after training. From pretest to posttest, across both training groups, for T1 there was an 18% increase, from 54 to 72% ($p = 0.009$); there was no change for T2, with accuracy rates of 47% ($p > 0.99$); T3 dropped 1% in accuracy from 24 to 23% ($p = 0.72$); and T4 made a marginally significant increase of 8% from 72 to 80% ($p = 0.062$). No other two- or three-way interactions were found.

4.2.2 Individual Tones in the Second Syllable Position ($\sigma 2$)

Figures 5 and 6 below illustrate the tone identification of the second syllable position in the disyllabic stimuli by English-speaking learners in the two training groups from pretest to posttest.

A three-way repeated measures ANOVA, with Test (pretest and posttest) and $\sigma 2_Tone$ (T1, T2, T3, T4) used as within-subjects factors, and Training Group (monosyllabic and disyllabic) as a between-subjects factor revealed a significant main effect of Test [$F(1,15) = 9.880, p = 0.007$]. This shows that, averaged across the two training groups and the four tones in the second syllable position, the learners did significantly better on the posttest with a 73% accuracy rate compared to a 67% accuracy rate on the pretest. The main effect of $\sigma 2_Tone$ [$F(3,45) = 5.354, p = 0.003$] suggests that there were significant differences among the four tones. The accuracy rates, from high to low, were: 80% for T4; 72% for T3; 69% for T1; and 58% for T2. The post hoc pairwise comparison shows that there was a significant difference between T4 and T2 ($p = 0.007$). However, there was no difference between T4 and T3 ($p = 0.459$), T4 and T1 ($p = 0.099$), T3 and T1 ($p > 0.999$), T1 and T2 ($p = 0.426$), and T2 and T3 ($p = 0.381$). A main effect of Training Group [$F(1,15) = 5.317, p = 0.036$] shows that, across the two tests, the learners in the monosyllabic training group did significantly better on identification of the tone in the second syllable position with a 77% accuracy rate than the learners in the disyllabic training group who had a 62% accuracy rate. Considering monosyllabic learners had higher starting

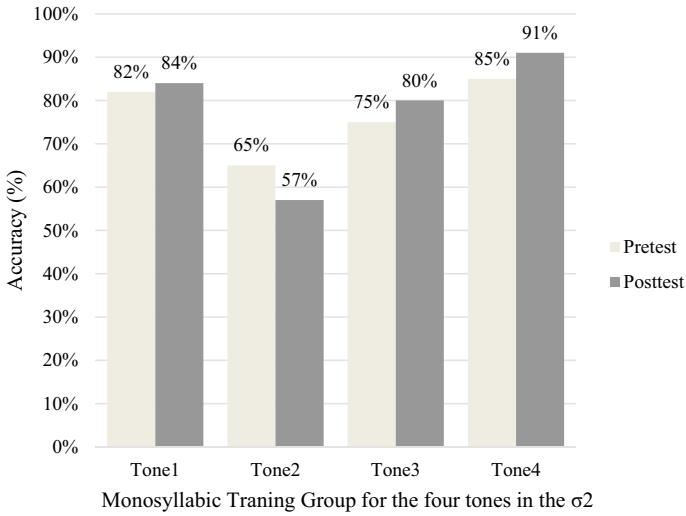


Fig. 5 Average accuracy in the second syllable (σ_2) of the disyllabic stimuli for the English-speaking learners in the monosyllabic training group on the pretest and posttest

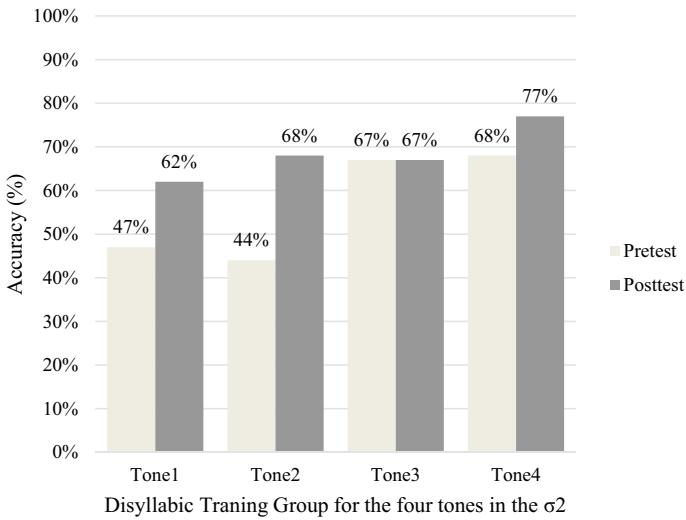


Fig. 6 Average accuracy in the second syllable (σ_2) of the disyllabic stimuli for English-speaking learners in the disyllabic training group on the pretest and posttest

accuracy rates of all four tones in pretest when comparing to disyllabic learners in the second syllable position, this better identification was not a surprise anymore. However, a significant interaction was found between Test \times Training Group [$F(1,15) = 7.200, p = 0.017$]. When this pretest to posttest interaction is broken down, there was substantially greater improvement (13%) by the disyllabic training group from 56 to 69% for the second syllable position compared to the 1% increase, from 77 to 78%, by the monosyllabic training group. There was no other interaction.

Research Question 2. *What are the problematic tones as well as the easily confused tone pairs in monosyllabic and disyllabic words for English-speaking learners in high-variability phonetic perceptual training?*

4.3 Tonal Confusion in the Monosyllabic Stimuli by the Training Groups

English-speaking learners' tonal confusion in identifying the four individual tones in the monosyllabic stimuli is presented in Tables 2 and 3. The error rates for each

Table 2 Confusion matrices of the four individual tones by the learners in the monosyllabic training group from pretest to posttest in percentages

Perceived \ Stimulus	Pretest Monosyllabic Stimuli				Posttest Monosyllabic Stimuli			
	Monosyllabic Training Group				Monosyllabic Training Group			
	T1	T2	T3	T4	T1	T2	T3	T4
T1	88	7	0	5	93	1	0	6
T2	4	82	11	3	4	87	6	3
T3	0	18	82	0	0	10	89	0
T4	2	3	0	95	1	0	0	99

Some rows total 99 or 101% due to rounding

Table 3 Confusion matrices of the four individual tones by the learners in the disyllabic training group from pretest to posttest in percentages

Perceived \ Stimulus	Pretest Monosyllabic Stimuli				Posttest Monosyllabic Stimuli			
	Disyllabic Training Group				Disyllabic Training Group			
	T1	T2	T3	T4	T1	T2	T3	T4
T1	74	11	2	13	87	1	2	10
T2	7	78	10	5	3	88	7	2
T3	0	16	82	2	0	19	81	0
T4	1	4	1	94	2	0	2	96

Some rows total 99 or 101% due to rounding

tone pair were investigated in two directions. For example, in tone pair T1 and T2, when T1 (heard) was misidentified as T2 (perceived), the percentage of errors for T1 → T2 represented the error rate in one direction; when T2 was misidentified as T1, the percentage of errors for T2 → T1 represented the error rate in the other direction. Table 2 shows the tonal confusion on the pretest and posttest by the monosyllabic training group. There were 216 stimuli for each tone (24 monosyllables × 9 learners) in total. Table 3 shows the tonal confusion in the monosyllabic stimuli by the disyllabic training group. There were 192 stimuli for each tone (24 monosyllables × 8 learners). All tonal confusion represented by the error numbers was converted to percentages.

The difference between the two groups' tone perception performance can be seen in the two tables below. The error rates for tone pair T2 and T3 in the monosyllabic training group dropped at least by half, from pretest to posttest, after training in both directions (T2 → T3: 11 vs. 6%; T3 → T2: 18 vs. 10%), but the error rate for tone pair T2 and T3 in the disyllabic training group improved only slightly in one direction (T2 → T3: 10 vs. 7%), while worsening slightly in the other direction (T3 → T2: 16 vs. 19%). This supports the repeated findings by previous studies that T2 and T3 is the tone pair most easily confused in the monosyllabic words.

The second most problematic tone pair is T1 and T4 for which both groups of learners showed asymmetrical confusion. T1 was misidentified as T4 more often than T4 as T1 in both types of stimuli across the two tests. The monosyllabic training group made more errors in the direction of T1 → T4 with 5% on the pretest and 6% on the posttest, while making only 2% on the pretest and 1% on the posttest in the direction of T4 → T1. Similarly, the disyllabic training group showed asymmetrical tonal confusion in the T1 and T4 pair. In one direction, the disyllabic group perceived T1 as T4 13% of the time on the pretest, and this difficulty persisted after training, with an error rate of 10% on the posttest. In the other direction, T4 → T1, the tonal confusion was quite low with only 1% on the pretest, and 2% on the posttest. These results suggest that the learners in both groups were successful in distinguishing T4 from T1 but not as successful in distinguishing T1 from T4.

The most easily distinguished tone pairs in the monosyllabic stimuli for learners in both groups were T1 and T3, and T3 and T4, which had extremely low error rates of zero, 1, or 2% in both directions before and after training. The distinctive pitch height and pitch movement of these two tone pairs in monosyllabic words might have contributed to these low error rates. T1 is a level tone with a high pitch while the low dipping tone in T3's pitch contour is easily perceived and would contrast clearly with the high falling T4.

4.4 Tonal Confusion in the Disyllabic Stimuli by the Training Groups

Confusion between tone pairs in each syllable position were examined in order to understand the mistakes that learners made in the tone identification tasks. The analyses include the tonal confusion of each syllable within one disyllabic stimulus by the two training groups. For example, for tone pair T1 and T2, when T1 (heard) was misidentified as T2 (perceived), the percentage of errors for T1 → T2 represents the error rate in one direction; and, when T2 was misidentified as T1, the percentage of errors for T2 → T1 represents the error rate in the other direction. In Mandarin, there are sixteen pairs of disyllabic tones (4 tones × 4 tones). However, due to the “third tone sandhi” rule, T3 changes to T2 when it precedes another T3 in perception, thus, all T3 + T3 in the tested disyllabic stimuli were coded as T2 + T3.

4.4.1 Tonal Confusion in the Disyllabic Stimuli by the Monosyllabic Training Group

The monosyllabic training group learners’ confusion of the four tones in the two syllables (σ_1 and σ_2) of the disyllabic stimuli are displayed in Tables 4 and 5, respectively. There were 108 stimuli (12 stimuli × 9 students = 108) in each of the first and the second syllable position of the disyllabic stimuli. The error numbers were converted to percentages in both Tables 4 and 5.

Comparing Tables 4 and 5, across the board, learners in the monosyllabic training group made fewer errors on the second syllable position (σ_2) than the first syllable position (σ_1). This is likely because there are fewer tonal variations in σ_2 than in σ_1 . The most difficult tone pair to distinguish in the first syllable position was T3 and T4 for the monosyllabic group learners with T3 → T4 having a 51% error rate on the pretest and 58% on the posttest. This increased error rate shows that learners seem to have more confusion after training when using the monosyllabic stimuli. In other words, using only monosyllabic words in training might not be

Table 4 Confusion matrices of the four individual tones in the first syllable position (σ_1) by the monosyllabic training group from pretest to posttest in percentages

Perceived \ Stimulus	Pretest disyllabic σ_1				Pretest disyllabic σ_1			
	Monosyllabic Training Group				Monosyllabic Training Group			
	T1	T2	T3	T4	T1	T2	T3	T4
T1	59	19	6	16	68	14	2	17
T2	21	56	7	17	24	48	4	24
T3	2	28	19	51	4	23	15	58
T4	12	13	2	73	8	9	1	81

Some rows total 99 or 101% due to rounding

Table 5 Confusion matrices of the four individual tones in the second syllable position (σ_2) by the monosyllabic training group from pretest to posttest in percentages

Perceived Stimulus	Pretest Disyllabic σ_2				Posttest Disyllabic σ_2			
	Monosyllabic Training Group				Monosyllabic Training Group			
	T1	T2	T3	T4	T1	T2	T3	T4
T1	82	6	5	7	84	10	0	6
T2	3	65	29	4	2	56	35	6
T3	0	23	75	2	2	17	80	2
T4	1	11	3	85	1	7	1	91

Some rows total 99 or 101% due to rounding

very helpful for distinguishing T3 from T4 in the first syllable position in disyllabic words. This consistent perceptual difficulty is probably because T3 at σ_1 is subject to T3 sandhi and changes to T2 when preceding another T3 in disyllabic words; and, when preceding any of the other three tones, it changes to a low falling tone T3 (21) (Lin 2007). This T3 alternation at σ_1 is difficult for the learners who have only been exposed to large numbers of the canonical forms of the dipping T3 in the monosyllabic training stimuli. When the monosyllabic training group heard the low T3 (falling), they mapped it onto T4, which they built in their phonetic category as a falling tone in training and probably in their class learning as well. Fortunately, this difficulty occurred only in one direction in the first syllable position. In the other direction, T4 \rightarrow T3, the learners did extremely well with the error rates at only 2% on the pretest, and 1% after training. Similarly, the monosyllabic training group did quite well for tone pair T3 and T4 in both directions in the second syllable position.

The second most confusing tone pair for the monosyllabic group was T2 and T3 in both syllable positions. In the first syllable position, the learners misidentified T3 \rightarrow T2 with error rates of 28% and 23% on the pretest and posttest, respectively. In the other direction, T2 \rightarrow T3, the error rates were 7% and 4% on the pretest and posttest, respectively. This asymmetrical error pattern of tone pair T3 and T2 was probably caused by the T3 sandhi rule in the first syllable position as described above. When T2, a high rising tone, is in the first syllable position, the pitch contour is quite easily distinguished from T3, a lower tone. The high error rate for the tone pair T2 and T3 was found again in the second syllable position in both directions. The error rate of T2 \rightarrow T3 was 29% on the pretest, and it increased to 35% after training. In the other direction, T3 \rightarrow T2, T3 was misidentified as T2 23% of the time on the pretest, and 17% on the posttest. It seems that the learners in the monosyllabic training group performed poorly in both directions on tone pair T2 and T3 in the second syllable position with more errors for T2 \rightarrow T3 than for T3 \rightarrow T2. Such persistent high error rates demonstrate that these tones are difficult for learners to perceive in general.

Three tone pairs, T1 and T2, T2 and T4, and T1 and T4 all showed higher error rates in the first syllable position than the second. This means that the monosyllabic training group was better at identifying these tones at σ_2 than at σ_1 . In σ_1 , the tone

pair T1 and T2 revealed persistent tonal confusion in both directions, T1 → T2 and T2 → T1. For the other two pairs, the error rates were asymmetrical. There were more T2 → T4 and T1 → T4 errors than T4 → T2 errors. That is to say, T4 in the first syllable position was probably the easiest tone to perceive among all the four tones by the disyllabic training group.

The most easily distinguished pair was T1 and T3 in both syllable positions for the monosyllabic group. Only 2 and 4% of T3 were misperceived as T1 on the pretest and posttest in σ_1 , and in σ_2 , 0 and 2% were misidentified. Such low error rates for this tone pair are possibly due to the distinctive pitch height of the two tones in the first syllable position. T1 (55) is a level tone with a high starting pitch while T3 (21) always starts from a low pitch.

4.4.2 Tonal Confusion in the Disyllabic Stimuli by the Disyllabic Training Group

The disyllabic training group's confusion of the four tones in the first syllable position (σ_1) and the second syllable position (σ_2) of the disyllabic stimuli are displayed on Tables 6 and 7, respectively. There were 96 stimuli (12 stimuli × 8 students) in each of the first and second syllables of the disyllabic stimuli. The error numbers have been converted to percentages in both tables below.

In Tables 6 and 7, for the disyllabic training group, T3 was the hardest tone to identify in the first syllable position while both T1 and T2 were most difficult to identify in the second syllable position. The learners misidentified T3 as T4 in the first syllable position most frequently: 60% of the T3 stimuli were perceived as T4 before training. Though the error rate dropped to 53% on the posttest, still, over half of the T3 stimuli were incorrectly perceived as T4. This extremely high error rate for T3 → T4 might be explained by the learners' incorrect perception of the low falling T3 in the first syllable position. This is due to the learners' mistaking the low falling tone T3 (pitch value: 21) for the high falling tone T4 (pitch value: 51) since the pitch directions of the two tones are falling, despite the distinctive onset pitch value. In the

Table 6 Confusion matrices of the four individual tones in the first syllable position (σ_1) by the disyllabic training group from pretest to posttest in percentages

Perceived \ Stimulus	Pretest Disyllabic σ_1				Posttest Disyllabic σ_1			
	Disyllabic Training Group				Disyllabic Training Group			
	T1	T2	T3	T4	T1	T2	T3	T4
T1	48	15	9	28	76	9	0	15
T2	23	38	12	28	18	44	16	23
T3	0	18	22	60	3	22	22	53
T4	9	9	11	70	10	6	5	78

Some rows total 99 or 101% due to rounding

Table 7 Confusion matrices of the four individual tones in the second syllable position (σ_2) by the disyllabic training group from pretest to posttest in percentages

Perceived Stimulus	Pretest Disyllabic σ_2				Posttest Disyllabic σ_2			
	Disyllabic Training Group				Disyllabic Training Group			
	T1	T2	T3	T4	T1	T2	T3	T4
T1	47	20	10	23	63	15	6	17
T2	11	44	34	10	4	68	25	3
T3	4	21	67	8	1	31	67	1
T4	11	11	9	68	10	10	2	77

Some rows total 99 or 101% due to rounding

other direction, T4 \rightarrow T3, the error rates in σ_1 were relatively low with only 11% on the pretest, and 5% on the posttest.

A similarly low error rate by the disyllabic training group was observed for tone pair T3 and T4 in the second syllable position. Eight percent of the T3 instances were misperceived as T4 on the pretest, and the error rate decreased to 1% on the posttest; 9% of the T4 instances were misidentified as T3, which decreased to 2% on the posttest. These unbalanced error rates in the two syllable positions were also observed in the performance of the monosyllabic training group.

However, the difference between the two groups' performance lies in the identification in the first syllable position of T3 \rightarrow T4. The monosyllabic training group learners seemed to misperceive more T3's as T4's after training with a posttest error rate of 58%, compared to the pretest error rate of 51%. On the other hand, the disyllabic training group made some improvements in identification after training, by decreasing their error rate from the pretest rate of 60% to the posttest rate of 53% despite their difficulties in distinguishing T3 from T4. This implies that disyllabic training is more effective in helping learners to identify the low falling T3 in the first syllable position of the disyllabic stimuli than the monosyllabic training is.

The second most problematic tone pair for the disyllabic group was T2 and T3. High error rates appear in both directions, T2 \rightarrow T3 and T3 \rightarrow T2, in both syllable positions. However, it is noteworthy that the T2 mean accuracy of the disyllabic training group improved from 38 to 44% in the first syllable position, and from 44 to 68% in the second syllable position, whereas the mean accuracy for T3 stayed the same at 22% even after the disyllabic training. This improvement in T2 identification did not occur in the monosyllabic training group, whose accuracy rates for both syllable positions actually dropped after training from 56 to 48% and 65% to 51%, respectively. Such observed improvement in one group but not the other may be due to the training effect.

Tone pairs, such as T1 and T2, T1 and T4, and T2 and T4, all showed error rates that are higher in one direction than the other. What is more interesting is that, despite the imbalanced patterns of error rates, all the error rates for these tones dropped to some degree on the posttest for the disyllabic training group learners. Such a drop in

cross-board error rates for these tones was not observed in the monosyllabic training group.

The tone pair T1 and T3 in both syllable positions was the easiest tone pair to distinguish by the disyllabic training group, which was similar to the results for the monosyllabic training group.

Research Question 3. During high-variability phonetic perceptual training, is the tone perception of disyllabic Chinese words by English-speaking learners affected by linguistic factors? Specifically, is the tone perception of learners affected by syllable position, tonal context, or tonal sequence?

Tone identification accuracy data was analyzed to examine learners tone perception through the lens of the three linguistic factors: syllable position (initial vs. final), tonal context (compatible vs. conflicting), and tonal sequence (same vs. different).

4.5 Training Effects on Syllable Position

Figure 7 displays the mean accuracy of the four tones in the two syllable positions by English-speaking learners in two training groups from the pretest to posttest. A three-way repeated measures ANOVA was conducted with Syllable Position (initial and final) and Test (pretest, posttest) as within-subjects factors, and Training Group (Monosyllabic and Disyllabic) as the between-subjects factor, and Accuracy as a

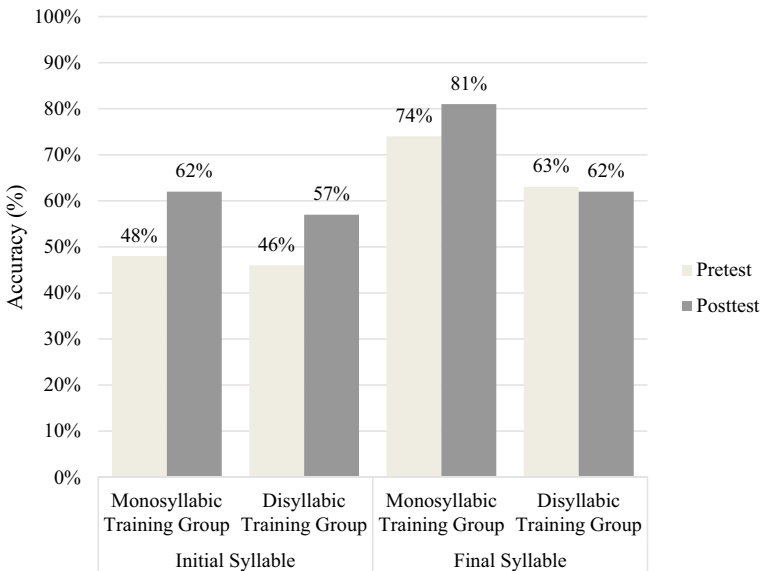


Fig. 7 Average accuracy in the initial and final syllable positions for English-speaking learners in the two training groups on the pretest and posttest

dependent variable. The results showed a main effect of Test [$F(1,15) = 18.797, p = 0.001$], suggesting that the learners identified Chinese tones significantly better on the posttest (66%) than the pretest (58%). We also found a main effect of Syllable Position [$F(1,15) = 85.530, p < 0.001$], suggesting learner identification was significantly better on tones in the final position (70%) than the tones in the initial position (53%).

The results also showed significant two-way interactions between Test and Syllable Position [$F(1,15) = 10.833, p = 0.005$], showing an improvement of 13% accuracy in the initial position from the pretest (47%) to the posttest (60%), which was significantly higher than the 2% improvement in the final position from the pretest (69%) to the posttest (71%). This difference in improvement suggests that after training, the learners' tone perception of the disyllabic stimuli improved more in the initial position than at the final position.

Additionally, a significant two-way interaction between Syllable Position and Training Group [$F(1,15) = 9.823, p = 0.007$] was found. This result suggests that, after perceptual training, the difference in the accuracy rate for the initial position between the monosyllabic training group (55%) and the disyllabic training group (52%) was significantly lower than the difference in the accuracy rate for the final position between the two groups with accuracy rate of 77% and 63% respectively. This interaction suggests that the learners in the monosyllabic training group did better at identifying tones in the final position than the disyllabic training group did. This is not surprising, considering that the monosyllabic training group's training stimuli only contained the citation form of the tones, which are more similar to tones in the final syllable position than those in the initial syllable position. Thus, it was more advantageous for the monosyllabic group to identify tones in final syllable positions when contrasting to the disyllabic training group, which were trained in coarticulated tones that resembles much less of the citation form of tones.

Overall, these results suggest that, in disyllabic stimuli, learners across groups were significantly more accurate when identifying tones in the final syllable position than in the initial position. Both types of training showed significant improvement for tones in the initial position from pretest to posttest.

4.6 Training Effects on Tonal Context

Figure 8 shows how English-speaking learners in the two training groups performed on the tone identification task in two tonal contexts, compatible and conflicting, from the pretest to the posttest. A three-way repeated measures ANOVA, with the Test (pretest, posttest) and Tonal Context (compatible, conflicting) as the within-subjects factors, and the Training Group as the between-subjects factor, and Accuracy as the independent variable was conducted. The results showed a main effect of Test [$F(1,15) = 5.552, p = 0.032$], suggesting that the learners did significantly better on the posttest after training (44%) than on the pretest (38%). A significant main effect of Tonal Context [$F(1,15) = 14.183, p = 0.002$] was also found, indicating that the learners did significantly better in the compatible tonal context (45%) than in the

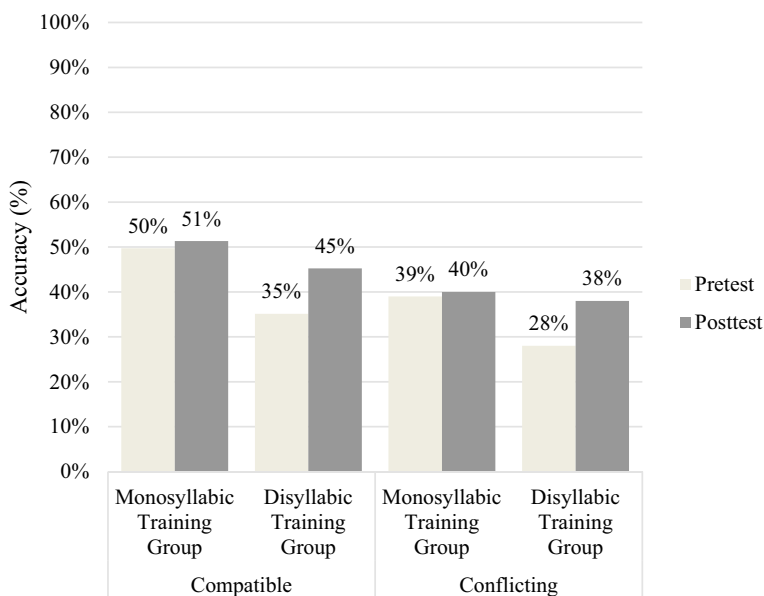


Fig. 8 Average accuracy in compatible and conflicting tonal contexts of English-speaking learners in the two training groups on the pretest and posttest

conflicting tonal context (36%) with a 9% higher accuracy rate. That is to say, the learners identified tones in the compatible tonal contexts more accurately than in the conflicting ones.

4.7 Training Effects on Tonal Sequence

Figure 9 shows how English-speaking learners in the two training groups performed on tone identification in two tonal sequences, namely the same tonal sequence and different tonal sequence, in disyllabic stimuli from the pretest to the posttest. A three-way repeated measures ANOVA, with Test (pretest and posttest) and Tonal Sequence (same and different) as the within-subjects factors, and the Training Group (monosyllabic training group and disyllabic training group) as the between-subjects factor was conducted with Accuracy as the dependent variable. The results showed a main effect of Tonal Sequence [$F(1,15) = 19.630, p < 0.001$], and a significant two-way interaction between Tonal Sequence and Training Group [$F(1,15) = 6.252, p = 0.024$].

The main effect of Tonal Sequence indicates that learners across the training groups and tests did significantly better in the same tonal sequence with an accuracy rate of 55% than in the different tonal sequence with an accuracy rate of 37%. At the same time, learners in the monosyllabic training group did substantially worse in the

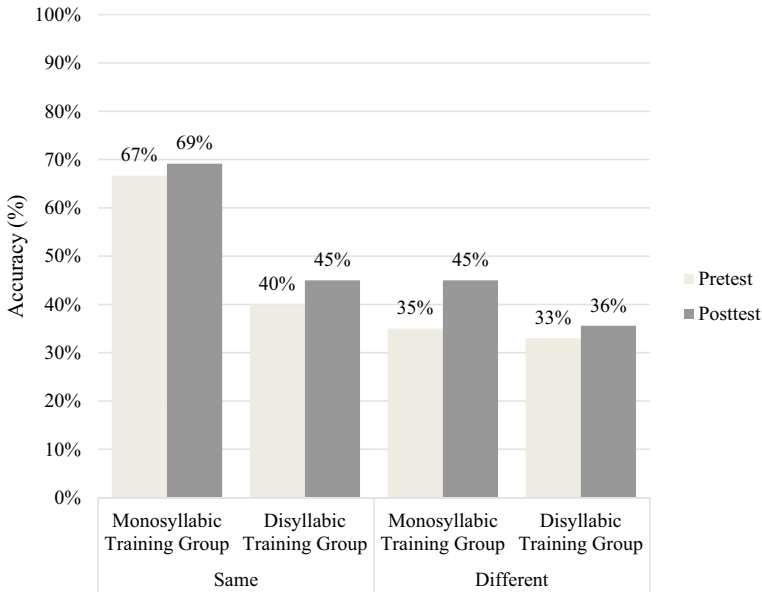


Fig. 9 Average accuracy for the same and different sequences for English-speaking learners in two training groups on the pretest and posttest

different tonal sequence (39%) than in the same tonal sequence (68%). This sizable difference of 29% between the same and different sequences shows that the learners who were trained using monosyllabic tone words were quite good at identifying tones in same tonal sequence (i.e. T1 + T1) but really bad at identifying tones in the different tonal sequence (i.e. T1 + T2). The learners in the disyllabic training group had a mean of 34% in the different tonal sequence and 44% in the same tonal sequence, which is nearly 20% less difference compared to the monosyllabic training group.

5 Discussion and Conclusion

5.1 Improved Tones

Although, overall, both training groups did significantly better on the posttest after training than on the pretest before training, when looking at performance on the four individual tones, the perceptual training effect for the two groups tended to be different. The monosyllabic training group did better on their overall tone identification task with an accuracy rate of 87% on the pretest that increased to 92% on the posttest for the monosyllabic test stimuli. However, there was no difference in

improvement on individual tones after training. The disyllabic training group also did statistically better on the overall tone identification task with an accuracy of 81% on the pretest and 88% on the posttest. What is more important is that the disyllabic training group made substantial improvement in the individual tones. T1 and T2 both improved statistically from the pretest to the posttest after training from 74 to 87% and 77% to 88%, respectively. This demonstrates that both types of perceptual training were helpful in tone learning but the disyllabic perceptual training was more effective than the monosyllabic training.

5.1.1 Tones in Monosyllabic Test Stimuli

For the four individual phonemic tones, the learners in both groups identified T4 (96%) with significantly more accuracy than T1 (86%), T2 (84%) and T3 (84%) after the training. This supports the findings of previous studies that adult learners do not perceive the four tones in isolation with equal accuracy. Sun (1998) found that American learners identified both T1 and T4 better than T2 and T3 in an isolated environment. Similarly, He (2010) also found that T2 was the most difficult tone to identify in monosyllabic stimuli, by both low-proficiency and high-proficiency American learners. This is because T1 and T4 share high onset pitch values that are perceptually salient and more easily identified by the learners than T2 and T3, which share low onset pitch values. Also, Lai and Zhang (2008) suggest that using the isolation point (IP) to examine the time difference in identifying the four tones, the IP is the fastest for T1 (a high register tone), followed by T4 (a high contour tone), and then by T2 and T3. In other words, the learners may also use faster perceptual processing when identifying the four tones on the test, with T1 and T4 easier to identify than T2 and T3.

5.1.2 Tones in Disyllabic Test Stimuli

For the disyllabic test stimuli, results show that the monosyllabic training group did not make significant improvement in overall accuracy from the pretest (43%) to the posttest (45%). However, the disyllabic training group made significant improvement ($p = 0.048$) from the pretest with 29% accuracy to the posttest with 39% accuracy on the disyllabic test stimuli. These results suggest that when trained with disyllabic stimuli (as was the disyllabic training group), beginning-level English-speaking learners learn the tones significantly more effectively than those trained with monosyllabic stimuli (as was the monosyllabic training group). For identifying tones in the disyllabic test stimuli, the disyllabic training was much more effective overall. This indicates that teaching learners the canonical form of Mandarin tones in monosyllabic stimuli is insufficient in helping to build a robust tonal category.

When identifying tones in the two syllable positions, the performance of the two training groups differed. The results show, from the pretest to the posttest, across the two groups, T3 was the most difficult tone to identify when in the first syllable

position (σ_1), with a low accuracy rate of 24%, followed by T2 (47%), T1 (62%) and T4 (76%); in the second syllable position (σ_2), T2 had the lowest accuracy rate (58%) among the four tones, then T1 (69%), T3 (72%) and T4 (80%).

5.2 *Problematic Tone Pairs*

The problematic tone pairs in all sixteen possible combinations are identified below in the disyllabic stimuli. Knowing which are the challenging tone pairs would help when teaching and learning tones.

5.2.1 **Tonal Confusion in Monosyllabic Words**

From the tone error results of the two training groups, it is clear that in the monosyllabic stimuli, the most confusing tone pair is T2 and T3 in both groups from the pretest to the posttest. This finding supports the claim in previous studies that T2 and T3 are the most difficult tones for learners to perceive in monosyllabic words (Sun 1998; He 2010; He and Wayland 2013). One interesting finding is that learners in the monosyllabic training group showed a decrease in error rates of the tone pair T2 and T3 in both directions, while learners in the disyllabic training group had a slightly higher error rate after the training than on pretest in identifying T3 as T2 (16% pretest error rate versus 19% posttest). This comparison suggests that the monosyllabic training seemed to help the learners distinguish between T2 and T3 when the two tones were in the stable citation form shown in monosyllabic words. However, since the disyllabic training used the coarticulated tones as shown in disyllabic stimuli, the tonal variations probably did not provide as many instances of stable input of T3 to the learners in the disyllabic training group as those in the monosyllabic training group; therefore, the disyllabic training group seemed to have more difficulty in distinguishing between T2 and T3 in monosyllabic stimuli even after training. However, it is necessary to remember that isolated T2 and T3 almost never appear in real life conversation.

The most easily distinguishable tone pairs in monosyllabic words were T1 and T3 and T3 and T4. This is probably due to the salient pitch difference in these two tone pairs, making them easy to perceive for all learners. In the monosyllabic stimuli, T3 has a falling and rising contour while T1 has a level tone with no contour. T4 is a high falling tone.

The performance of the two training groups on the tone pair T1 and T4 was different. The monosyllabic training group's performance of less than 5% error rate in both directions from the pretest to the posttest indicates that the learners in this group were excellent at distinguishing T1 from T4, and T4 from T1 in the monosyllabic stimuli before and after training. A similarly high rate of accuracy appeared only when the learners in the disyllabic training group distinguished T4 from T1, but not

T1 from T4. This asymmetrical misidentification was probably caused by the shared high pitch value at the onset of the two tones in monosyllabic words.

5.2.2 Tonal Confusion in Disyllabic Words

The results showed that across both training groups, among all possible tone pairs in disyllabic words, the learners made the most errors in perceiving T3 as T4 in the first syllable position. The error rates of the monosyllabic group for T3 → T4 were 51 and 58%, while those of the disyllabic group were 60% and 53%, respectively, before and after training. Such a high error rate of misidentification of T3 → T4 in the first syllable position was probably caused by the “half-third sandhi” tone rule. In this rule, T3’s pitch value of 213 is reduced to 21, and it becomes a low falling tone before any tone other than another T3 (Zhang 2007; Zhang and Lai 2010). This tonal variation was not introduced to the learners in the monosyllabic training group. They were only exposed to the T3 citation form with a falling and rising pitch contour at pitch value 213. Therefore, when the monosyllabic group heard a low falling T3 (pitch value of 21) in σ_1 , their monosyllabic stimuli-trained tonal category mapped it as a falling tone, which resembled the contour movement of another falling tone, T4 (pitch value 51). The low-falling T3 misidentified as T4 in perception was, in fact, observed by Gottfried and Suiter (1997). They reasoned that this type of error was related to the phonological change in the stimuli, since in the first syllable, T3 has a low-falling tone instead of the dipping-rising pattern as it has in isolation. When the American listeners paid more attention to the movement/direction, they confused these two tones. This probably explains this highest level of difficulty in distinguishing T3 from T4 only in the first syllable position which was shared by both groups. This finding partially agrees with what Sun (1998), He (2010), and He and Wayland (2013) found in their studies: that T2 and T3 were the most difficult tones to identify in the first syllable position.

Fortunately, that performance did not appear in the other direction (T4 → T3), nor in the second syllable position. Both groups did quite well at identifying T4 → T3 in the first syllable position. This is not only because of the offset pitch differences in these two tones (T4 is high while T3 is low), but is also likely because T4 has a distinctively high falling pitch that the beginning learners might subconsciously map onto their native (English) language sound system as a sentence-end falling intonation that they are very familiar with. In other words, they mastered the falling sound T4. This was also demonstrated by fact that T4 identification had the highest accuracy rate in this study.

Learners in both groups performed well when distinguishing T3 and T4 in the second syllable position before and after training. This is probably due to the canonical-like forms of these two tones in the second syllable position in disyllabic words which most closely resembles the standard isolated tones that the learners easily mapped onto their stored tonal category.

For the monosyllabic group learners, after training, the identification of some tones improved, such as T1 and T4 in the first syllable position and T1, T3 and

T4 in the second syllable position, while other tones worsened, such as T2 and T3 in the first syllable position and T2 in the second syllable position. Unlike the monosyllabic group's mixed performance after training in all four tones at both syllable positions, the disyllabic group's identification accuracy rates in all four tones increased from the pretest to the posttest except for T3, which showed no change. These results not only support the previous finding of T3 generally being the most difficult among the four tones, but they also show that disyllabic training demonstrates more improvement in tone identification than monosyllabic training for each individual tone in the disyllabic stimuli across the syllable positions.

T2 and T3 were the most difficult tones to distinguish in the second syllable position across the two training groups before and after the perceptual training. This finding confirms the claims of previous studies. The easiest tone pair to differentiate by all learners in the two syllable positions is T1 and T3. This is probably due to the clear difference embedded in the phonetic characteristics. For instance, T1 has a high onset while T3 has a low onset; T1 is a level tone without change in the pitch contour, but T3 (21) is a low falling tone in the first syllable position and a falling-rising tone in the second syllable position (213).

Overall, across the board from the pretest to the posttest, the learners in both training groups made many more tone errors in both syllable positions in the disyllabic stimuli than in isolated tones in the monosyllabic stimuli. This is due to the tonal coarticulation and variations in disyllabic stimuli which differ greatly from the stable, canonical tones in the monosyllabic stimuli. Moreover, unlike the learners in the disyllabic training group, the learners in the monosyllabic training group were not exposed to variable tonal exemplars in the disyllabic stimuli context, so it seems that the monosyllabic group learners made more errors after the training when identifying tones in the disyllabic stimuli compared to their performance before training. On the other hand, although the learners in the disyllabic training group also made more errors before training, they did show improvement in many tone pairs, such as T1 and T2, T3 and T4, T2 and T3, T2 and T4. These improvements likely resulted from the perceptual training.

5.3 Linguistic Factors

Learners' tone identification performance for syllable position, tonal context, and tonal sequence are examined below.

5.3.1 Syllable Position

Overall, the English-speaking learners identified tones in the second syllable position more accurately than in the first syllable position. The significance of tonal accuracy in the final syllable echoes findings by Sun (1998), and He and Wayland (2013) for tone identification in disyllabic words. This pattern may be attributed to the

following: first, in disyllabic stimuli, the tone in the final syllable tends to have a longer duration than those in the first syllable in natural production (Xu and Wang 2009). Thus, the shape of the tone is more fully represented in the final position than in the initial position. Second, there may be a recency effect such that the tones in the final syllable are heard more recently by learners as compared to the tones in the initial syllable, so the learners are able to identify the tones in the final syllable with greater accuracy.

Overall, the learners made more improvement in the initial tones than the final tones. For example, the monosyllabic training group increased their accuracy rates on tones in the initial position from 48% on the pretest to 62% after training, while the increase in the final position went from 74% on the pretest to 81% on the posttest. Similar tonal improvement appeared for the disyllabic training group, but only for tones in the initial syllable. The disyllabic learners increased their accuracy rates from the pretest 46% to the posttest 57% in the initial position, but there was no significant improvement for the final position. These results demonstrate that perceptual training is effective, especially for the tones in the initial syllable position.

5.3.2 Tonal Context

This study looks at the results from the pretest to the posttest using disyllabic stimuli in two tonal contexts: compatible and conflicting. The learners performed significantly better in compatible tonal contexts (45%) than in conflicting tonal contexts (36%) with a 9% increase ($p = 0.002$). This may be because the degree of adjustment between the two adjacent tones is relatively small in compatible contexts as compared to conflicting contexts (Xu 1994). As Xu found, a conflicting tonal context can substantially change the original tonal contours to the extent that they resemble other tone categories. Thus, it is more difficult for learners to identify tones that are distorted by conflicting contexts tones in compatible contexts. The coarticulated tones that contain tonal variations are difficult for learners to acquire within a short training period. This finding confirms the results of He and Wayland (2013), in which American learners identified tones in compatible tones better than in conflicting tones, across the proficiency levels.

In general, the learners performed better after training. The learners in the disyllabic training group in particular improved more from the pretest to the posttest than those in the monosyllabic training group, in both tonal contexts. From the pretest to the posttest, the disyllabic learners in the compatible tonal contexts increased their accuracy rate by 10% (35–45%), while the monosyllabic learners made very little improvement (50–51%). Similarly, in the conflicting tonal contexts, the disyllabic learners increased their tone identification accuracy rate by 10% (28–38%), while the monosyllabic learners made almost no improvement (39–40%). Overall, it seems that the disyllabic training helped the learners to a greater degree than did the monosyllabic training when identifying both the compatible and conflicting tones in disyllabic stimuli.

5.3.3 Tonal Sequence

The accuracy rates for the same and different tonal sequences in disyllabic stimuli were analyzed. It was found that the learners across the training groups did significantly better ($p < 0.001$) in the same tonal sequences (55%) than they did for the different tonal sequences (37%). However, this finding is different from that of He (2010) who found no difference between the same and different tonal sequences by her American learners of Mandarin.

In the current study, the advantages demonstrated in the perception of tones in the same tonal sequence may for two reasons. The first is that the high variability phonetic training, especially in disyllabic training with only disyllabic stimuli, provided many exemplars of each tone to the learners so that they could develop more robust tonal categories for all four phonemic tones after training, despite the contextual difference in these tone combinations, such as T1 + T1, T2 + T2, and T4 + T4. The learners in this study made great gains in tones in these same tonal sequences. The second may simply be the tonal repetition. The beginning learners, who only had limited exposure to the target language before participating in the study, appeared to perceive the same tonal sequences better after training. This, again, speaks to the importance of building a robust tonal category for beginning learners.

These findings demonstrate that the learners were generally good at perceiving tones in the same tonal sequences but not at identifying those in sequences of different tones—sequences which embody many tonal coarticulation and variations. At the same time, the learners identified tones in compatible tonal contexts significantly better than in conflicting tonal contexts. Moreover, the learners perceived tones in the final syllables significantly better than in the initial syllables. All the results suggest that to improve English-speaking learners' tonal perception of coarticulated tones, providing the learners with more perceptual training time on (1) tones in different tonal sequences than in the same tonal sequences, (2) more tones in different tonal contexts than in the same tonal contexts, and (3) more tones in the initial syllable position than in the final syllable position is highly efficacious.

6 Pedagogical Implications

The current study investigated native English-speaking learners' tonal behavior in monosyllabic and disyllabic words before and after perceptual training. The results show a positive training effect due to high variability phonetic training on tonal perception for those learners. The improved tones, difficult tones, and tone pairs were analyzed, and the linguistic factors in tones were studied in hopes of helping with the teaching and learning of tones.

The results demonstrated that all learners improved their accuracy of tone identification after the perceptual training. Findings supported the hypothesis that the disyllabic training allowed for more improvement in tone identification than the

monosyllabic training did, especially in disyllabic words. The mainstream classroom tone teaching was captured by Xing (2006) and Orton (2013). In their long-term observation of the teaching and learning of Mandarin tones in the United States from public schools to universities, they both found that tone teaching was given little attention in Mandarin Chinese language classrooms nationwide. Tones were introduced primarily as isolated tones in monosyllabic words, not as coarticulated tones in disyllabic words as presented in the current perceptual training. This presents a problem because disyllabic words compose at least 70% of all words used in the modern Mandarin vocabulary (Zhou et al. 1999, p. 526; Duanmu 1999). Disyllabic words and their connected tones are used more often in the daily lives of Chinese speakers than monosyllabic words with their isolated tones. The tones in disyllabic words mirror the tones perceived and produced at the sentence level in real conversation more than isolated tones do. Therefore, the teaching of tones in disyllabic words is urgently needed in Chinese classrooms not only across the US, but also in other English-speaking countries. For Chinese language teachers, it seems useful and necessary to incorporate a short perceptual training of disyllabic tones into their teaching labs to help Mandarin learners acquire tones that sound more native-like. It is suggested that when the Chinese language teachers introduce tones, they can introduce the four tones in isolation briefly, but then they should emphasize introducing and practicing tones using disyllabic words, which carry many more tonal variations and coarticulation as in real conversations. Moreover, for challenging tone pairs identified above, minimal pair practice with these tones embed in disyllabic words would help the learners in tone learning. When more tonal exemplars in disyllabic words are provided, learners are exposed to more variations of the four phonemic tones in different syllable positions. This will benefit the shaping of the learners' tone category.

7 Limitations and Future Research

The present study describes the tonal perceptions of native English-speaking learners in both monosyllabic and disyllabic stimuli before and after perceptual training. The findings strongly suggest that when teaching tones in Chinese classes, the focus should be shifted from teaching isolated tones by using monosyllabic stimuli to teaching coarticulated tones by using disyllabic stimuli, which better simulates natural, realistic learning environments for improving learners' tone identification. In the current study, all participants were beginning-level native English learners of Mandarin at a Midwestern university in the US. They had fewer than two semesters of studying the target language and were at a novice level of proficiency. While the results of this study cannot be generalized to learners whose native language is not English, it is expected that similar patterns would be observed. Nor can the current results be generalized to learners whose Chinese language proficiency is above or below the novice level. Future studies might investigate learners at different language proficiency levels and groups of learners other than native English speakers, using

the same perceptual training to facilitate the effects on improving tonal perception. It is hypothesized that similar improvements will be found.

References

- Chang, Y. H. S. (2011). Distinction between Mandarin Tones 2 and 3 for L1 and L2 Listeners. In Z. Jing-Schmidt (Ed.), *Proceedings of the 23rd North American conference on Chinese linguistics (NACCL-23)* (Vol. 1, pp. 84–96). Eugene: University of Oregon.
- Chao, Y. R., & Pian, R. C. (1955). *Mandarin primer* (p. 25). Folkways Records.
- Duanmu, S. (1999). Stress and the development of disyllabic words in Chinese. *Diachronica*, 16(1), 1–35.
- Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics*, 36(2), 268–294.
- Gottfried, T. L., & Suiter, T. L. (1997). Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of Phonetics*, 25(2), 207–231.
- Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 24(2), 95–112.
- Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, 40(2), 269–279.
- He, Y. (2010). *Perception and production of isolated and coarticulated Mandarin tones by American learners* (Doctoral dissertation, University of Florida).
- He, Y., & Wayland, R. (2013). Identification of Mandarin coarticulated tones by inexperienced and experienced English learners of Mandarin. *Chinese as a Second Language Research*, 2(1), 1–21.
- Herd, W., Jongman, A., & Sereno, J. (2013). Perceptual and production training of intervocalic/d, r, r/in American English learners of Spanish. *The Journal of the Acoustical Society of America*, 133(6), 4247–4255.
- Jongman, A., Wang, Y., Moore, C. B., & Sereno, J. A. (2006). In: P. Li, L. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *Perception and production of Mandarin Chinese tones. Handbook of Chinese Psycholinguistics*. Cambridge University Press.
- Kingston, J. (2003). Learning foreign vowels. *Language and Speech*, 46(2–3), 295–348.
- Lai, Y., & Zhang, J. (2008). Mandarin lexical tone recognition: The gating paradigm. *Kansas Working Papers in Linguistics*, 183–198.
- Lee, C. Y., Tao, L., & Bond, Z. S. (2010a). Identification of multi-speaker Mandarin tones in noise by native and non-native listeners. *Speech Communication*, 52(11), 900–910.
- Lee, C. Y., Tao, L., & Bond, Z. S. (2010b). Identification of acoustically modified Mandarin tones by non-native listeners. *Language and Speech*, 53(2), 217–243.
- Li, Y., Lee, G., & Sereno, J. (2019). Comparing monosyllabic and disyllabic training in perceptual learning of Mandarin. In Nyvad, Anne Mette (Eds.), *A sound approach to language matters in honor of Ocke-Schwen Bohn* (pp. 303–319). Aarhus University Library/Royal Danish Library—AU Library Scholarly Publishing Service.
- Lin, Y. H. (2007). *The sounds of Chinese*. Cambridge University Press.
- Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English/r/and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, 94 (3), 1242–1255.
- Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English/r/and/l: A first report. *The Journal of the Acoustical Society of America*, 89(2), 874–886.
- Miracle, W. C. (1989). Tone production of American students of Chinese: A preliminary acoustic study. *Journal of the Chinese Language Teachers Association*, 24(3), 49–65.
- Orton, J. (2013). Developing Chinese oral skills—a research base for practice. *Research in Chinese as a Second Language*, 3–26.

- Shen, X. S. (1989). Toward a register approach in teaching Mandarin tones. *Journal of the Chinese Language Teachers Association*, 24(3), 27–47.
- Shen, X. N. S. (1990). The prosody of Mandarin Chinese. *Linguistics* (Vol. 118). Berkeley, California: University of California Press.
- Shen, G., & Froud, K. (2016). Categorical perception of lexical tones by English learners of Mandarin Chinese. *The Journal of the Acoustical Society of America*, 140(6), 4396–4403.
- Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech*, 34(2), 145–156.
- Sun, S. H. (1998). *The development of a lexical tone phonology in American adult learners of standard Mandarin Chinese* (No. 16). University of Hawaii Press.
- Tagliaferri, B. (2008). Paradigm: Perception research systems [Computer Program]. Retrieved from <https://www.paradigmexperiments.com>.
- Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, 106(6), 3649–3658.
- Xing, J. Z. (2006). *Teaching and learning Chinese as a foreign language: A pedagogical grammar* (Vol. 1). Hong Kong University Press.
- Xu, Y. (1994). Production and perception of coarticulated tones. *The Journal of the Acoustical Society of America*, 95, 2240.
- Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25(1), 61–83.
- Xu, Y., & Wang, M. (2009). Organizing syllables into groups—Evidence from F0 and duration patterns in Mandarin. *Journal of Phonetics*, 37(4), 502–520.
- Zhang, J. (2007). A directional asymmetry in Chinese tone sandhi systems. *Journal of East Asian Linguistics*, 16(4), 259–302.
- Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, 27(01), 153–201.
- Zhou, X., Marslen-Wilson, W., Taft, M., & Shu, H. (1999). Morphology, orthography, and phonology reading Chinese compound words. *Language and Cognitive Processes*, 14(5–6), 525–565.

Tone Category Learning Should Serve Tone Word Learning: An Experiment of Integrating Pronunciation Teaching in the L2 Chinese Curriculum



Jiang Liu and Cheng Xiao

Abstract Tones are the primary focus in teaching L2 Chinese pronunciation. Seemingly, less effort is made to integrate the tone category and tone word learning. The current study tested whether the use of (near) minimal pairs formed by monosyllabic and disyllabic words that the learners have previously learned can direct L2 Chinese learners' attention to tonal contrast, thus, improve their tone production and memorization of the tone words. 66 beginner-level learners were assigned to a dictation only (traditional) group and a perception plus production training (experimental) group in which minimal and near-minimal pairs that included tonal contrasts were used as training stimuli. Both groups recorded the target words in a pretest, an immediately administered posttest, and a delayed posttest. Using native speakers' comprehensibility judgment as assessment, we found that the experimental group had better comprehensibility than the traditional group in the immediate and delayed posttests. The comprehensibility ratings seemed to vary across words. Participants also had two dictations in the pretest and delayed posttest respectively. The words used in the recording tasks had a significantly higher dictation score than those not used in the recording tasks in the posttest. The pedagogical significance of these findings is discussed.

Keywords Pronunciation teaching · Perception · Tone category · Tone word learning

1 Introduction

One goal in pronunciation teaching research in the domain of Instructed Second Language Acquisition (Instructed SLA) is to test how effective various pedagogical claims are through empirical studies. This chapter first reviews research in both applied and psycholinguistic fields to find a connection between research and pedagogical practice of teaching pronunciation. By doing so, we argue that teaching

J. Liu (✉) · C. Xiao
University of South Carolina, Columbia, SC, USA
e-mail: jiangliu@mailbox.sc.edu

lexical tones to Chinese as Second Language (CSL) learners should serve the purpose of helping their word learning. Based on previous research findings, we then designed a pronunciation teaching method that aimed to direct learners' attention to the tonal contrast in monosyllabic and disyllabic words and integrated it into the Chinese curriculum. To test its effectiveness, we compared the integrated approach to the traditional classroom instruction where there is no systematic training on the perception and production of tonal contrasts in class teaching. The result obtained from the empirical study provides us with some direction for pronunciation teaching in the instructed L2 Chinese classroom.

In the literature, there is a consensus that perception plays a critical role in production in L2 speech learning. For example, the Perceptual Assimilation Model (PAM) posits that speakers employ articulatory gestures as the basis of speech perception (Best 1995; Best and Tyler 2007). The Speech Learning Model (SLM) claims that well-formed phonological representation at the perception level is essential for target-like sensory motor skills and accurate L2 speech production (Flege 1995; Flege, Schirru and MacKay 2003). In recognizing the importance of perception in L2 speech learning, various perceptual training paradigms aim at improving learners' perception of L2 speech sounds. The methods include High Variability Phonetic Training or HVPT (Lively, Logan and Pisoni 1993), auditory-visual stimuli (Hardison 2005), and hyper-articulated/exaggerated speech stimuli method (Iverson et al. 2005). Studies have shown that the production accuracy of L2 learners can benefit from perception training (Bradlow et al. 1997; Hardison 2005; Thomson 2011; Wang et al. 2003). Those findings imply that L2 learners can transfer their knowledge in perception trained on a specific set of words to the perception and likely the production of new words (with similar phonological contexts). So far, less research has discussed these perceptual training studies in the context of L2 learners' lexical development. From a pedagogical perspective, the ability to perceive (categorize and differentiate) phonemes in L2 needs to be employed for L2 word recognition, including phonological form, semantic form and orthography. Sometimes, even if learners improve their perception of phonemes or lexical tones provided with sound category contrast, they may rarely see or hear such contrast in word learning because the words with phonemic contrast are rarely learned at the same time (e.g., rake versus lake in English; 学xue2 'to learn' versus 血xue3 'blood' in Chinese). The 'phonetic-phonological-lexical continuity' is a complex learning process where learners need to apply their phonetic knowledge (e.g., differentiate phonemes) to learn lexical items by memorizing those items (Wong and Perrachione 2007). In the current study, we aim to integrate the perceptual training for tonal contrast into the word learning so as to build a cycle of tone learning and word learning. In this way, on the one hand, highlighting phonemic contrasts is likely to help learners differentiate, identify and memorize the spoken words better. On the other hand, given two words that have different meanings, learners are likely to notice that it is the phoneme/toneme that makes different meanings. In the current study, we tested whether this potential mutually enhancing effect exists for pronunciation teaching and word learning.

Another related topic in L2 pronunciation teaching is to decide what pronunciation features to teach. Darcy et al. (2012) once stated, 'there is no agreed upon system of

deciding what [pronunciation features] to teach, and when and how to do it.’ Against this ‘give-up’ viewpoint, we argue that directions and guidelines in pronunciation teaching can be found based on previous research findings. To answer the *what* (to teach) question, based on PAM and SLM, some L2 sound categories are easier to learn than others. For example, in terms of segment acquisition, numerous phonetic studies investigated the effectiveness of various perceptual training paradigms for improving Japanese speaking L2 English learners’ perception and production of ‘r’ /ɹ/ and ‘l’ /l/ in L2 English. The reason to target teaching the pronunciation of /ɹ/ and /l/ in L2 English is that in Japanese, there is only one liquid sound ‘r’ that sounds more similar to English /l/ (Lively et al. 1993; Bradlow et al. 1997; Iverson et al. 2005). When it comes to L2 Chinese pronunciation teaching, tonal contrast in various contexts should be targeted (Yang 2019; Zhang 2018). Some more specific questions could be: Are tones always difficult, or are they only difficult in long, multi-syllable words or sentences? What is difficult about tones, hearing them, remembering them, or noticing them in meaningful speech? In the current study, we focus on learners’ difficulty with tonal contrast in disyllabic words in L2 Chinese. To address the *when* (to teach) question, it has been argued that inaccurate perceptual performance is more likely a matter for L2 beginners, which causes the production problems (Trofimovich et al. 2009). Colantoni and Steele (2008) also agree that it is less contentious in the field of study that the strength of the link between L2 speech production and perception varies at different proficiency levels where the link at the early stage of L2 learning tends to be stronger than a late stage. Based on the empirical evidence and pedagogical guideline, in the current study, we focused on teaching pronunciation to beginner-level L2 Chinese learners. In terms of the *how* (to teach) question, the selective perception model (SPM) claims that online speech perception by adults is processed via highly overlearned selective perceptual routines (Strange 2006). Due to L1 experience, adults pay more attention to phonetic features that are present in L1. When learning L2, learners need to learn how to redirect their perceptual attention to phonetic features in L2 that are not frequently (if at all) used in their L1 (Chandrasekaran et al. 2010; Iverson et al. 2005; Lim and Holt 2011). Although the ability to modify L2 speech perception patterns is maintained well into adulthood (Flege et al. 2003), the ‘retuning’ procedure to develop targetlike perception patterns is not something learned incidentally. Instead, it appears to require a great deal of language exposure and explicit training (Bradlow 2008). Therefore, in L2 pronunciation teaching, we should use training methods to retune L2 learners’ perception as the first step.

1.1 Tone Category Versus Tone Word Learning in L2 Chinese

Having reviewed the perception-pronunciation relation and some pedagogical aspects of L2 pronunciation teaching so far, we now turn to a specific area to which we argue L2 Chinese pronunciation teaching should pay attention. That is the distinction between tone category and tone word learning (Wiener et al. 2018; Pelzl 2019). When CSL learners learn Mandarin Chinese (Chinese henceforth), they not only

need to learn the tonal contrasts, they also need to use the tones linguistically to recognize and produce words. Therefore, to address the challenge of L2 Chinese tone learning faced by CSL learners, we should consider both tone category learning and tone word learning.

Standard Mandarin Chinese known as Putonghua has four lexically contrastive tones that differ in their pitch height (low, high) and contour (rising, falling, or dipping). By convention, the four tones are labeled with numbers, and most learners and teachers talk about the tones using either the numbers (1,2,3,4) or tone diacritic (ā á ǎ à). The first tone (T1) is a high-level tone (e.g., mā ‘mom’). The second tone (T2) is a rising tone (e.g., má ‘hemp’). The third tone (T3) is realized as a low-falling tone in most contextualized speech, but also can occur as a low-dipping tone in isolation or at the end of a phrase (e.g., mǎ ‘horse’). The fourth tone (T4) is a falling tone (e.g., mà ‘to scold’). Most of the minimal pairs in Chinese are monosyllabic morphemes/characters (e.g., 大 da4 ‘big’ versus 打 da3 ‘to hit’). In modern Chinese, most words are disyllabic words (Duanmu 2007). So far, there is no estimate about the percentage of minimal pairs formed by disyllabic words contrasted by tones (e.g., 大水 da4-shui3 ‘flood’ versus 打水 da3-shui3 ‘to fetch water’). Based on some mainstream Chinese pronunciation teaching textbooks (e.g., Di and Rong 2012), we estimate that there is a small portion of disyllabic minimal pairs that are contrasted by tones. It is much easier to find near-minimal pairs contrasted in tones among disyllabic words (e.g., 大水 da4-shui3 ‘flood’ versus 打牌 da3-pai2 ‘play cards’ where the first syllables contrast in tones) than finding disyllabic minimal pairs. To teach pronunciation of tones in L2 Chinese, the simple citation forms of tones may be useful at the beginning. Ultimately, learners need to deal with contextual tone changes in disyllabic or multi-syllabic words. If we consider the simplest contextual case of two syllables, a tonal coarticulation will cause the contour tone not to be fully articulated, or it takes extra time for a tone of one syllable to be fully realized within itself, thus, realized across syllable boundary, known as peak delay (Xu 1997). The clearest examples are when a tone with a high or low offset precedes a tone with the opposite onset (e.g., a falling T4 ends low, and a level T1 starts high), in this case, the shape of both tones could be strongly influenced. Research with native speakers (Xu 1997) and some work with L2 learners (Yang 2016) has begun to investigate this type of circumstance.

Numerous studies have shown that disyllabic tone identification is more difficult than monosyllabic tone identification, and that initial syllables cause more difficulty overall than final syllables (Broselow, Hurtig, and Ringen 1987; Chang and Bowles 2015; Hao 2012, 2018; Sun 1998). Given the fact that perceiving tones in disyllabic words is more challenging, it should not be surprising that when instructors give dictation of disyllabic words to CSL learners in class, many beginner-level learners can write the consonant and vowel letters for the pinyin (romanization) of Chinese words, but are unable to provide the tone diacritics (ˉ ˊ ˇ ˋ). In other words, it is possible that a learner can recognize a Chinese word without perceiving or knowing the tones of the syllables in the word. We consider this phenomenon to be incomplete learning for both tone category and tone word learning, but it is nevertheless a real phenomenon that captures a possible stage in the L2 learning of any given Chinese

word. Given that the outcome most learners and teachers really care about is word recognition, it is not enough to focus just on tone category learning. However, research on the relation between tone category and tone word learning is overall scarce.

Tone word learning has a second important component. The learner needs to be able to encode the tone with a word's mental representation in long-term memory so that the tone can contribute to successfully recognizing that word when it is heard again in the future. The psycholinguistic model of lexical representations claims that lexicon in the first language (L1) and second language (L2) words consist of a lemma (i.e., syntax, semantics) and lexeme (i.e., morphology, phonology, orthography; see Levelt 1993). Jiang (2000) proposes that the first or formal stage of an L2 lexical entry involves only phonological and orthographical information. L2 lexical development is characterized by its conscious and deliberate learning. For most adult L2 learners, this lexical learning initially results in weak connections between L2 form and meaning (see Jiang 2018 for a review). By taking into account the lexical development, we aim to use pronunciation teaching to facilitate CSL learners' vocabulary learning by enhancing the association among pronunciation, orthography (both pinyin and written characters), and meaning.

On the one hand, we recognize the fruitful phonetic and psycholinguistic research on the acquisition of Chinese tones in the past few decades. On the other hand, the research findings somehow are not directly translated into the pedagogical application to a large extent. Interestingly, the choice of settings for pronunciation teaching research appears to have changed. Namely, studies of pronunciation teaching have migrated over time from laboratories to classrooms, a shift often seen in other social sciences where experimental effects are explored in low-stakes environments before testing them in applied contexts such as classrooms (Oswald and Plonsky 2010; Lee et al. 2015). The current study follows the trend as we experiment with a systematic pronunciation teaching paradigm throughout a semester of a beginning Chinese course. The data we present in this chapter is a subset of the data we collected where all the target words used in the pronunciation teaching are the words students have learned in the textbook.

1.2 Current Study: Integrate Pronunciation Teaching in the Curriculum

As many studies have consistently found that disyllabic tone identification is more difficult than monosyllabic tone identification, and that initial syllables cause more difficulties overall than final syllables do (Broselow et al. 1987; Chang and Bowles 2015; Hao 2012, 2018; Li 2016; Sun 1998), the current study tries to improve beginning CSL learners' perception and production of disyllabic words in L2 Chinese. As mentioned earlier, one goal of the current study is to make pronunciation teaching a part of the curriculum. In the literature on pronunciation instruction, it is not uncommon to see an overarching research question: what are the *most* effective

training techniques in L2 pronunciation? We think this is a question impossible to answer. We would rather ask an alternative question: given the resources available such as manpower, number of course credit hours, language lab, software, application, and so on, how can instructors customize the pronunciation teaching so as to fit in their curriculum? Plonsky and Oswald (2014) warn that a strong correlation between production training length and effect size may put into question the practicality of such interventions. In other words, instructional costs (time and energy) must be weighed against their potential benefits for L2 learners. It has been argued that teaching metalinguistic knowledge of Chinese phonology and phonetics can facilitate CSL learners' production (Liu 2019). Depending on the make-up of the instructor team for a Chinese course, often instructors may have never taken a Chinese linguistics course, let alone a phonetic course. Therefore, it requires some sort of teacher training for all instructors for them to teach metalinguistic knowledge about pronunciation to learners. But in reality, based on our experience and discussion with instructors from other Chinese programs, the inexperienced instructors often deliver explicit pronunciation teaching not as good as we expected. That is the cost and problem from the delivering end. On the receiving end—we often found students were easily bored with metalinguistic terminology (e.g., tone register feature, contour, syllable, etc.) unless certain students are really interested in Chinese linguistics. Even worse, some students cannot tolerate listening to their own recordings. With all these practical issues in pronunciation teaching, the current study tries an alternative form of explicit teaching, utilizing a self-paced learning paradigm on Blackboard (a centralized learning platform widely used in the USA) to direct learners' attention to the specific tonal contrasts in disyllabic words. We asked the learners to listen and produce the target tone words during the training. As the spectrum of explicit instruction for speech perception and production can range from providing minimal corrective feedback (e.g., informing the correctness of the responses) to teaching metalinguistic knowledge about phonology and phonetics (Chandrasekaran et al. 2016; Lee et al. 2015), in the current, study we tried out the attention-directing approach by using both audio and orthography prompts to highlight the tonal contrast. We tried to inform the learners that when a syllable is combined with different tones, their meanings change together with their orthography and the context where they occur. As all the training stimuli and instruction were made online in the self-paced training paradigm, it minimizes the instructors' roles and maximizes learners' exposure to the tonal contrasts in monosyllabic and disyllabic words. We split beginner-level CSL learners in five different sections into two groups. One is the traditional teaching group considered as the control group as they did not receive any extra perceptual or production training and the other is the experimental group who underwent the self-paced training.

In terms of the training stimuli used for the experimental group, with the idea of using pronunciation teaching to help learners' tone word learning, we selected a list of monosyllabic and disyllabic words from the chapters students have learned in the textbook (Integrated Chinese Level 1 Part 2, 4th Edition, Liu et al. 2009). We selected 4–6 disyllabic words from each chapter as the target words. The syllable of the first or second character in a word chosen overlapped with the syllable of a character in

another disyllabic word that was learned in the previous chapters. For example, in a chapter where the theme is weather, we selected a disyllabic word 滑冰 hua2-bing1 ‘to skate’ as the target word and used another previously learned disyllabic word 花钱 hua1-qian2 ‘to spend money’ to form a near-minimal pair in which the first syllables in the two disyllabic words share the same syllable ‘hua’ but contrast in tones (e.g., hua2 versus hua1). Both groups audio recorded the target words in a pretest, immediately administered posttest, and delayed posttest. By using the tonal minimal pairs (e.g., hua2-hua1) and near-minimal pairs (e.g., hua2-bing1 versus hua1-qian2), we expect the experimental group to be able to direct their attention not only to the tonal contrast in citation forms but also in disyllabic context.

On the one hand, we tried to use the disyllabic words that carry tonal contrasts for certain syllables to improve the learners’ tone production accuracy. On the other hand, we tried to enhance learners’ word learning such as the memorization of the orthography-pronunciation-meaning association of the newly learned words. As previous research has found, beginning-level CSL learners are surprisingly good at using their prior phonological and lexical knowledge to learn new tone words in L2 Chinese (Liu and Wiener 2020). We believe that tone categorization and tone word recognition are intertwined with each other during the learning process. It is entirely plausible that using recently learned words as training stimuli in pronunciation teaching can benefit both tone category and tone word learning. To test whether the perceptual and production training can also improve learners’ word learning, we had both groups dictate the newly learned words (e.g., 滑冰 hua2-bing1 ‘to skate’) twice, one before the tone perception and production training as a pretest and one after the training as a posttest. We used the dictation performance to measure the word learning. We expect the experimental group to retain the memory of the newly learned words better because they had been exposed to the words in the training while the traditional group did not receive the training.

1.3 Technological Component of the Current Study

The self-paced tone perceptual and production training paradigm was implemented on Blackboard (a centralized online learning platform widely used by universities in the USA) and the VoiceThread (2020) app embedded in Blackboard. In SLA research, sometimes technology has been used to complement teacher- or researcher-delivered instruction (Lord 2008); in others, a computer program is the sole provider of instruction (Hardison 2005). We put all the model speech sound stimuli recorded by a female native Chinese speaker on the Blackboard website for a tone identification task to which corrective feedback was provided. Then we asked learners to record the target words at three different times in VoiceThread where all recorded stimuli can be uploaded by students and downloaded by the teacher (see details in the methods section). All these tasks were completed in lab sessions.

1.4 Pronunciation Assessment

We then used the Qualtrics web survey tool to ask inexperienced native listeners (native Chinese speakers who rarely listened to foreigners' speaking of Chinese) to rate the word comprehensibility (how easy an utterance can be understood, see Derwing and Munro 2015). The reason we used inexperienced listeners' comprehensibility rating as the assessment of the pronunciation is two folds. First, Levis' (2005) proposal to abandon the goal of being nativelike in L2 pronunciation, intelligibility, comprehensibility, and foreign accent have been widely adopted in the assessment of L2 pronunciation (see Saito and Plonsky 2019 for a review). Therefore, the current study adopted comprehensibility as the production measurement. Second, previous research has shown that experience with foreign accent affects comprehensibility rating (e.g., Issacs and Thompson 2013; Saito and Plonsky 2019). Typically, foreign-accented Chinese tend to be more comprehensible to L2 Chinese instructors than to native Chinese speakers who have little experience with accented Chinese. Ultimately, the learners will go outside the classroom to communicate with native Chinese speakers who are likely inexperienced listeners. Therefore, we want to see how effective our training method is in terms of increasing learners' comprehensibility judged by inexperienced listeners.

1.5 Word-Specific Effect on Pronunciation

In tone study, in general, researchers find some tones are more difficult to learn than others, and there is an order of tonal acquisition. Although the orders of perception and production are not exactly the same, a consistent finding across a wide variety of studies is that, at least in isolated syllables, T2 is the most difficult tone for learners to identify, typically followed closely by T3 (Hao 2012; Lee et al. 2013; Pelzl et al. 2019; Sun 1998). At the same time, concerning T3, studies have occasionally reported dramatically different results in which it appears to be the easiest tone (Chang and Bowles 2015). Few studies explored or reported whether there exists a tendency that when the same tones combined with different syllables, the comprehensibility varies. In the current study, we examined whether such a word-specific effect on comprehensibility rating existed. However, we were only able to describe the comprehensibility ratings across different disyllabic words because with the current word set we selected, it is hard to tease apart the segmental and tonal contribution to the comprehensibility rating.

In short, in the current study, we recycled the vocabulary that learners newly and previously learned as the stimuli for pronunciation teaching as we aimed to use pronunciation teaching to facilitate both tone category and tone word learning. In training, we used minimal and near-minimal pairs together with prompts to direct learners' attention to specific tonal contrasts in monosyllabic and disyllabic contexts. We try to answer the following three research questions.

2 Research Questions

1. Does the experimental group (a combination of perception and repetition tasks) outperform the traditional group (dictation only) in terms of disyllabic word comprehensibility in the immediate and delayed posttest?
2. Is there a tendency that comprehensibility ratings vary across words?
3. Does pronunciation teaching help to improve word learning assessed in the dictation tasks?

3 Methods

3.1 Participants

66 native English speakers (30 male; 36 female; mean age = 20; SD = 0.8; age range: 18-24) enrolled in a second semester beginner-level Mandarin Chinese class at a public US university participated in the study. All participants started to learn Chinese from college and had completed roughly 11 weeks of formal classroom instruction at the time of the study. All participants self-reported normal hearing and normal or corrected-to-normal vision. The participants were split among two groups defined by the training conditions. 32 participants were assigned to the traditional teaching group (traditional) and 34 participants were assigned to the experimental group (tone identification + word repetition). In the data analysis, we discarded 16 participants' data either because they are heritage speakers, or some of them did not participate in all three recording sessions. The remaining 50 participants represent data across the two groups (25 in the traditional group and 25 in the experimental group).

3.2 Training Materials and Test Instruments

Throughout the semester, we had a total of six pronunciation teaching sessions to train the experimental group's pronunciation of the words selected from four chapters in the textbook (Integrated Chinese). Right after students finished a chapter, we set up a lab session to practice the listening and pronunciation of the words learned in that chapter. Two weeks later, we had another lab session that taught the same target words again but with shortened duration so that it allowed time to teach the pronunciation of the newly learned words from another chapter. Thus, the words selected from a chapter were trained in two lab sessions with two weeks in between. For the current study, we reported the pronunciation teaching result from one chapter.

To prepare the model speech, we asked a female native speaker to record four monosyllables (e.g., 滑 hua2 'to skate') in citation form together with another four

monosyllabic words that only differ in tones (e.g., 花 hua1 ‘to spend’). Then we asked the speaker to record four disyllabic words in which the first or second morphemes were the monosyllabic words that have been recorded (e.g., 滑冰 hua2-bing1 ‘to skate’ and 花钱 hua1-qian2 ‘to spend money’). Half of the words were just learned in the new chapter and the other half were learned in the previous chapters. Thus, all words are meaningful to learners (see Appendix for the full list of words used). All the disyllabic words formed near-minimal pairs except one pair was a minimal pair (回去 hui2-qu4 ‘to return’ versus 会去 hui4-qu4 ‘will go’). In total, eight monosyllabic words and eight disyllabic words were recorded as target words. These recorded words were used in the tone identification and word repetition tasks.

For the pronunciation task, both groups read aloud and audio recorded the 16 words in a pretest, immediate posttest, and delayed posttest by using VoiceThread. Participants clicked the ‘comment’ on each prompt slide and recorded themselves using a headset. When recording, all words were presented in pairs with pinyin, characters and English translation in slides within VoiceThread (see Fig. 1). At the beginning of the recording task, we had instructions shown on the screen that explicitly asked the learners to pay attention to the tonal contrast of a syllable highlighted at the top. Learners recorded the stimuli in pairs. All the recorded sound files were downloaded in mp3 format with good sound quality. We then used Praat (Boersma and Weenink 2018) to make individual word audio files for comprehensibility rating. All sound stimuli were scaled to 70 dB to ensure comparable volume.

3.3 Procedure

Both groups were assigned a recording homework to record all the 16 target words by using VoiceThread. This recording homework was used as the pretest for the pronunciation task. Within the same week, when learners completed the pretest, the



Fig. 1 Screenshot of prompt used in the recording task in VoiceThread

experimental group participated in the lab session where they were first given a tone identification task for all 16 target words by marking the tones of all syllables in the monosyllabic and disyllabic words. To make the learners focus on the tonal contrast, we presented both monosyllabic and disyllabic words in pairs. The learners only needed to mark the tone number (i.e., 1, 2, 3, 4) next to each syllable. This tone identification task was carried out on Blackboard websites. The feedback informed the learners whether their answers were right or wrong and provided with the correct answers. Following the tone identification task, the experimental group logged onto VoiceThread and started to practice the pronunciation of the target words by repeating model speech utterances provided to each prompt (e.g., Figure 1). During the repetition task, they did not record themselves. They could replay the model speech if they wanted to. After finishing the word repetition task, we asked the learners to redo the tone identification task. Then they were asked to record the target words on VoiceThread. This time there was no model speech. This round of recording was served as the immediate posttest. The lab session lasted about 40 min. On the same day, the traditional group also came to the lab session and did the same recording task as the experimental group did. But they did not do tone identification and word repetition tasks. Two weeks after the first lab session, both traditional and experimental groups came to another lab session. They recorded the target words on VoiceThread for the third time as the delayed posttest.

Before the pronunciation pretest, both traditional and experimental groups were given a dictation of four disyllabic words that were used in the recording task (Disyllabic words in column 2 in the Appendix, e.g., 回去 *hui2-qu4* ‘return,’ which was newly learned but not its counterpart 会去 *hui4-qu4* ‘will go,’ which was learned in previous chapters in the textbook) and another four disyllabic words that were learned in the same chapter but were not used in the recording task (e.g., 暖和 *nuan3-huo* ‘warm’). This dictation served as a pretest of the word learning task. One week after the pronunciation delayed posttest, the same dictation was given for a second time to serve as the delayed posttest for the word learning task.

3.4 Data Coding

To measure the comprehensibility of target words, we used Qualtrics online survey tool to construct an online survey for comprehensibility rating. For the rating task, we recruited five inexperienced native listeners who were college students living in mainland China at the time of study who rarely interacted with foreigners who speak Chinese. The raters listened to the speech tokens produced by the L1 baseline (the female speaker who recorded the target words) and the CSL learners. The eight monosyllabic words and eight disyllabic words produced by each participant in each test were extracted from each participant’s utterances and then digitalized at 44,100 Hz using Praat (Boersma and Weenink 2018). As a result, a total of 2416 (16 words × 50 participants × 3 tests = 2400, 16 words × 1 L1 baseline = 16) were extracted. In a pilot study, we found a ceiling effect of monosyllabic words in terms

of comprehensibility rating (the comprehensibility rating of learners' monosyllabic word tokens was close to native speakers' word tokens). Thus, in the current study, we only reported the comprehensibility rating of 1208 tokens (1200 disyllabic tokens produced by learners + 8 L1 baseline tokens). The 1208 tokens were divided into four blocks (308 tokens in Block 1, 300 tokens for Block 2, 3 and 4). The tokens were randomized in each block. For each token, the L1 raters were given a nine-point Likert scale appeared along with the following instruction: 'Judge how good the pronunciation is between 1 (very easy to understand) and 9 (cannot understand at all).' For each token, the Chinese orthography of the words was given. The raters could replay the audio files and were allowed to change their choices until they submitted their response. It took about 40 min to complete one block. To make sure the consistency of the rating criterion, we asked the L1 raters to rate the comprehensibility of all four blocks. But we sent the survey links to the raters on four consecutive days. They rated one block per day to avoid fatigue and familiarity effect. The raters received compensation for doing the comprehensibility rating.

For the dictation, learners were asked to write down pinyin, characters and meaning of the target words. In the data analysis, we only counted the number of correct responses on tones and meanings, leaving out the character writing because we focused on the spoken word-to-meaning mapping. When learners transcribed the pinyin (segments + tone), overall they wrote the segmental part (initials + finals) correctly but with some minor errors. That is why we did not include the segments' accuracy in the data analysis. There were two conditions in the dictation: words used in the recording tasks and words not used in the recording tasks. We only counted the correct responses on tones and meanings by ignoring the segments (4 disyllabic words \times 2 tones in each word, thus, 8 target tones + 4 word meanings). Each correct tone and word meaning received 1 point. Thus, for each condition, the full score is 12. Learners had the same dictation at two different time points with three weeks apart.

4 Results

4.1 Production Results

The Cronbach α was calculated in order to verify in interrater agreement among the five inexperienced raters. The raw Cronbach α was 0.8 (95% CI: 0.77–0.82). The reliability indexes were considered acceptable, following the benchmark value of 0.70–0.80 in L2 research studies (Larson-Hall and Herrington 2010). Thus, by averaging all raters' scores, one mean score for each target word at three testing times was computed as the comprehensibility score for each word per participant (see Lee and Lyster 2017 for the same data verification procedure). In addition, the L1 baseline participant showed ceiling effects for the target words (Mean = 1.2, SD = 0.03).

Figure 2 plots the comprehensibility ratings of each word token (points with jitter added), along with density, group means (solid lines), and group 95% confidence intervals (white box) for both groups in each test. We can see that the comprehensibility score gets lower (thus easier to understand) for the experimental group in the immediate and delayed posttests, whereas the traditional group’s comprehensibility score did not go down in the two posttests.

To test whether the condition, test or their potential interaction affected the comprehensibility ratings, a mixed-effects regression model was built using the lme4 package in R (version 3.6.2; R Core Team 2019). The model contained test as a continuous variable, condition a sum coded factor (1, -1). These two main effects and all corresponding two-way interactions were included in the model. Random word and subject intercepts were included. Table 1 reports the model and R code along with 95% confidence intervals for the coefficient and a standardized coefficient.

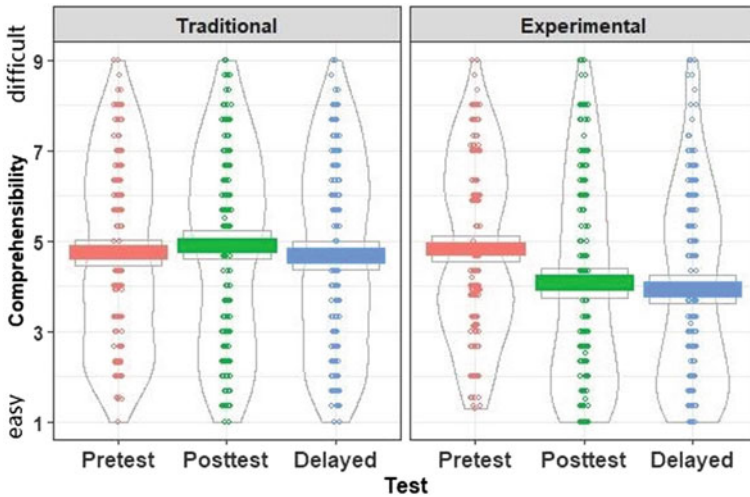


Fig. 2 Comprehensibility rating by participant (point), test (color), condition. White boxes represent 95% confidence intervals with solid color line representing mean. ‘1’ indicates very easy to understand, ‘9’ indicates cannot understand at all

Table 1 Mixed-effects linear regression model output for comprehensibility rating

Parameter	β	SE	95% CI	t	p	Std_β
(Intercept)	5.02	0.31	[4.41, 5.63]	16.12	<.001	0.01
Test	-0.24	0.07	[- 0.37, -0.11]	-3.63	<.001	-0.09
Condition	0.14	0.20	[- 0.24, 0.52]	0.72	0.46	-0.12
Test:Condition	-0.20	0.07	[- 0.33, -0.07]	-3.10	<.001	-0.08

```
lmer(Comprehensibility ~ Test*Condition + (1|Words) + (1|Subject))
```

The model revealed a main effect of Test and an interaction of Test: Condition. To explore the locus of the interaction, analyses were carried out in the pretest, immediate posttest and delayed posttest respectively. In the pretest, there was no difference between traditional and experimental groups ($\beta = 0.076$, $SE = 0.23$, $t = 0.32$, $p = 0.74$). This non-significant difference indicated that the two groups had similar pronunciation performance initially. In the immediate posttest, the experimental group had significantly lower comprehensibility rating score (better comprehensibility) than the traditional group ($\beta = -0.87$, $SE = 0.41$, $t = -2.1$, $p = 0.04$). In the delayed posttest, again the experimental group had significantly lower comprehensibility rating score than the traditional group ($\beta = -0.79$, $SE = 0.39$, $t = -2.0$, $p = 0.04$). The result indicated that the experimental group significantly improved their pronunciation after the training and the gain was retained in the delayed posttest whereas the traditional group did not show any gain in the two posttests.

4.2 Comprehensibility Rating Across Words

Although the current study was not designed to test whether learners' pronunciation varies across different words, we just want to report how the comprehensibility ratings may vary across different disyllabic words. Figure 3 illustrates the comprehensibility ratings by test, condition, and words.

As shown in Fig. 3, numerically, it seems that the comprehensibility rating varies across words in all three tests for both groups. For example, the disyllabic words 花钱 *hua1-qian2* 'to spend money' and 滑冰 *hua2-bing1* 'to skate' in general had higher comprehensibility (lower comprehensibility rating score) than other words in all three tests among both groups. The word-specific effect was also manifested in the degree of comprehensibility improvement. For the experimental group, certain words such as 汽车 *qi4-che1* 'automobile' had larger comprehensibility improvement than other words in the posttests. Although we cannot make a strong claim about the

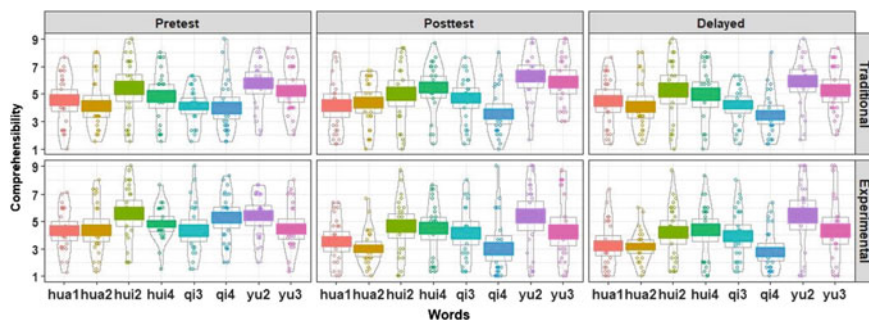


Fig. 3 Comprehensibility rating by tests, condition, and words. White boxes represent 95% confidence intervals with solid color line representing mean. The syllable + tone (e.g., *hua1*) indicates the target syllables that had tonal contrasts in disyllabic words

word-specific effect on the comprehensibility rating based on the current data, the general pattern provides us a glimpse of how comprehensibility could vary across words.

4.3 Word Learning (Dictation) Results

Figure 4 summarizes the dictation results of the two groups in the pretest (before the first recording task) and the delayed posttest (after the third recording task). Here, we included whether the words in the dictation were used in the recording task as an additional fixed factor coded as ‘untrained’ versus ‘trained.’

A mixed-effects linear regression model was built in R following the previously outlined approach and variable coding. The model contained test as a continuous variable, trained status, and condition were sum coded factors (1, -1). Table 2 reports the final model output and R code with 95% confidence intervals for coefficient and a

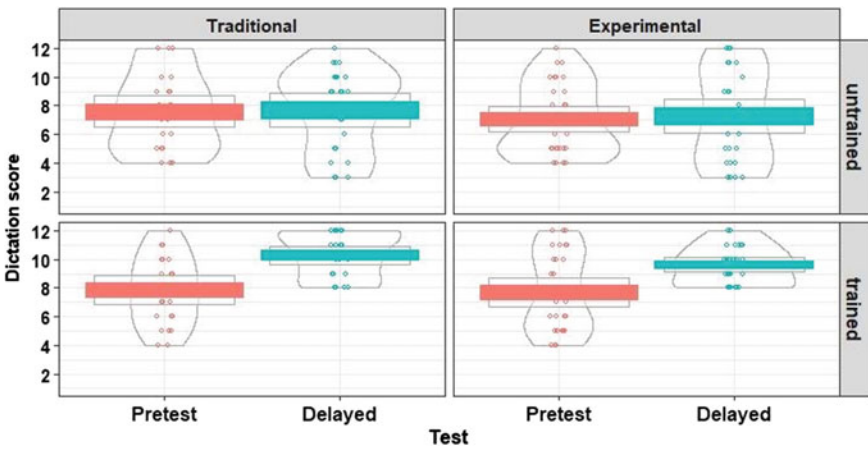


Fig. 4 Dictation score by test, condition and word status of whether appearing in the recording tasks. White boxes represent 95% confidence intervals with solid color line representing the mean

Table 2 Mixed-effects linear regression model output for dictation scores

Parameter	β	SE	95% CI	t	p	Std_β
(Intercept)	8.14	0.17	[7.80,8.48]	47.06	0	0.00
Trained	0.74	0.17	[0.40,1.07]	4.27	<.001	0.27
Test	0.58	0.17	[0.24,0.91]	3.34	<.001	0.21
Condition	-0.22	0.17	[-0.56,0.12]	-1.28	0.20	-0.08
Trained:Test	0.49	0.17	[0.15,0.83]	2.85	<.01	0.18

lmer(Dictation ~ Trained + Test + Condition + Trained:Test + (1|Subject))

standardized coefficient. The model revealed main effects of trained status (positive) and test (positive). A significant trained status by test interaction was found.

To explore the interaction, we subset the data into untrained and trained for analyses. We found the dictation scores did not differ between the untrained and trained words ($\beta = 0.24$, $SE = 0.25$, $t = 0.97$, $p = 0.33$) in the pretest. In the delayed posttest, the trained words had a significantly higher score than the untrained words ($\beta = 1.22$, $SE = 0.21$, $t = 5.9$, $p < 0.001$). The result showed that the condition did not affect word recognition accuracy in both pretest and posttest. But the trained status of the words affected the word recognition accuracy.

5 Discussion

In this study, we set out a voyage to explore how pronunciation teaching can be integrated into a Chinese curriculum to improve learners' pronunciation and memorization of disyllabic words that included tonal contrasts. Rather than focusing exclusively on the often-debated segmental/suprasegmental distinction, the results of our study support an approach that uses learned vocabulary as the stimuli to train CSL learners' perception and production of tones so that it aligns with learners' needs and proficiency. From an acquisition point of view, CSL learners not only need to learn the tone categories but also need to store the tone category in their mental lexicon. In other words, learners need to associate the tone categories with words in long-term memory. With this overarching theme in mind, we selected a set of syllables (e.g., *hua*) from each chapter in the textbook. Those syllables were combined with different tones to form different morphemes. Those morphemes then appeared in the same position in disyllabic words that learners just learned in a chapter or had learned in previous chapters. Then we asked learners to record these disyllabic words at three time points. The experimental group received tone identification and word repetition training while the traditional group did not receive any extra training. Before and after the pronunciation teaching, both groups did the same dictation of a list of words, half of which appeared in the pronunciation task (recording task) while the other half did not appear in the pronunciation task, so that we can test whether pronunciation teaching helped to improve word recognition. We had three major findings.

First, we found that doing tone identification and word repetition tasks can improve the beginner-level learners' pronunciation of disyllabic words significantly right after the training session. More importantly, such gain was maintained in a delayed posttest two weeks after the training session. The null result found in the traditional group, on the other hand, indicates that merely showing the disyllabic words in pairs with the tonal contrast highlighted in the prompts in a recording task cannot help to improve learners' pronunciation. This is what the aggregated data analysis showed to us. However, by looking at individual learners, we did find four learners in the traditional group had better comprehensibility ratings (lower rating scores) for most of the words in the delayed posttest. Overall, pronunciation practice (recording task) without tone listening exercises or word repetition did not increase learners' comprehensibility.

But for some learners, when being provided with very limited amount of instruction that was used to direct their attention to the tonal contrast in monosyllabic and disyllabic contexts, somehow they picked it up anyway. Of course, it is unknown to us whether they listened to the words by themselves outside the classroom and noticed the tonal contrast patterns in the recording task. But in any case, the improvement shown by those learners in the traditional group encourages us to make an effort to highlight the pronunciation features in class so that the motivated learners can try to improve their pronunciation at their own pace. As a recent meta-analysis study conducted on a large number of pronunciation instruction studies showed, previous research has shown rarely with exception that almost any form of training or extra teaching can improve L2 pronunciation (Lee et al. 2015). Thus, we should realize that doing something is better than doing nothing about pronunciation teaching.

The second finding in this study was that there seems to be a tendency that the comprehensibility rating varies across different words, as shown in Fig. 3. For example, previous tone study usually makes a general claim that T2 is the most difficult to perceive and produce for L2 Chinese learners (e.g., Sun 1998). The fact tends to hold for our data in the pretest. Figure 3 shows that numerically *hui2* had worse comprehensibility than *hui4* in the minimal pair (回去 *hui2-qu4* ‘to return’ versus 会去 *hui4-qu4* ‘will go’) and *yu2* tends to have poorer comprehensibility than *yu3* in the near-minimal pair (小鱼 *xiao3-yu2* ‘little fish’ versus 下雨 *xia4-yu3* ‘to rain’) in the pretest for both traditional and experimental groups. However, as learners did more pronunciation practice, the comprehensibility of *hui2* became more similar to *hui4* in the immediate posttest and delayed posttest whereas the comprehensibility of *yu2* was still worse than that of *yu3* in the two posttests. As for *hua2*, which also carries T2, it had comparable comprehensibility to *hua1* from the very beginning. These cases indicate that the difficulty of a certain tone can be affected by the context in which it appears (e.g., vowel, position in a disyllabic word, etc.). In general, previous pronunciation instruction (PI) research found relatively homogeneous effects of PI on different pronunciation features (Saito 2012). In other words, the relative effects of PI across a range of targeted linguistic features are more or less the same. Our finding though seems not to be aligned with that claim as seemingly there was a word-specific effect on comprehensibility rating. We have to point out that we are not making strong claims about this word-specific effect on syllable + tone production as the current study did not systematically investigate this topic. We mainly described the pattern here for researchers to further explore this issue.

The third finding is that we found simply asking learners to do multiple pronunciation practice sessions (the recording tasks) with a time interval (one week or two) can help to enhance learners’ long-term memory of those words. With a three week gap, learners recognized words that were used in the recording tasks significantly better than the words that never appeared in the recording tasks for both traditional and experimental groups. It indicates the importance of recycling vocabulary in L2 Chinese teaching. Previous research has shown that adult L2 learners, including L2 Chinese learners use statistical learning in word learning (Liu and Wiener 2020; Pelzl

et al. 2019; Wiener et al. 2019). Therefore, the more frequent exposure to previously learned words, even if it is just orthographic as in the traditional group, the better the L2 word recognition could be.

6 Pedagogical Implication

The findings in the current study have several pedagogical implications we want to discuss. The first and foremost is that when instructors decide to implement pronunciation teaching in the classroom, they do not have to depend on ‘fancy’ technology. The advantages of using self-developed computer-assisted training programs for tone learning (e.g., Liu et al. 2011; Wang 2013; Wang et al. 2003) include that the researchers can target at specific tone features and try to boost learners’ perception and production of those target features within a short period of time. The disadvantage of those sophisticated training programs is that it is hard to distribute their training programs to other institutions or even within the same institution as not every Chinese instructor is ‘tech-savvy.’ Therefore, it may not be so easy to incorporate those computer-assisted programs in a Chinese language program. Recently, some research has tried to use existing online learning tools to enhance CSL learners’ tone perception (Xu et al. 2019). The idea of using a publicly accessible tool for pronunciation teaching should be encouraged. The reason we use VoiceThread on Blackboard to carry out the current study is that it is very easy for students to use the online platform to practice their pronunciation and refresh their memory of learned vocabulary. Learners can do the recording tasks inside and outside the classroom. Such asynchronous instruction will get more and more popular given the pandemic period that mankind has been experiencing.

In L2 Chinese teaching, it is easy to notice that most beginning-level learners can perceive and produce the tones in monosyllabic words much easier than the tones in disyllabic words. The simple recast in the classroom may elicit correct or near correct pronunciation of the disyllabic words, but the pronunciation gain is usually short-lived. The same pronunciation errors can easily reoccur. As shown in the current study, using a systematic pronunciation teaching integrated into the curriculum with regular lab sessions in a fixed time interval not only improved learners’ disyllabic word pronunciation right after the training but also helped to retain their pronunciation gain after two weeks of the training. It indicates a long-term benefit of pronunciation teaching is achievable if we can help learners form a habit of doing pronunciation practice while focusing on the specific tonal contrasts. We also see the systematic pronunciation teaching benefited learners’ word recognition in the long run. After three weeks from the initial dictation, learners recognized and memorized the words used in the pronunciation teaching significantly better than those not used in the pronunciation teaching. There has been some study that shows the phonological memory of words is important for vocabulary learning (Martin and Ellis 2012). So highlighting the phonological form of the previously learned words is expected to benefit the long-term memory of the vocabulary.

In pronunciation teaching, the most time-consuming part is to provide feedback to learners because instructors need to listen to the students' recordings first and then either give oral or written feedback. If somehow instructors can find an effective way of collecting native speakers' comprehensibility ratings of students' recorded speech samples and use that as a form of feedback to students, it may help students understand how well their speech can be understood by native Chinese speakers. The comprehensibility feedback is quasi-communicative oriented. We encourage instructors to let inexperienced native listeners judge the comprehensibility of learners' pronunciation instead of just using instructors' intuitive judgment.

All the previous research has informed us that doing something is better than nothing for helping with learners' pronunciation. To teach pronunciation, instructors need to be aware of a range of phonetic features in the target language. A good language teacher should be a good observer first. Through observing learners' pronunciation in the classroom, the teachers will gain an idea about what pronunciation errors are common, and in what context pronunciation errors occur more frequently. Above all, if a systematic pronunciation teaching module can be incorporated into a course curriculum and instructors can identify which words students need to spend more time on training, then students will benefit from the pronunciation instruction.

Appendix

Monosyllabic minimal pairs	Disyllabic word 1 (newly learned)	Disyllabic word 2 (previously learned)	Word1-word2
滑 hua2 'to skate' – 花 hua1 'to spend'	滑冰 hua2-bing1 'to skate'	花钱 hua1-qian2 'to spend money'	Near-minimal pair
汽 qi4 'steam' – 起 qi3 'to get up'	汽车 qi4-che1 'automobile'	起床 qi3-chuang2 'to get up from bed'	Near-minimal pair
雨 yu3 'rain' – 鱼 yu2 'fish'	下雨 xia4-yu3 'to rain'	小鱼 xiao3-yu2 'little fish'	Close to minimal pair
回 hui2 'to return' – 会 hui4 'will'	回去 hui2-qu4 'to return'	会去 hui4-qu4 'will go'	Minimal pair

Note underscored syllables are the syllables with tonal contrast

References

- Best, C. (1995). A direct realist view of cross-language speech perception. *Speech Perception and Linguistic Experience*, pp. 171–206.
- Best, C., & Tyler, M. (2007). Nonnative and second-language speech perception. In O. Bohn & M. Munro (Eds.), *Language experience in second language speech learning: In honour of James Emil Flege* (pp. 13–34). Amsterdam, Netherlands: John Benjamins.
- Boersma, P., & Weenink, D. (2018) Praat: Doing Phonetics by Computer [Computer Program]. Version 6.0.43. Retrieved from <http://www.praat.org>.
- Bradlow, A. R. (2008). Training non-native language sound patterns: Lessons from training Japanese adults on the English /r/-/l/ contrast. In J. G. H. Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 287–308). Amsterdam: John Benjamins.
- Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, 101, 2299–2310.
- Broselow, E., Hurtig, R. R., & Ringen, C. (1987). The perception of second language prosody. In G. Ioup & S. H. Weinberger (Eds.), *Interlanguage phonology: The acquisition of a second language sound system* (pp. 350–364). New York: Newbury House Publishers.
- Chandrasekaran, B., Sampath, P. D., & Wong, P. C. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America*, 128(1), 456–465.
- Chandrasekaran, B., Yi, H., Smayda, K., & Maddox, W. T. (2016). Effect of explicit dimension instruction on speech category learning. *Attention, Perception, & Psychophysics*, 78, 566–582.
- Chang, C. B., & Bowles, A. R. (2015). Context effects on second-language learning of tonal contrasts. *Journal of the Acoustical Society of America*, 136(6), 3703–3716.
- Colantoni, L., & Steele, J. (2008). Integrating articulatory constraints into models of second language phonological acquisition. *Applied Psycholinguistics*, 29, 489–534.
- Darcy, I., Ewert, D., & Lidster, R. (2012). Bringing pronunciation instruction back into the classroom. An ESL teachers' pronunciation 'toolbox' In J. Levis, & K. Lavelle (Eds.), *Proceedings of the 3rd pronunciation in second language learning and teaching conference* (pp. 93–108). Iowa State University.
- Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals*. Evidence-based perspectives for L2 teaching and research. Amsterdam, Benjamins.
- Di, C., & Rong, J. (2012) 现代汉语语音教程 A course for Mandarin Chinese pronunciation. Peking University Press.
- Duanmu, S. (2007). *The phonology of standard Chinese*. OUP Oxford.
- Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 92, 233–277.
- Flege, J. E., Schirru, C., & MacKay, I. R. (2003). Interaction between the native and second language phonetic subsystems. *Speech Communication*, 40(4), 467–491.
- Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, 40(2), 269–279. <https://doi.org/10.1016/j.wocn.2011.11.001>.
- Hao, Y.-C. (2018). Contextual effect in second language perception and production of Mandarin tones. *Speech Communication*, 97, 32–42. <https://doi.org/10.1016/j.specom.2017.12.015>.
- Hardison, D. M. (2005). Second-language spoken word identification: Effects of perceptual training, visual cues, and phonetic environment. *Applied Psycholinguistics*, 26(4), 579.
- Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, 10(2), 135–159.
- Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English/r/-/l/ to Japanese adults. *Journal of the Acoustical Society of America*, 118(5), 3267–3278.

- Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, 21(1), 47–77.
- Jiang, N. (2018). *Second language processing: An introduction*. Routledge.
- Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, 31(3), 368–390.
- Lee, A. H., & Lyster, R. (2017). Can corrective feedback on second language speech perception errors affect production accuracy? *Applied Psycholinguistics*, 38(2), 371.
- Lee, C.-Y., Tao, L., & Bond, Z. S. (2013). Effects of speaker variability and noise on Mandarin tone identification by native and non-native listeners. *Speech, Language and Hearing*, 16(1), 46–54. <https://doi.org/10.1179/2050571X12Z.0000000003>.
- Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, 36(3), 345–366.
- Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). Cambridge, MA: MIT Press.
- Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *Tesol Quarterly*, 39(3), 369–377.
- Li, Y. L. (2016). *Effects of high variability phonetic training on monosyllabic and disyllabic Mandarin Chinese tones for L2 Chinese learners* (Doctoral dissertation, University of Kansas). Retrieved from ProQuest Dissertations and Theses Global. (1800269699).
- Lim, S. J., & Holt, L. L. (2011). Learning foreign sounds in an Alien World: Videogame training improves non-native speech categorization. *Cognitive Science*, 35(7), 1390–1405.
- Liu, J. (2019) Teaching Chinese pronunciation: Explanation, expectation, and implementation. In C. Shei, M. E. M. Zikpi & D. L. Chao (Eds.), *the routledge handbook of Chinese language teaching*. Routledge.
- Liu, J., & Wiener, S. (2020). Homophones facilitate lexical development in a second language. *System*, p. 102249.
- Liu, Y., Wang, M., Perfetti, C. A., Brubaker, B., Wu, S., & MacWhinney, B. (2011). Learning a tonal language by attending to the tone: An in vivo experiment. *Language Learning*, 61(4), 1119–1141. <https://doi.org/10.1111/j.1467-9922.2011.00673.x>.
- Liu, Y. H., Yao, T.-C., Bi, N.-P., Ge, L. Y., & Shi, Y. H. (2009). *Integrated Chinese (中文听说读写) level 1 part 2*. Boston, MA: Cheng & Tsui Company.
- Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, 94(3), 1242–1255.
- Lord, G. (2008). Podcasting communities and second language pronunciation. *Foreign Language Annals*, 41(2), 364–379.
- Martin, K. I., & Ellis, N. C. (2012). The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition*, 34(3), 379–413.
- Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, 30, 85.
- Pelzl, E. (2019). What makes second language perception of Mandarin tones hard? A non-technical review of evidence from psycholinguistic research. Chinese as a Second Language. *The Journal of the Chinese Language Teachers Association, USA*, 54(1), 51–78.
- Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2019). Advanced second language learners' perception of lexical tone contrasts. *Studies in Second Language Acquisition*, 41(1), 59–86. <https://doi.org/10.1017/S0272263117000444>.
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912.
- Saito, K. (2012). Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. *TESOL Quarterly*, 46, 842–854. <https://doi.org/10.1002/tesq.67>.
- Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, 69(3), 652–708.

- Strange, W. (2006). Second-language speech perception: The modification of automatic selective perceptual routines. *Journal of the Acoustical Society of America*, 120, 3137.
- Sun, S. H. (1998). *The development of a lexical tone phonology in American adult learners of standard Mandarin Chinese*. Honolulu, HI: Second Language Teaching & Curriculum Center.
- Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *CALICO Journal*, 28, 744–765.
- Trofimovich, P., Lightbown, P. M., Halter, R. H., & Song, H. (2009). Comprehension-based practice: The development of L2 pronunciation in a listening and reading program. *Studies in Second Language Acquisition*, 31(4), 609–639.
- VoiceThread (2020). Retrieved from <https://voicethread.com/myvoice/>.
- Wang, X. (2013). Perception of Mandarin tones: The effect of L1 background and training. *The Modern Language Journal*, 97(1), 144–160.
- Wang, Y., Jongman, A., & Sereno, J. (2003). Acoustic and perceptual evaluation of Mandarin tone production before and after training. *Journal of the Acoustical Society of America*, 113, 1033–1043.
- Wiener, S., Ito, K., & Speer, S. R. (2018). Early L2 spoken word recognition combines input-based and knowledge-based processing. *Language and Speech*, 61(4), 632–656.
- Wiener, S., Lee, C. Y., & Tao, L. (2019). Statistical regularities affect the perception of second language speech: Evidence from adult classroom learners of Mandarin Chinese. *Language Learning*, 69(3), 527–558.
- Wong, P. C., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, 28(4), 565.
- Xu, Y. (1997). Contextual tonal variation in Mandarin. *Journal of Phonetics*, 25, 61–83.
- Xu, H., Li, Y., & Li, Y. J. (2019). Using online applications to improve tone perception among L2 learners of Chinese. *Journal of Technology and Chinese Language Teaching*, 10(1), 26–56.
- Yang, B. (2019). Some explicit linguistic knowledge for Chinese pronunciation teaching. In C. Shei, M. E. M. Zikpi & D. L. Chao (Eds.), *The Routledge Handbook of Chinese Language Teaching*. Routledge.
- Yang, C. (2016). *The acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice* (Vol. 1). John Benjamins Publishing Company.
- Zhang, H. (2018). Current trends in research of Chinese sound acquisition. In C. Ke (Ed.), *The Routledge handbook of Chinese second language acquisition*. London & New York: Routledge Taylor & Francis Group.

“Repeat After Me”: Is There a Better Way to Correct Tone Errors in Teaching Mandarin Chinese as a Second Language?



Nan Meng

Abstract Blocked practice outperforms random practice in sports and musical training. This study examines whether the principle of blocked practice and random practice applies to correcting L2 Mandarin learners' tone errors. Three treatments were designed: (1) repeating the target word (blocked practice), (2) repeating the target word in the original context (blocked practice combined with random practice), and (3) repeating the target word in a new context (random practice). Nineteen L2 Mandarin learners received treatments to correct their tone errors and then participated in the post-treatment assessments. The results showed that the third treatment was the best among the three, which indicated that repeating the target words without context is not as effective as repeating them in a context when correcting the tone errors. A pedagogical modal is proposed based on the findings. The limitations of the current study and methodological refinement are also discussed.

Keywords L2 tone correction · Pronunciation teaching · L2 mandarin acquisition

1 Introduction

In recent years, there has been a great deal of second language acquisition (SLA) studies focusing on the instruction of L2 pronunciation. When it comes to teaching Mandarin Chinese as a second language, many teachers and researchers focus on L2 tone acquisition as it is one of the most challenging areas for L2 learners, especially native English speakers. Various innovative pedagogical techniques have been discussed in recent studies of this area. Shih et al. (2010) adopted a computer-aided pronunciation training (CAPT) program and found varied input helped L2 Mandarin learners to improve tone recognition. Similarly, Godfroid et al. (2017) examined L2 learners' perception of tones and claimed that using visual aids plus audio input provided better chances for learners to acquire tones. Morett and Chang's

N. Meng (✉)
University of Connecticut, Storrs, CT 06269, USA
e-mail: nan.meng@uconn.edu

study (2015) indicated that pitch gestures enhanced English speakers' discrimination Mandarin words differing in tone.

Perception of Mandarin tones is closely connected with production. Elliot (1991) studied the relationship between perception and production of Mandarin tones by L2 speakers and found eighty-five percent of the subjects perceived tones better than they produced the same tones. Yang (2012) further examined the gap between the perception and production of tones by American learners of Mandarin and argued that tones are perceived at the phonological level and produced at the phonetic level.

As for L2 tone production, many researchers collected data with L2 adult learners in both laboratory and classroom environments. Zheng et al. (2018) investigated the impact of metaphoric actions—head nods and hand gestures—in producing Mandarin tones for first language and second language speakers. They found hand gestures helped Tone 4 production by L2 learners, and light head nods modestly benefit Tone 3. Besides metaphoric actions, some researchers are interested in orthographic effect on L2 tone production. For example, Mok et al. (2018) found Pinyin system was more beneficial for processing monosyllabic words whereas disyllabic words were better processed in Chinese characters. Their study also revealed orthography effects varied according to tasks, materials, and proficiency levels. Wiener et al. (2020) examined how explicit instruction of tone contours and high variability phonetic training affect the production of L2 Mandarin tones. Their findings suggested an overall accuracy improvement of Tones 2 and 4 after the explicit instructions were used. This study is in line with other previous studies about the targeted pronunciation instruction that improves L2 learner's speech (Lee et al. 2014). Another study about tone production was conducted by Wiener et al. (2019) about non-speech auditory analogs of Mandarin tone categories and incidental learning videogames. They found incidental learning transferred to affect learners' reading aloud of tones and resulted in more native-like tonal contours. Non-speech "perceptual building block" appears to support classroom learning of difficult-to-acquire L2 speech sounds.

There has been an abundance of studies about the importance of correcting pronunciation errors by L2 learners of English (ESL or EFL). Gumbaridze (2013) emphasized the correction technique was essential in EFL speaking classrooms because if it was not chosen in a proper way it can unintentionally upset students' confidence in fluency. When correcting tone errors made by Chinese as second language (CSL) learners, many teachers use the audio-lingual method: identifying the error in speech, providing a model of the target word, and asking learners to repeat after the model. Duff and Li found (2004) that L2 learners held a strong view about error correction, repetition, or modeling in Mandarin language instruction. Repetition with a model addressing the target word may be effective right away, but it could still recur when the target appears again in a different context.

This problem makes many CSL learners experience the obstacle in their learning. The relative difficulty of retaining the correctness is related to the contextual tonal variability. This was proved for learning disyllables (Chang and Bowles 2015), and third tone variants (Zhang 2018). Xu (1993) claimed the adjacent pitch values disagreeing across syllable boundaries may greatly change a tone from its canonical

form, sometimes severely enough to even alter the direction of the tonal contour. In addition, according to Yang (2016, pp. 21), “intonation or other pragmatic functions may influence the actual realization of tones”; namely, suprasegmental can completely override the lexical tones.

Tones themselves are difficult, and context makes it even more troublesome. What are the effective ways to correct tone errors and help L2 learners of Mandarin Chinese retain the accuracy despite changing contexts? If we jump out of the SLA box, we will find people in other fields struggling with the same problem. For example, athletes are trained to shoot the ball and score the goals in different positions, namely contexts. Musicians must practice numerous times in order to get a perfect pitch or tone in different pieces. According to Gebrian (2016), performing on stage is totally a different context, which often can be a mountainous challenge for string players. Both professional athletes and musicians must practice in the right way in order to achieve reliable performance in games or on stages.

Just like L2 learners, athletes and musicians use different kinds of drills. Blocked practice is mechanically repeating the same target movement/skill in a certain period. That is the adage of “practice makes perfect.” Blocked practice belongs to lower level of cognitive interference and builds up stabilization, whereas random practice refers to practice different target movements/skills in a mixed manner, which involves a high-level cognitive interference and adaptation to learning process. According to Corrêa et al. (2014), adaptation occurs when the context changes, challenging its stability and causing uncertainties. Therefore, in order to achieve higher levels of learning and retention in different contexts, random practice outperforms blocked practice.

Researchers analyze the underlying psychological reasons and argue that random practice is a better way than blocked practice when preparing for a reliable performance, for example, playing soccer (Williams and Hodges 2005), playing the viola (Gebrian 2016), and practicing medical surgeries (Goldin et al. 2014). An increasing number of studies in language disorder also show this principle applies to complex tasks such as language learning (Cherney et al. 2018). However, the extension to which the random practice principle applies to the retention of the tone accuracy in L2 Mandarin learning requires investigation.

This study thus aims at examining if the principle of blocked practice and random practice applies to correcting L2 learners’ tone error. There are various ways of providing corrective feedback to L2 learners, but this study only focuses on one type, modeling and repeating after it. Repeating in this paper refers reiterating after a native speaker’s demonstration of the correct articulation. Also, the paper does not touch upon tone production or perception, but only addresses tone error correction. Three treatments for the tone errors are designed: repeating the target word (blocked practice), repeating the target word in the original context (blocked practice combined with random practice), and repeating the target word in a new context (random practice). The question is which treatment results in a higher correction rate in the post-treatment assessment, which is to adapt the target words in new contexts. In this paper, tone accuracy refers to the acceptability by native speakers.

2 Research Method

2.1 Participants

Nineteen L2 learners of Mandarin Chinese were recruited to participate in this study from a large public university in the USA. Because they were all college students, the researcher categorized them as adult L2 learners. Nine of them had already finished first-year Chinese course sequence, *Elementary Chinese I and II*, and ten finished the second-year sequence, *Intermediate Chinese I and II*. All participants were non-heritage L2 learners, and eighteen were English native speakers. There was one participant whose first language was Vietnamese. Fourteen participants had the experience of learning another foreign language prior to studying Chinese.

a. Procedures

The read-aloud stimuli used in this study contained ten sentences taken from the first year textbook used in the Chinese program at the university where the participants were recruited. The sentences were written in Chinese characters and listed with Pinyin and English translations (see Appendix at the end of the paper). After signing the consent forms, the participants were asked to read aloud the ten sentences and audio-record themselves. Three words containing tone errors from each participant's recording were selected to receive treatments. These words varied in length, either two or three syllables long, and had different parts of speech, such as nouns, verbs, and pronouns. All participants received all three treatments, but they were randomly assigned to the three words, aiming to correct the tone errors. These three treatments were as follows:

Treatment 1: Repeating the word containing the target syllable three times in a row after the model.

Treatment 2: Repeating the word containing the target syllable twice and then repeating it in the original context after the model.

Treatment 3: Repeating the word containing the target syllable twice and then repeating it in a new context after the model.

For example, if the participant made an error in the third tone 美 měi as in 美国人 měi guó rén, *American*, the three treatments will be:

1. Repeating the word containing the error 美国人 měi guó rén after the model.
2. Repeating the word 美国人 měi guó rén twice and then repeating it in the original context, for example, 我是美国人 wǒ shì měi guó rén “*I am an American*” after the model.
3. Repeating the word 美国人 měi guó rén twice and then repeating it in a different context, for example, 有几个美国人 yǒu jǐ gè měi guó rén “*There are several Americans*” after the model.”

After all the treatments, the participants were asked to read aloud four sentences containing the words that received each treatment, twelve sentences in total, as the

post-treatment assessment. In these four sentences, the words containing the target syllables appeared in different positions, at the beginning, in the middle, and at the end. These sentences represented functions and structures taught at the beginning level of Mandarin Chinese. This way the participants had no difficulty understanding them. In addition, the post-treatment assessment consisted of different sentence types: declarative, imperative, exclamative, and interrogative. This was because the contour and intonation of a sentence may influence the tones as well. Some sentences were negative while others were affirmative or positive, again to provide a more varied sentence pattern. In some cases, two or more sentences would serve as a short conversation so that they sound like natural speech.

All the sentences used in the post-treatment assessments were presented to the participants in Chinese characters with Pinyin and English translations (see the samples in the Appendix). Therefore, if the participants knew all the characters, they do not need to use Pinyin at all. The participants were given three to five minutes to read through these sentences to themselves, either silently or aloud, before they were ready to “formally” read aloud the sentences as the assessment.

The whole process, pre-treatment reading, the three treatments, and the post-treatment reading aloud were conducted by the researcher via Zoom and were audio-recorded using the same platform. All of the instructions were given to the participants in English during the data-collecting process. They were told that if they had any questions, they could stop and ask the researcher at any time. After the post-treatment assessment was finished, the researcher helped correct the tone errors in the assessment, which was not recorded because it was not part of the research.

A Chinese native speaker with linguistic background served as the evaluator in the post-treatment assessment. The audio recordings of the post-treatment assessments were played to him only once with 5 s intervals in-between sentences. There were no scripts presented as the audio played back, but the evaluator had a list of the target words so that he would have a clear idea of what to focus on. The evaluator was tasked to decide whether the target words in these sentences had tone errors and then mark his judgment on the list. There are only two choices for the evaluator: correct and incorrect.

3 Results

After the evaluator finished grading the list of each participant, a correct percentage for each individual treatment was calculated. ANOVA was used to determine whether there were statistically significant differences between the three kinds of treatments. The averages of the correct percentages for these three treatments are all higher than 50%, which means the targeted instruction and correction were generally effective in correcting tone errors. The third treatment has the highest average correction rate (78%). The average correction rate of the second treatment (57%) is slightly higher than that of the first (54%). However, the variance of the second treatment is the highest (0.124) of all three and that of the third is the lowest (0.041), which

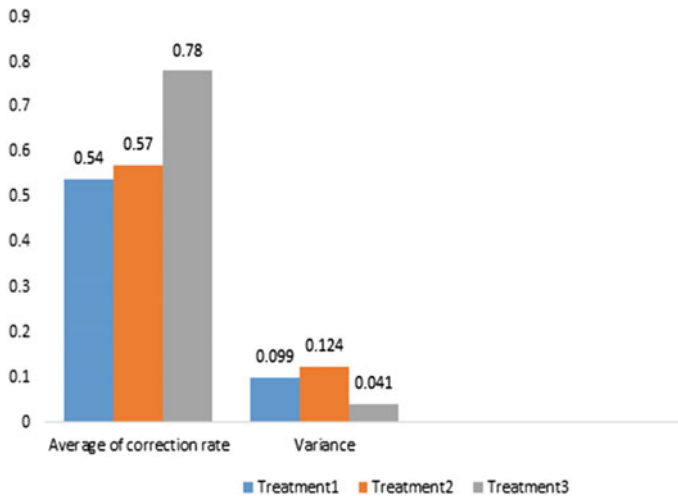


Fig. 1 Averages of correct rates and variances

means the third treatment is consistently more effective than the other two. Among these three treatments, the second treatment is the least consistent despite having the average correct rate higher than that of the first treatment (Fig. 1).

ANOVA conducted on the correct rates of tone productions by the three groups confirmed the main effect of treatment. The following pairwise Tukey HSD showed that the Treatment 3 outperformed both Treatment 1 ($p < 0.01$) and Treatment 2 ($p < 0.05$).

To summarize, the best treatment in this study is repeating the target word and applying it in a new context. The post-treatment assessment results of this treatment are statistically different from those of the first treatment, repeating the target only, and it is more effective than the second treatment. The second treatment, repeating the target word and using it in the original context, also shows statistical difference from the first treatment but it is not consistent.

4 Discussion

Based on the results of the above statistical analysis, it is concluded that repeating the target words without context is not as effective as repeating them in a context when correcting the tone errors. Repeating the target words with the model is helpful but not effective in maintaining the correctness. It could be the first step to correct the tone errors, but it is not enough for L2 learners to retain the accuracy and truly improve the pronunciation.

The results are in accordance with research about blocked practice versus random practice in training motor and music skills mentioned in the introduction. When

correcting tone errors, linking the correctness with new contexts is a form of random practice, which occurs in the high-level learning and thus will help learners to maintain the correct form.

Practice makes perfect, and practice also makes permanent. If L2 learners keep repeating in a wrong way, the errors will be fossilized. Only perfect practice makes perfect. The coaches and music teachers always aim at training and developing muscle memory. In this sense, learning a foreign language, especially acquiring the correct tones in Mandarin, is the same with playing sports or musical instruments. When learners are being corrected, the link between the wrong tone production and the context is broken. If the new link is not created through correcting, the new muscle memory will not be generated. The difference between L2 learning and training in sports and music is that L2 learners will be creative when using the language. Therefore, it is especially important to train them to handle the uncertain situations by random practice.

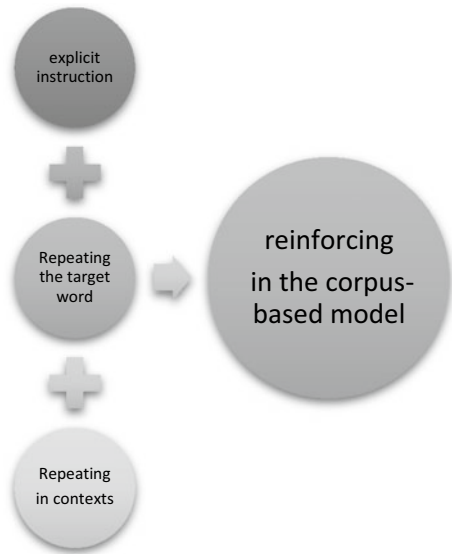
Context is shown to play an important role in correcting tone errors. The findings of the current study are in line with the previous studies that emphasized the context in tone acquisition (Yang 2016; Xu 1993). Therefore, it is important to include suprasegmentals in teaching pronunciations such as stress, rhythm, intonation, and word juncture. It is essential to repeat the target words in the original context and then apply it in a new context. Correcting the tone errors and trying it in a new context have been proved to have greater rates of consistency in retaining the accuracy among the participants in this study.

One technique to help learners to acquire these suprasegmentals is backward buildup, which is to build up long utterances from the last elements, adding one element at a time. This way, it is easy to master word juncture, rhythm, stress, and intonation. In addition, learners will not get confused by tone sandhi. For example, when a third tone, *nǚ*, is followed by another third tone *hǎo*, the first one *nǚ* is changed to the second tone *ní*. If learners are instructed to pronounce *hǎo* with the third tone and then add *ní*, the changing process from third tone to second tone can be avoided. In addition, backward buildup helps learners to learn the words within a context, not as isolated items. This will facilitate the learning process of not only tones but also grammar, word order, and meaning.

5 A Pedagogical Model

Incorporating the pedagogical strategy of random practice discussed above in L2 instruction helps to retain more of the tone accuracy than when this strategy is not used. In L2 Mandarin classes, teachers often only ask students to repeat the target words when correcting the errors. If the correction process stops here, it is very likely that the errors would occur again in different sentences. Based on the results of this study and previous ones, a pedagogical model for correcting tone errors is proposed as shown in Fig. 2. Wiener et al. (2020) found that the explicit instructions of tone facilitate to improve tone production. Hence, this pedagogical model starts

Fig. 2 A pedagogical model



with the explicit instructions. Repeating the target with the model in contexts, as the second step, is essential in correcting tone errors. It is important that these two steps happen in a classroom setting and in a timely manner, which creates L2 learners' first impressions of being corrected. The next step, reinforcing in different contexts, could happen outside classroom if there is limited time in class.

Figure 3 demonstrates a pedagogical cycle of correcting tone errors consisting of three steps: performing, learning/rehearsal, and performing. It is an upward spiral. When learners make tone errors in a performance, the teacher helps to correct the

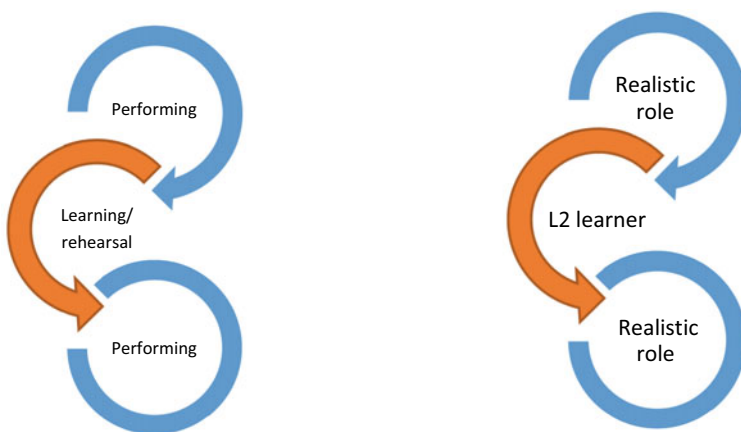


Fig. 3 Pedagogical cycle

errors, which turns the performing into learning and rehearsing mode. Without performing again, applying the correctness in a new context, the pedagogical cycle will not be completed. In this pedagogical cycle, a learner is assigned a realistic role when performing and then is pulled out and changed into L2 learner's role when being corrected. The learning is not finished until the learner goes back to the realistic role.

A corpus-based database is proposed to realize this last step. The database of sentences with audio can be constructed online, providing L2 learners a tool to search for a target word in various contexts. To save time in class, teachers can assign reinforcement tasks for L2 learners to do after class. For example, learners would be able to choose six sentences from the corpus-based database to study, then listen to the model, and repeat the whole sentence after it. The sentences in this database could be categorized based on sentence types, parts of speech, position in the sentence, sentence length, etc. The searching in the database can be modified as searching for three questions containing the target word. This corpus-based tool enables L2 learners to experience the target words in authentic contexts.

6 Limitations and Future Studies

First, the participants in this study were all beginning to intermediate level L2 learners of Mandarin Chinese, so this conclusion cannot be generalized to advanced L2 learners. Regardless, because the advanced L2 learners have mastered more vocabulary, structures, and functions than beginning to intermediate level learners, it is possible that they will need to apply the treated words in new contexts when being corrected on tone errors.

Second, this study only involved adult L2 learners, so the conclusions cannot be generalized to other groups. There could be different results if K-12 learners served as participants. Future studies could compare across the age groups, genders, the levels of proficiency, etc. In addition, the number of participants is comparatively small and they are from the same university. The researcher notes that under an ideal condition the study would have had a larger size of sample participants with more diverse learning backgrounds.

Treatment time in this study was relatively short. It would be better if the treatment is more intensive and if the post-treatment assessment is conducted at a later time, not right after the treatment. In Wiener et al.'s study (2020), an artificial language was used to avoid the influence from the preexisting knowledge. It could be used in the future study so that the researcher could have a better control of the contextual variables.

References

- Chang, C. B., & Bowles, A. R. (2015). Context effects on second-language learning of tonal contrasts. *The Journal of The Acoustical Society of America*, *138*, 3703–3716.
- Cherney, L., van Vuuren, S., Hitch, R., Hurwitz, R., & Kaye, R. (2018). Evaluating the impact of practice conditions (randomized vs. blocked) and schedule (distributed vs. massed) on script training in aphasia. *Aphasiology: International Aphasia Rehabilitation Conference (IARC) September 2018*, *32*(sup1), 45–46. <https://doi.org/10.1080/02687038.2018.1487922>.
- Corrêa, U. C., Walter, C., Torriani-Pasin, C., Barros, J., & Tani, G. (2014). Effects of the amount and schedule of varied practice after constant practice on the adaptive process of motor learning. *Motricidade*, *10*(4), 35–46. [https://doi.org/10.6063/motricidade.10\(4\).2905](https://doi.org/10.6063/motricidade.10(4).2905).
- Duff, P. A., & Li, D. (2004). Issues in Mandarin language instruction: Theory, research, and practice. *System*, *32*, 443–456.
- Elliot, C. E. (1991). The relationship between the perception and production of mandarin tones: An exploratory study. *University of Hawai'i Working Papers in ESL*, *10*(2), 177–204.
- Gebrian, M. (2016). Interleaved practice: The best practice method for reliable performance. *Journal of The American Viola Society*, *32*(2), 37–41.
- Godfroid, A., Lin, C. H., & Ryu, C. (2017). Hearing and seeing tone through color: An efficacy study of web-based, multimodal Chinese tone perception training. *Language Learning*, *67*, 819–857.
- Goldin, S. B., Horn, G. T., Schnaus, M. J., Grichanik, M., Ducey, A. J., Nofsinger, C., et al. (2014). FLS skill acquisition: A comparison of blocked vs interleaved practice. *Journal of Surgical Education*, *71*(4), 506–512. <https://doi.org/10.1016/j.jsurg.2014.01.001>.
- Gumbaridze, J. (2013). Error correction in EFL speaking classrooms. *Procedia—Social and Behavioral Sciences*, *70*, 1660–1663. <https://doi.org/10.1016/j.sbspro.2013.01.237>.
- Lee, J., Jang, J., & Plonsky, L. (2014). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*, 345–366.
- Mok, P. P., Ki, L., Albert, Li., Jingwen, J., & Xu, R. B. (2018). Orthographic effects on the perception and production of L2 mandarin tones. *Speech Communication*, *101*, 1–10. <https://doi.org/10.1016/j.specom.2018.05.002>.
- Morett, L. M., & Chang, L. Y. (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, *30*, 347–353.
- Shih, C., Lu, H. Y. D., Sun, L., Huang, J. T., & Packard, J. (2010). An adaptive training program for tone acquisition. In *Proceedings of the 5th International Conference on Speech Prosody* (paper 981). Baixas, France: International Speech Communication Association.
- Wiener, S., Chan, M. K. M., & ITO, K. . (2020). Do explicit instruction and high variability phonetic training improve nonnative speakers' mandarin tone productions? *The Modern Language Journal*, *104*(1), 152–168.
- Wiener, S., Murphy, T. K., Goel, A., Christel, M. G., & Holt, L. L. (2019). Incidental learning of nonspeech auditory analogs scaffolds second language learners' perception and production of Mandarin lexical tones. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th international congress of phonetic sciences, Melbourne, Australia 2019* (pp. 1699–1703). Canberra, Australia: Australasian Speech Science and Technology Association Inc.
- Williams, A. M., & Hodges, N. J. (2005). Practice, instruction and skill acquisition in soccer: Challenging tradition. *Journal of Sports Sciences: Preparation and Training for Soccer*, *23*(6), 637–650. T. P. Reilly and A. M. Williams (Guest Editors). <https://doi.org/10.1080/02640410400021328>.
- Xu, Y. (1993). *Contextual tonal variation in mandarin Chinese*. University of Connecticut.
- Yang, B. (2012). The gap between the perception and production of tones by American learners of Mandarin—An intralingual perspective. *Chinese as a Second Language Research*, *1*(1), 33–53. <https://doi.org/10.1515/caslar-2012-0003>.
- Yang, C. (2016). *The acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice* (Bilingual Processing and Acquisition (BPA), Volume 1).

- Zhang, H. (2018). *Second language acquisition of Mandarin Chinese tones : Beyond first-language transfer*. Brill Rodopi.
- Zheng, A., Hirata, Y., & Kelly, S. D. (2018). Exploring the effects of imitating hand gestures and head nods on L1 and L2 mandarin tone production. *Journal of Speech, Language, and Hearing Research, 61*(9), 2179–2195. https://doi.org/10.1044/2018_JSLHR-S-17-0481.

Prosody

Duration of Disyllabic Words Produced by Russian Learners of Chinese



Jing Yang and Bei Yang

Abstract The current study explores the performance of disyllabic words' duration in Chinese produced by advanced L2 learners. Eight native speakers and eight Russian learners of Chinese were recruited. They participated in two read-aloud tasks. In one task, they read aloud disyllabic words at the sentence-medial position, and in the other, they read aloud disyllabic words at the sentence-final position. Each task included thirty sentences. The duration of disyllabic words produced by L2 learners was analyzed by ANOVA. The results indicated that L2 learners adequately produced the duration of disyllabic words at the sentence-final position. The extra lengthening occurred on the second syllable within a disyllabic word at the sentence-medial position. Moreover, at the sentence-medial position, learners and native speakers had significant differences in duration of *Tone 1* and *Tone 2* of the first syllable. The findings suggest that multiple factors influence L2 syllable duration at various word/sentence positions, including the prosodic chunking ability of learners, L1 prosodic structure and physiological mechanism.

Keywords Russian learners of chinese · Disyllabic words · Duration · Production

1 Introduction

Prosody is important in identifying foreign accents in Second Language Acquisition (Mareüil and Vieru-Dimulescu 2006). It involves multiple levels above linguistic units, including syllables, words, phrases, and sentences (Kent and Read 1992). The principal prosodic features include stress, tone, and intonation (Ladefoged and Johnson 2011). Duration is also an important prosodic feature, which is an important embodiment of learners' learning results and learning ability in reading aloud (Zhou and Chen 2014). Prosodically and rhythmically, Russian is a stress-timed language (Abercrombie 1967) because stressed syllables "tend to come at more or less evenly recurrent intervals" (Pike 1945: 35), while Chinese is closer to a syllable-timed

J. Yang (✉) · B. Yang
Sun Yat-Sen University, Zhuhai, Guangdong Province, China
e-mail: yangj533@mail2.sysu.edu.cn

language (An 1997; Zhou 2008) and “all syllables seem to have the same duration in these languages” (Ladefoged and Johnson 2011: 252).

In terms of syllable duration, differences exist between “stress-timed” languages and “syllable-timed” languages (Abercrombie 1967). For instance, stressed syllables and unstressed syllables in Russian are different in duration. The unstressed syllable duration is highly weakened in Russian, but in Mandarin, the duration differences among syllables are not obvious (Zhu et al. 2001), except the light syllables. Besides, the syllable duration is affected by the tone types in Chinese, whereas Russian as a non-tone language (Wang 1982), the syllable duration is affected by stress. Because the most commonly used standard feet in Chinese consist of two syllables (Wang 2000), we investigate the duration patterns in disyllabic words in Chinese produced by Russian learners, as compared to native speakers of Chinese.

In the field of second language acquisition, more and more researchers have focused on the duration of linguistic units. Some studies found that late bilinguals produce longer second language (L2) sentences than early bilinguals, and early bilinguals tend to produce L2 sentences with longer durations than native speakers, because “late bilinguals needed to expend more resources to suppress their native language subsystem” (Mackay and Flege 2004: 373). Aoyama and Guion (2007) found that absolute durations of syllables and utterances tended to be longer in non-native speakers (NNS) utterances than in native speakers (NS). While the sentence duration produced by NNS (especially late bilinguals) is longer than that by native speakers, does the same hold true for advanced learners when they produce disyllabic words in Chinese?

There are not many studies on the prosodic aspects, especially duration, of Chinese as a second language, except tones. One of the differences between L2 speech and native speech is prosodic boundaries. Deng et al. (2005) analyzed the duration of disyllabic words in Mandarin and Taiwanese Mandarin at the sentence-medial and the sentence-final positions. The results showed that the clause boundary effect worked on both Chinese Mandarin and Taiwanese Mandarin, leading to the obvious lengthening of syllable duration at the sentence-final position. Chen (2013) found that American learners of different proficiency levels had different duration output of the syllable at the sentence boundary. As a result, there was no difference between advanced learners and NS. Liu and Chen (2016) examined the syllable duration at the prosodic boundary hierarchy by NS and Korean learners at different proficiency levels. The result showed that there was no significant difference between NS and the learners when the syllable was at the intonation phrase boundary and prosodic phrase boundary. Gao and Wang (2018) investigated the syllable duration at phrase boundaries and the Chinese prosodic chunking abilities of L2 learners. The results showed that there were an average of 3.9 syllables in a L2 chunking unit and an average 6 syllables in a NS chunking unit.

Furthermore, the syllable duration is also related to tones. Shi and Liao (1986) pointed out that there were obvious changes and differences in the syllable duration in Chinese produced by NS and American learners of Chinese. The result showed that the syllable duration in Chinese was directly related to the tone types. Deng, Shi, and Lv (2005) suggested that the syllable duration of *Tone 1* and *Tone 2* was lengthened

more obviously than *Tone 3* and *Tone 4* with regard to syllables at the sentence-final position. Zhou and Chen (2014) found that NNS was not able to acquire the duration of *Tone 4* well.

Most previous studies focused on learners whose native language is English. How about learners of Chinese whose L1 is another language, such as Russian? Do the word boundary and sentence boundary affect L2 learners in a similar way? Are they related to prosodic chunking abilities of L2 learners or L1 prosodic structure? Do tone contours affect the duration of disyllabic words in L2 speech? If so, which tones are not good at producing by L2 learners?

Therefore, the current study compares Russian learners of Chinese with native speakers of Chinese to answer following research questions:

- Is the duration of disyllabic words significantly different between NS and NNS?
- Would the duration of disyllabic words show any difference between the sentence-medial position and sentence-final position?
- How do different positions within a disyllabic word affect the syllable duration? What are the differences between NS and NNS?
- Are there any duration differences among different tones?

In order to answer these questions, the current study examines the syllable duration within a disyllabic word, words at different positions in a sentence, and analyzes how tones affect syllable duration.

2 Methods

2.1 Participants

There are two groups of participants. One group consists of eight Russian adults with a advanced level of Chinese, who have passed the *Hanyu Shuiping Kaoshi* (HSK) Level 5 and have spent more than two and a half years studying Chinese in China. These Russian learners (age mean: 22; age range: 20–25; four males and four females) do not have any Chinese language background. The other group is a control group that has eight native speakers of Chinese. The native speakers (age mean: 24.5; age range: 20–26; four males and four females) have passed the second-level in National Mandarin Test (*Putonghua Shuiping Ceshi*, PSC).

2.2 Experiment Instrument and Procedure

The disyllabic words were selected based on the fifteen Chinese tone combinations proposed by Wu (1982). Each combination included two syllables. 30 disyllabic

words were selected from *HSK1*, *Short-term Spoken Chinese Threshold*, and *Short-term Spoken Chinese Elementary*. Meanwhile, “我去____” (I go to____) and “我去____学习” (I go to ____to study) were chosen as the carrier sentences, namely disyllabic words occur at both the sentence-medial and sentence-final positions respectively. It is easy for NNS to read aloud these disyllabic words and carrier sentences. To avoid unfamiliar words that may affect pronunciation, *Pinyin* was marked for each syllable.

Two read-aloud tasks were performed. In the first task, both NS and NNS were required to read disyllabic words carried by the sentence “我去____” (I go to____). In the second task, both NS and NNS were required to read the disyllabic words carried by the sentence “我去____学习” (I go to____ to study) (see appendix for details). All readings were recorded.

The experiments were conducted in a university phonetic laboratory or a quiet room. Each speaker scrutinized the carrier sentences for 1–2 min before recording. Each sentence was required to be read twice in a most natural way. The order of sentences was randomized. CoolEdit was used to record the pronunciations of speakers. The participants' utterances were digitized at 44.1 kHz with 16-bit resolution. Totally, $15 \times 2 \times 2 \times 16 = 960$ sentences in which the target words located at the final position and $15 \times 2 \times 2 \times 16 = 960$ sentences in which the target words located at the medial position were recorded.

In order to analyze the prosodic chunking abilities of L2 learners, a follow-up perceptual evaluation task was conducted. In this task, additional five native speakers of Chinese were recruited to mark the perceived pauses in each sentence produced by NS and NNS. In total, $15 \times 2 \times 2 \times 16 = 960$ sentences in which the target words at the medial were evaluated, so that we could focus on word boundary effect yet avoid sentence/clause boundary effect.

2.3 Analyses

The duration of the whole syllables was measured by Praat (Shi and Liao 1986; Deng et al. 2005). The closure duration at the beginning of stops and affricates were included.

The Analyses of Variance (ANOVA) were conducted on data of the duration of disyllabic words via SPSS. The between-group factor was nationality (Chinese and Russian), the within-group factors were the position of a sentence (sentence-medial and sentence-final), syllable position within a disyllabic word (first and second syllable), tones (*Tone 1*, *Tone 2*, *Tone 3*, and *Tone 4*). The critical p-value for ANOVA and post hoc tests was set to $p < 0.05$.

The visualization method proposed by Wagner (2007) was used to identify the difference in speaking style-related rhythmical preferences between Chinese and Russian speakers. To compare the timing relation across different speakers (Chinese and Russian), the duration was normalized via z-score. The z-score duration of the first syllable was plotted on the horizontal axis, while that of the second syllable was

plotted on the vertical axis. As shown in Fig. 1, the two-dimensional diagram could be divided into four quadrants numbered from 1 to 4, which represent long-long, short-long, short-short, and long-short syllables respectively. Data points of stress-timed languages tend to locate in the second (iambic) and the fourth (trochaic) quadrants. Data points of syllable-timed languages tend to locate in the first (spondaic) and the third (reduced) quadrants (Wagner 2007). As shown in Fig. 2, data points of global

Fig. 1 Rhythm related timing relations in a two-dimensional diagram

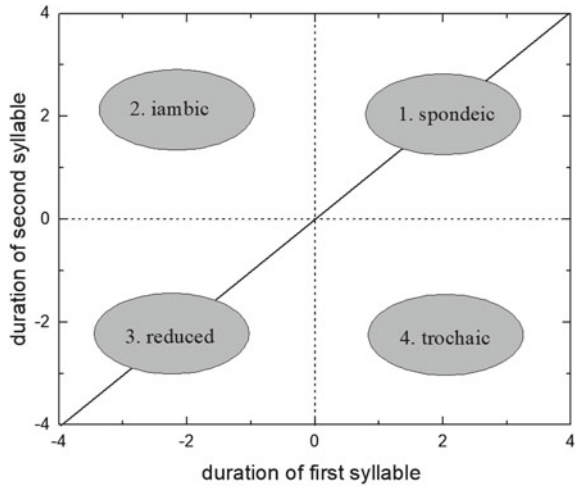
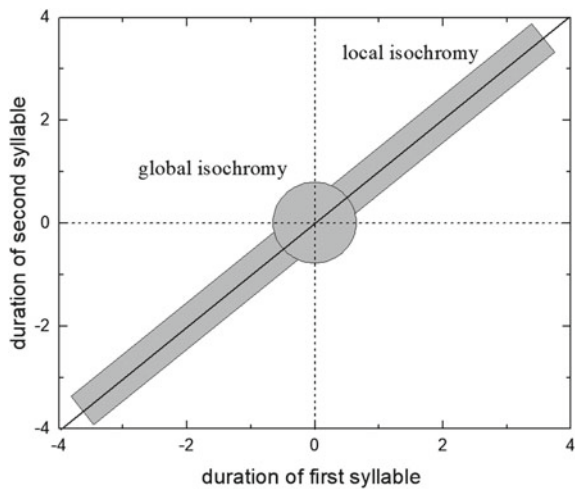


Fig. 2 Distribution of global isochrony and local isochrony



isochrony¹ mainly locate in the circular region around (0, 0), while data points of local isochrony² are distributed around the bisecting line, i.e. $y = x$.

3 Results

3.1 Duration of Disyllabic Words Between NS and NNS

Four-factor composite analysis of variance was used to analyze the data (details in 2.3). Statistical results showed that there was no significant difference in the duration of the disyllabic words between native speakers and L2 learners. The overall result indicated that advanced L2 learners had a good grasp of disyllabic words duration. However, all the within-group factors were significant: for the position of a sentence (medial/final), $F(1,14) = 60.15$, $p < 0.001$; for the position of the syllable within a disyllabic word (first/second syllable), $F(1,14) = 64.547$, $p < 0.001$; for tones (*Tone 1*, *Tone 2*, *Tone 3*, *Tone 4*), $F(3,42) = 13.898$, $p < 0.001$.

3.2 Duration Affected by Different Positions in a Sentence

The significant effect of the position in a sentence was further explored with Bonferoni post hoc tests. The analyses showed that NS and NNS had different performance when they read aloud the disyllabic words at the sentence-medial position ($p = 0.039$), yet no significant effect at the sentence-final position. To more clearly show the similarities and differences between NS and L2 learners at the sentence-medial and the sentence-final positions, the two-dimensional diagrams of visualization were drawn in Fig. 3.

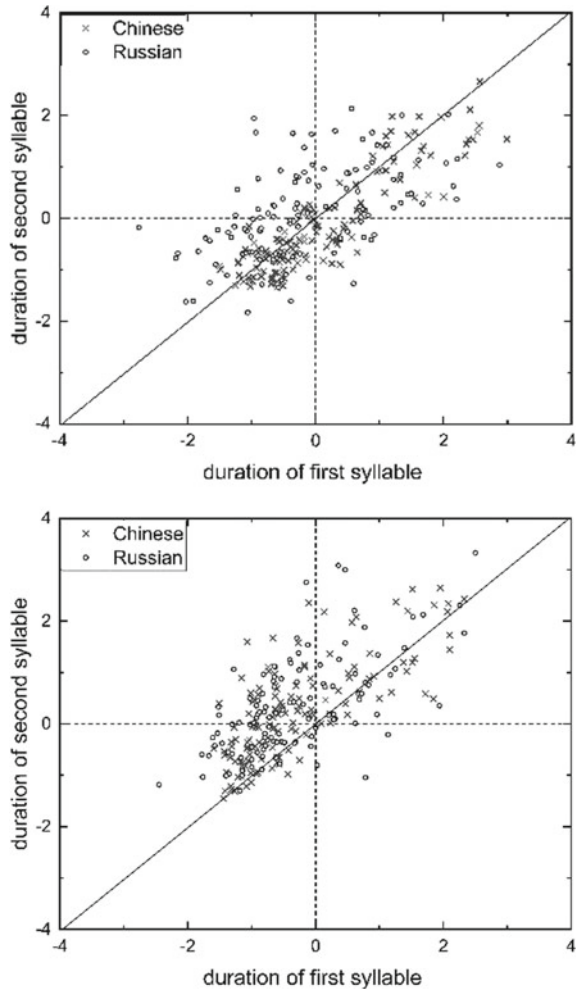
3.2.1 Disyllabic Words at the Sentence-Medial Position

On the left panel of Fig. 3, the data points (the red color or the symbol “×”) of Chinese production are more concentrated on the diagonal. Compared with NS, the distribution of Russian learners’ data points (the blue color or the symbol “o”) is more dispersed. Both red (“×”) and blue (“o”) points are mixed in quadrant 1 and 3, and some points (both red “×” and blue “o”) are mixed in quadrant 4 near the point (0, 0). The data combined in these three quadrants mentioned above indicated that the duration of most syllables produced by NNS was similar to that by NS. In other

¹Global isochrony means that the duration of the first syllable is equal to that of the second syllable within a word, and the durations of different words are similar.

²Local isochrony means that the duration of the first syllable is equal to that of the second syllable within a word.

Fig. 3 Visualizations of timing relations of the two syllables within disyllabic words in Chinese. Note: The left Fig. presents the relation of the two syllables within disyllabic words by NS and NNS at the sentence-medial position and the right Fig. presents that at the sentence-final position

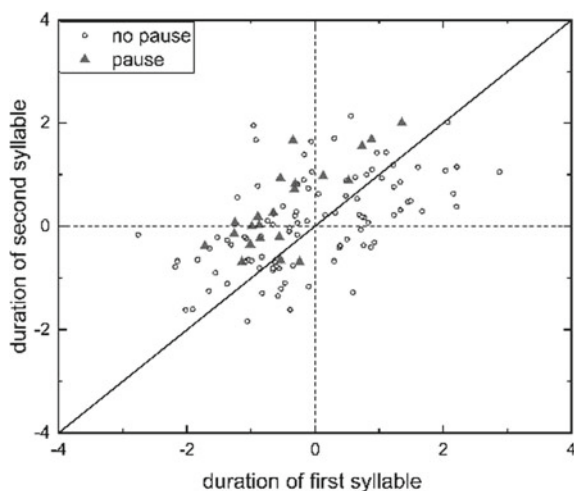


words, the duration of the first syllable was similar to that of the second syllable, which reflected the feature of the syllable-timed language, Chinese.

However, only blue (“o”) points occur in quadrant 2, and some blue points locate far above the diagonal in quadrants 1 and 3; but there are not any red (“x”) points that occur in these areas. This showed that the duration of some disyllabic words produced by NNS was different from that by NS, i.e. the duration of the second syllable was longer than that of the first syllable. These data reflected the feature of the stress-timed language. Meanwhile, it revealed that NNS preferred elongating the second syllable.

The results of the follow-up perceptual evaluation task showed that the perceived pauses only annotated in NNS sentences. Among the NNS sentences, only four sentences were annotated as a “2 + 2+2” pattern, which meant a pause were perceived

Fig. 4 Visualizations of timing relations of two syllables within disyllabic words by NNS at sentence-medial positions



in every two syllables, e.g. [我去][罗马][学习], 23 sentences were annotated as a “2 + 4” pattern, e.g. [我去][罗马学习], and 58 sentences were viewed as a “4 + 2” pattern, e.g. [我去罗马][学习]. To explore why some data points of NNS appear in the second quadrant and observe the word boundary effect, we divided the whole materials of L2 learners into two parts: one is the words without perceived pauses, the other is the words with perceived pauses at the boundary of the target word (Fig. 4).

As can be seen from Fig. 4, the circular points represent the disyllabic words without a pause at the word boundary, while the triangular points represent those with a pause. The triangular and the circular points are mixed above the diagonal, whereas some circular points are below the diagonal, only two triangular points below yet near it. Compared with triangular points, the distribution of circular points is more dispersed. These indicated that the duration of some disyllabic words with a pause was different from those without a pause. In other words, almost all words with a pause (triangular points) presented the “short-long” feature, whereas words without a pause (circular points) showed features of “short-long”, “short-short”, and “long-long”. Further, as mentioned above, compared with data points of Chinese (“x”), only data points of NNS (“o”) occur in quadrant 2 (the left panel of Fig. 3). From Fig. 4, we can see that data points of both disyllabic words without a pause and disyllabic words with a pause at the word boundary appear on the second quadrant. Further analysis shown in discussion.

3.2.2 Disyllabic Words at the Sentence-Final Position

As shown on the right panel of Fig. 3, the data points (red “x” and blue “o”) from Chinese and Russian are mixed and more points are above the diagonal and appear in the second quadrant (“short-long”) compared to the left panel of Fig. 3. It indicated

that the duration of the second syllable by NS and NNS was prolonged, possibly influenced by the sentence boundary effect (Yang 1997; Wang, Yang, and Lu 2004) when learners have not acquired the phonological structure completely. Meanwhile, different from the second quadrant on the left panel, red and blue points are mixed in the second quadrant on the right panel of Fig. 3, which showed a short-long duration pattern for both NS and NNS at the sentence-final position. ANOVA results also revealed that the differences between NS and NNS were reduced when disyllabic words at the sentence-final position ($p = 0.098 > 0.05$). So how does the sentence boundary specifically affect the duration of disyllabic words?

Repeated Measures ANOVA was used for analyzing the duration of disyllabic words produced by NS and NNS separately. The within-group factors were the position of a sentence (medial/end), tones (*Tone 1* to *Tone 4*), the syllable position within a disyllabic word (first/second syllable). The results showed that the duration of the disyllabic words had significant differences between the two locations, i.e., sentence-medial and sentence-final positions by NS ($F(1, 7) = 39.95, p < 0.001$). The average duration differed by 50 ms. There was a significant difference between the two locations by L2 learners ($F(1, 7) = 22.176, p = 0.002$). The average duration differed by 39 ms. The rate of prolongation is defined as:

$$C_{\text{prolongation}} = \frac{T_{\text{end}} - T_{\text{medial}}}{T_{\text{medial}}} \times 100\% = \frac{\Delta T}{T_{\text{medial}}} \times 100\% \tag{1}$$

where T_{end} is the duration of disyllabic words at the sentence-final position, and T_{medial} is the duration of disyllabic words at the sentence-medial position. Table 1 showed the details.

Table 1 shows that the mean duration of L2 learners is longer than that of NS, and the rate of prolongation by NS is higher than that by L2 speakers, which indicated the sentence boundary effect on duration applied to the speech by both NS and L2 learners. This would cause a nonsignificant effect at the sentence-final position between NS and NNS.

Compared with the sentence-final position, the sentence-medial position could better represent the different production between NS and NNS. Based on this, we conducted a further analysis.

Table 1 Comparison of NS and NNS duration of disyllabic words at different positions in a sentence

Nationality	Mean duration (ms, medial)	Mean duration (ms, final)	Prolongation (%)
Chinese (NS)	286 ± 19	335 ± 16	17 ± 8.7
Russian (NNS)	346 ± 18	385 ± 23	11 ± 8.3

3.3 *Duration Affected by Different Positions Within a Disyllabic Word According to the Positions in a Sentence*

The significant effect of position within a disyllabic word was further explored with Bonferroni post hoc tests. The result identified that the duration of the first syllable at the sentence-medial position was not significantly different between NS and NNS, while the duration of the second syllable at the sentence-medial produced by NS was significantly shorter than NNS ($p < 0.015$). The duration of the first and the second syllables at the sentence-final position was not significantly different between NS and NNS. In order to explore the boundary effect on L2 learners in detail, the duration of the first and the second syllables were investigated in NNS and NS respectively.

The result showed that there was no significant difference in duration between two syllables within a disyllabic word at the sentence-medial position by NS and NNS, while there was a significant difference at the sentence-final position (NS: $p < 0.001$, $T_{\text{first}} - T_{\text{second}} = -48$ ms; NNS: $p < 0.001$, $T_{\text{first}} - T_{\text{second}} = -57$ ms). The duration of the second syllable was longer than the first one. It indicated that for disyllabic words, the second syllable was affected by the sentence boundary obviously than the word boundary. The syllable duration at the sentence-final position is prolonged since it conveys the end-of-discourse information (Cao 2005). Both NS and L2 learners were influenced by this effect.

3.4 *Duration Affected by Tones According to the Positions in a Sentence*

The post hoc tests revealed that at the sentence-medial position, there were significant differences on the duration of *Tone 1*, *Tone 2*, and *Tone 4* between NS and L2 speakers ($p_{\text{Tone1}} = 0.015$; $p_{\text{Tone2}} = 0.044$; $p_{\text{Tone4}} = 0.035$). More specifically, there were significant differences on the duration of *Tone 1* and *Tone 2* of the first syllable between NS and NNS ($p_{\text{Tone1}} = 0.005$, $p_{\text{Tone2}} = 0.005$), and there were significant differences on the duration of *Tone 1*, *Tone 3* and *Tone 4* of the second syllable between NS and NNS ($p_{\text{Tone1}} = 0.005$, $p_{\text{Tone3}} = 0.015$, $p_{\text{Tone4}} = 0.016$). However, there was no significant difference on the duration of *Tone 2* of the second syllable between NS and NNS, the p-value was 0.051, which was very close to the critical p-value, 0.05. This meant that almost all the durations of the second syllables by Russians were significantly different from that of NS. In addition, the duration from *Tone 1* to *Tone 4* produced by L2 learners were all longer than NS.

At the sentence-final position, there was a significant difference on the duration of *Tone 1* between NS and L2 speakers ($p_{\text{Tone1}} = 0.038$), and only the duration of *Tone 1* of the first syllable showed a significant difference between NS and NNS ($p_{\text{Tone1}} = 0.033$). The average duration of *Tone 1* produced by NS was 70 ms, which was significantly less than that by Russians.

3.5 Summary

The overall statistical analysis could not reflect the differences in the duration of disyllabic words between NS and NNS, and there were significant differences for within-group factors and the interactions including group*position and group*tone type.

The duration of disyllabic words of NNS was significantly longer than that of NS at the sentence-medial position. In detail, the duration of the second syllable by NNS was significantly longer than NS.

When the target word at the sentence-final position was affected by the sentence boundary, the duration of disyllabic words was prolonged. The rate of prolongation for NS was higher than that for NNS, which led to a reduction in the difference between NS and NNS. Hence, the duration of the first/second syllable showed non-significant difference between NS and NNS.

At the sentence-medial position, NS and NNS had significant differences in duration of *Tone 1* and *Tone 2* of the first syllable, and almost all the durations of second syllables were prolonged by NNS. According to the tone types, except for *Tone 1* of the first syllable, the results did not show any significant differences between NS and NNS when target words at the sentence-final position.

4 Discussion

4.1 Differences at the Sentence-Medial Position

There was a significant difference between NS and NNS when the target words were at the sentence-medial position. More specifically, the duration of the second syllable by NNS was significantly longer than NS. Why?

On one hand, it is related to the Chinese prosodic chunking abilities of NNS. According to the results of the follow-up perceptual evaluation task, L2 learners can produce sentences fluently without a pause. The carrier sentence, as a L2 chunking unit, carries 6 syllables. This contrasts with the study by Gao and Wang (2018), which suggested an average of 3.9 syllables in a L2 chunking unit, and an average 6 syllables in a NS chunking unit. Gao and Wang (2018) conducted the memorization task and each sentence included 12–17 syllables, so L2 learners might need more time to process and organize the utterance. However, a few L2 learners probably further divide the sentence into smaller chunks, such as phonological phrases, like [我去罗马][学习] or [我去][罗马][学习]. In this case, the extra lengthening of the target word's second syllable of L2 productions would reflect simple phrase-final lengthening. And “the silent pause could cause time delay, so it is sometimes perceived as dysfluency when it occurs in speech production.” (Yang 2017: 21). This reveals that the prosodic chunking abilities of NNS could influence L2 prosodic phrasing, so that

the salient feature at the surface level is the lengthening of second syllable within a disyllabic word at the sentence-medial position.

On the other hand, L1 prosodic feature might also influence the L2 production. The duration of disyllabic words is influenced less by the word boundary when they are at the sentence-medial position, so prosodic words lengthening in Chinese is weak (Qian et al. 2001; Wang et al. 2004), which reveals the prosodic feature of Chinese as a syllable-timed language more obviously (see Fig. 3). Russian is a stress-timed language, but its default stress position has been extensively argued in the literature, and the stress position of Russian is flexible. Jouravlev and Lupker (2014) explored the distribution of metrical patterns in Russian disyllabic words based on a Russian word corpus. They found that there is no clear regular pattern overall. Further, experiments on the production of disyllabic words indicated that, for verbs and nouns, there is no advantage for initial or final stress. Molczanow et al. (2019) investigated stress errors produced by speakers diagnosed with acquired surface dyslexia, and found initially stressed words presented more errors than finally stressed words, thus suggesting the final stress as the metrical default in Russian. This view supported the theoretical model proposed by Alderete (1999) who pointed out the metrical default of Russian is post-stem prominence. Besides, the experimental results of Molczanow et al. (2019) are in line with the results of some previous studies on the production of nonce words and unfamiliar borrowings, such as Mayer (1976), Crosswhite et al. (2003), etc. However, Molczanow et al. (2013) suggested a trochee to be the metrical default in Russian by employing event-related potentials (ERPs). This result supports the proposals of Melvold (1989), Idsardi (1992), Halle (1997) and Revithiadou (1999). The current study is based on production experiments. It seems that the iambic pattern at the sentence-medial position in L2 is transferred from L1. From these perspectives, the current study supports the final stress as the metrical default in Russian.

4.2 Non-significant Differences at the Sentence-Final Position

At the sentence-final position, there was no significant difference regarding the duration of disyllabic words between NS and NNS.

Firstly, due to the effect of sentence boundary on duration, the rate of prolongation by NS was higher than that by L2 speakers, which reduced the differences between NS and NNS (details in 3.2.2). Meanwhile, the sentence-final position could be more prominent than the medial position (Zhou and Chen 2014), so Russians pay more attention to pronunciation when the target words at the final position, i.e., the lengthening associated with prominences preceding intonational boundaries (Price et al. 1992). Thirdly, the NS data (see 3.3) showed that the duration of the second syllable was longer than the first when the disyllabic words at the sentence-final position. Therefore, the disyllabic word showed a “short-long” prosody feature, which

was similar to the prosodic feature in Russian (the final stress as the metrical default). Hence, it is relatively easy for L2 learners to produce disyllabic words at the sentence-final position. Therefore, disyllabic words at the sentence-final position are produced well by advanced learners. This result is consistent with Chen (2013), Liu and Chen (2016), Gao and Wang (2018).

4.3 *Tone Effect*

The data showed that there were significant differences between NS and NNS in the duration of *Tone 1* and *Tone 2* of the first syllable when the target words at the sentence-medial position. And there was a significant difference between NS and NNS in the duration of *Tone 1* of the first syllable when the target words at the sentence-final position. The duration of almost all the second syllable showed a significant difference between NS and NNS when target words at the sentence-medial position, whereas there was no significant difference at the sentence-final position.

From the perspective of tone types and physiological mechanism of pronunciation, *Tone 1* and *Tone 2* are leveling and ascending, and their terminal characteristics are both “high”. However, *Tone 3* and *Tone 4* are falling, and the final features are both “low”. The limitation of physiology mechanism makes it impossible to maintain the characteristics of “low” for a long time, so the duration extensions of falling tones (*Tone 3* and *Tone 4*) are less than that of ascending and leveling tone (*Tone 1* and *Tone 2*) (Deng et al. 2006). Therefore, in terms of the physiology mechanism of pronunciation, the duration of *tone 1* and *tone 2* of the first syllable is easier to prolong than *Tone 3* and *Tone 4* for NNS. Besides, it also might be affected by metrical patterns of nouns in Russian. As mentioned above, there is not any advantage for initial or final stress for nouns in Russian (Jouravlev and Lupker 2014). Although the current study preferring the final stress as the metrical default in Russian, it is undeniable that there are still some or even more disyllabic words carrying the initial stress. The duration of stressed syllables is longer than non-stressed syllables (Wang 1982; Zhu et al. 2001). For L2 learners, the initial stress of nouns’ metrical pattern could affect L2 learners’ production, so the duration of the first syllable within a disyllabic word is lengthened.

Further, compared with the duration of different tones carried by the first syllable, the duration of almost all the second syllable at the sentence-medial position showed significant differences between NS and NNS, which indicated that the word boundary and the metrical default of Russian affect the duration of the second syllable prominently than tones. It is similar to the situations when target words at the sentence-final position, the sentence boundary effect plays a more important role than tone types in the duration of the second syllable.

From discussions above, the production of disyllabic words by NNS are related to their Chinese prosodic chunking abilities, physiological mechanism of pronunciation and prosodic characteristics of Russian. Therefore, there are some pedagogical implications in the current study. The duration of a disyllabic word at the sentence-medial position can be used for evaluating the acquisition of NNS compared with NS. Meanwhile, L2 learners should be reminded not to pause too long during speaking when there is not a big break, such as a intonational phrase boundary or a sentence boundary. More attention should be paid to *Tone 1* and *Tone 2* while teaching, in order to avoid extra lengthening of *Tone 1* and *Tone 2*. Additionally, it would be better to introduce different prosody features in Chinese and Russian. The prosodic feature, duration, reflects the features of some related abilities and L2 development, including L2 chunking abilities which influence pauses, L1 transfer and physiological mechanism.

4.4 Future Research Directions

The current study, which includes some controlled experiments, did not involve rhythmic perception experiments, and the participants only included advanced L2 learners. Therefore, we do not know whether rhythmic perception and production on duration in the connected speech will lead to the same results. Meanwhile, we do not know whether the performance is different among L2 learners at different Chinese proficiency levels or not. Moreover, the disyllabic words at the sentence-initial position were not observed in the current design. All issues mentioned above are worth further examining.

5 Conclusion

The findings of the current study contribute to the understanding of the different performance of disyllabic words between NS and NNS in terms of the duration. Advanced learners produce the duration of the disyllabic words well overall. However, the duration of the second syllable within a disyllabic word at the sentence-medial position, and the duration of *Tone 1* and *Tone 2* of the first syllable by NNS are significantly longer than that by NS. Besides, disyllabic words at the sentence-final position are produced well by advanced learners, which is similar to that by NS.

Appendix

1. The Disyllabic Words at the Sentence-final Position

1. 我去罗马(luó mǎ)。
2. 我去南方(nán fāng)。
3. 我去美国(měi guó)。
4. 我去广州(guǎng zhōu)。
5. 我去客厅(kè tīng)。
6. 我去中国(zhōng guó)。
7. 我去食堂(shí táng)。
8. 我去韩国(hán guó)。
9. 我去日本(rì běn)。
10. 我去公司(gōng sī)。
11. 我去长沙(cháng shā)。
12. 我去首都(shǒu dū)。
13. 我去学校(xué xiào)。
14. 我去桂林(guì lín)。
15. 我去上海(shàng hǎi)。
16. 我去大连(dà lián)。
17. 我去宾馆(bīn guǎn)。
18. 我去故宫(gù gōng)。
19. 我去抚顺(fǔ shùn)。
20. 我去法国(fǎ guó)。
21. 我去宿舍(sù shè)。
22. 我去新疆(xīn jiāng)。
23. 我去门口(mén kǒu)。
24. 我去教室(jiào shì)。
25. 我去公园(gōng yuán)。
26. 我去武汉(wǔ hàn)。

27. 我去重庆(chóng qìng)。
28. 我去青岛(qīng dǎo)。
29. 我去书店(shū diàn)。
30. 我去单位(dān wèi)。

2. The Disyllabic Words at the Sentence-medial Position

1. 我去罗马(luó mǎ)学习。
2. 我去南方(nán fāng)学习。
3. 我去美国(měi guó)学习。
4. 我去广州(guǎng zhōu)学习。
5. 我去客厅(kè tīng)学习。
6. 我去中国(zhōng guó)学习。
7. 我去食堂(shí táng)学习。
8. 我去韩国(hán guó)学习。
9. 我去日本(rì běn)学习。
10. 我去公司(gōng sī)学习。
11. 我去长沙(cháng shā)学习。
12. 我去首都(shǒu dū)学习。
13. 我去学校(xué xiào)学习。
14. 我去桂林(guì lín)学习。
15. 我去上海(shàng hǎi)学习。
16. 我去大连(dà lián)学习。
17. 我去宾馆(bīn guǎn)学习。
18. 我去故宫(gù gōng)学习。
19. 我去抚顺(fǔ shùn)学习。
20. 我去法国(fǎ guó)学习。
21. 我去宿舍(sù shè)学习。
22. 我去新疆(xīn jiāng)学习。
23. 我去门口(mén kǒu)学习。
24. 我去教室(jiào shì)学习。

25. 我去公园(gōng yuán)学习。
26. 我去武汉(wǔ hàn)学习。
27. 我去重庆(chóng qìng)学习。
28. 我去青岛(qīng dǎo)学习。
29. 我去书店(shū diàn)学习。
30. 我去单位(dān wèi)学习。


References

- Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh University Press.
- Alderete, J. D. (1999). *Morphologically governed accent in optimality theory* [Doctoral dissertation, University of Massachusetts]. <https://doi.org/10.7282/T3J965B0>.
- An, Y. (1997). 等时等长的汉语节奏原则. *Chinese Language Learning*, (5), 26–29.
- Aoyama, K., & Guion, S. G. (2007). Prosody in second language acquisition: Acoustic analyses of duration and F0 range. In O. Bohn & M. J. Munro (Eds.), *The role of language experience in second-language speech learning* (pp. 281–297). In Honor of James Emil Flege: John Benjamins Publishing Co.
- Cao, J. (2005). 音节时长伸缩与话语韵律结构. In *the 8th National Conference on Man-Machine Speech Communication*. Technical Acoustics, 354–359.
- Chen, M. (2013). The Characteristics of Oral Chinese Prosodic Boundaries of American English Speaking CSL Learners. *Chinese Teaching in the World*, 27(1), 97–106.
- Crosswhite, K., Alderete J., Beasley T., & Markman V. (2003). Morphological effects on default stress in novel Russian words. In G. Garding & M. Tsujimura (Eds.), *Proceedings of the 22nd West Coast Conference on Formal Linguistics*. Cascadilla Press, 151–164. <http://roa.rutgers.edu/files/630-1103/630-CROSSWHITE-0-0.PDF>.
- Deng, D., Shi, F., & Lv, S. (2005). 国语在台湾双音节词的韵律表现. In *the 8th National Conference on Man-Machine Speech Communication*. Technical Acoustics, 350–353.
- Deng, D., Shi, F., & Lv, S. (2006). 普通话双音节韵律词时长特性研究. In *the 13th Phonetic Conference of China*, 111–117.
- Gao, S., & Wang, J. (2018). Characteristics of prosodic grouping in the speech production by learners of Chinese from South Korea. *Journal of Yunnan Normal University (Teaching & Studying Chinese as a Foreign Language Edition)*, 16(1), 17–27. <http://doi.org/10.3969/j.issn.1672-1306.2018.01.004>.
- Halle, M. (1997). On stress and accent in Indo-European. *Language*, 73(2), 275–313. <https://doi.org/10.2307/416020>.
- Idsardi, W. J. (1992). *The computation of prosody* [Doctoral dissertation, Massachusetts Institute of Technology]. <http://dspace.mit.edu/handle/1721.1/12897>.
- Jouravlev, O., & Lupker, S. J. (2014). Stress consistency and stress regularity effects in Russian. *Language, Cognition and Neuroscience*, 29(5), 605–619. <https://doi.org/10.1080/01690965.2013.813562>.
- Kent, R. D., & Read, C. (1992). *The acoustic analysis of speech*. Singular Publishing Group. <http://doi.org/10.13140/RG.2.1.2449.0404>.
- Ladefoged, P., & Johnson, K. (2011). *A course in phonetics* (6th ed). Cengage Learning.
- Liu, F. & Chen, M. (2016). A Study of Chinese Prosodic Boundary Characteristics of Chinese as Second Language Learners. *TCSOL Studies*, (3), 1–16. <http://doi.org/10.3969/j.issn.1674-8174.2016.03.001>.

- Mackay, I. R. A., & Flege, J. E. (2004). Effects of the age of second language learning on the duration of first and second language sentences: The role of suppression. *Applied Psycholinguistics*, 25(3), 373–396. <https://doi.org/10.1017/S0142716404001171>.
- Mareüil, P. B., & Vieru-Dimulescu, B. (2006). The contribution of prosody to the perception of foreign accent. *Phonetica*, 63(4), 247–267. <https://doi.org/10.1159/000097308>.
- Mayer, G. L. (1976). The stress of foreign place names in Russian. *Slavic and East European Journal*, 20(4), 451–459.
- Melvold, J. L. (1989). Structure and stress in the phonology of Russian. *Massachusetts Institute of Technology*, 447(1), 124–8.
- Molczanow, J., Iskra, E., Dragoy, O., Wiese, R., & Domahs, U. (2019). Default stress assignment in Russian: Evidence from acquired surface dyslexia. *Phonology*, 36(1), 61–90. <https://doi.org/10.1017/S0952675719000046>.
- Molczanow, J., Domahs, U., Knaus, J., & Wiese, R. (2013). The lexical representation of word stress in Russian: Evidence from event-related potentials. *The Mental Lexicon*, 8(2), 164–194. <https://doi.org/10.1075/ml.8.2.03mol>.
- Pike, K. L. (1945). *The intonation of American english*. University of Michigan Press.
- Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1992). The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America*, 90(6), 2956–2970. <https://doi.org/10.1121/1.401770>.
- Qian, Y., Chu, M., & Pan, W. (2001). 普通话韵律单元边界的声学分析. In *the 5th National Conference on Modern Phonetics*. Tsinghua University Press, 70–74.
- Revithiadou, A. (1999). *Headmost accent wins: head dominance and ideal prosodic form in lexical accent systems* [Doctoral dissertation, Rijksuniversiteit te Leiden]. <https://doi.org/10.7282/T3M53R31>.
- Shi, F., & Liao, R. (1986). 中美学生汉语塞音时值对比分析. *Language Teaching and Linguistic Studies*, (4), 67–83.
- Wagner, P. (2007). Visualizing levels of rhythmic organization. In Trouvain, J. & Barry, W. J. (Eds.), *XVIth international congress of the phonetic sciences*. Pirrot GmbH, pp. 1113–1116. <http://pub.uni-bielefeld.de/publication/1785380>.
- Wang, B., Yang, Y., & Lv, S. (2004). Acoustic analysis on prosodic hierarchical boundaries of Chinese. *Acta Acustica*, 29(1), 29–36.
- Wang, H. (2000). The prosodic word and prosodic phrase of Chinese. *Studies of the Chinese Language*, (6), 525–536.
- Wang, X. (1982). 论俄汉语音, 语调, 重音—节律的差异 (下). *Foreign Language Research*, (4), 55–61.
- Wu, Z. (1982). 普通话语句中的声调变化. *Studies of the Chinese Language*, (6), 439–450.
- Yang, B. (2017). Prosodic features, self-monitoring, and dysfluency in native and non-native mandarin speech. *Chinese as a Second Language*, 52(1), 3–27. <https://doi.org/10.1075/csl.52.1.01yan>.
- Yang, Y. (1997). Prosodic cues to syntactic boundaries. *Acta Acustica*, 22(5), 414–421.
- Zhou, Q., & Chen, M. (2014). Syllable Duration of Oral Reading Prosody in Chinese as Second Language. In *the 11th Phonetic Conference of China*, 203–213.
- Zhou, S. P. (2008). The Acquisition of Chinese Rhythm by American Students. *Journal of College of Chinese Language and Culture of Jinan University*, (2), 38–44. <http://doi.org/10.3969/j.issn.1674-8174.2008.02.005>.
- Zhu, G., Xu, L., Jiang, Y., & Pan, L. (2001). *汉俄语音对比实验研究*. Nanjing University Press.

How Does Mandarin Learning Experience Modulate Second-Language Learners' Phonological Knowledge of Tone 3 Sandhi in Word Production?



Zhen Qin 

Abstract Tone 3 (T3) in Mandarin is one of the most difficult tones for second-language (L2) learners given its variants in different contexts. While previous studies investigated L2 learners' acquisition of T3 in the sandhi context, it remains unclear how L2 learners' learning experience modulates their phonological knowledge of T3 sandhi in producing different types of words. This study used a wug production test to investigate the effect of Mandarin learning experience on the production of the sandhi form (a rising tone) by experienced and inexperienced Korean-speaking learners. The acoustic analyses showed that experienced Korean-speaking L2 learners were better at using their phonological knowledge of T3 sandhi than less experienced learners when producing pseudo and novel words, but not real Chinese words. The experienced learners had higher pitch values and a steeper rising slope for the T3 sandhi form of pseudo words, and higher pitch values for that of novel words, than the inexperienced learners. The findings suggested that learners' increased experience with Mandarin facilitated their use of the phonological knowledge of T3 sandhi. Given learners' difficulties with the T3 sandhi rule, language teachers are suggested to develop teaching materials with different word types to promote the generalization of the rule.

Keywords Tone 3 sandhi · Phonological knowledge · WUG test · L2 Korean learners

1 Introduction

Mandarin Chinese (henceforth, Mandarin) is a language which is rapidly gaining in importance on the international scene. Accordingly, it has become commonly taught in foreign language programs outside of Mandarin-speaking areas (e.g., USA) according to the National Council of Less Commonly Taught Languages (NCOLCTL), and an increasingly large number of students are learning Mandarin.

Z. Qin (✉)

Division of Humanities, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
e-mail: hmzqin@ust.hk

© Springer Nature Singapore Pte Ltd. 2021

C. Yang (ed.), *The Acquisition of Chinese as a Second Language Pronunciation*, Prosody, Phonology and Phonetics, https://doi.org/10.1007/978-981-15-3809-4_9

Importantly, it differs from many other languages in the types of information they use to convey meaning in words: Mandarin uses lexical tones (i.e., pitch movement) to contrast word meanings (e.g., /pā/ “eight” (Tone 1 (T1)), /pá/ “to pull out” (T2), /pǎ/ “to hold” (T3), /pà/ “father” (T4) (Yip 2006)). Learning tones pose difficulties for adult second-language (L2) learners of Mandarin who speak non-tonal languages as their mother tongue (Pelzl 2019; Wiener, Ito, and Speer 2018). Among the four Mandarin tones, T3 is one of the most challenging tones for English-speaking L2 learners (e.g., Hao 2012; Yang 2019). One possible reason is that T3 is involved in a tone sandhi rule with allophonic variations and thus poses great difficulty in learning its different variants depending on the tone contexts (Yip 2006).

Specifically, different from other Mandarin tones, Mandarin T3 is involved in a phonological alternation as follows: T3 (214; the dipping tone) in Mandarin becomes T2 (35; the rising tone) before another T3. This phonological alternation of tones is called the T3 sandhi rule. It can be written as follows:

214 (T3, the citation form) → 35 (T2, the sandhi form) / ____ 214 (T3, the citation form).

The T3 sandhi rule must apply in Mandarin disyllabic words, for instance, *yǔ sǎn* “umbrella” (Yip 2006). The L2 acquisition of T3 in the sandhi and non-sandhi contexts has been researched in several studies (e.g., Zhang 2016, 2017; Zhang and Xie 2020). For instance, Zhang (2017) tested the production of different allophonic variants of T3 as follows: (a) the citation form (214) of T3 in the final position; (b) the half-form (21) of T3 when preceding T1, T2, and T4; and (c) the sandhi form (35) of T3 when preceding another T3. By examining the production of Mandarin disyllabic words by English-speaking learners with different learning experience, the study showed that the learning experience of Mandarin modulated the learners’ use of the T3 half-form. Specifically, inexperienced learner groups (beginner, intermediate level) mispronounced a T3 half-form as a citation form of T3, when it preceded T1, T2, and T4, more frequently than the experienced learner group (advance-level). Pronunciation errors were also found for the T3 sandhi form when it preceded another T3 across learner groups. While the findings of the above-mentioned studies suggested an effect of Mandarin learning experience on L2 learners’ acquisition of T3 in different forms, the error-based approach (on real Chinese words) of the studies did not allow us to build a deeper understanding of L2 learners’ phonological knowledge of T3 sandhi (i.e., the correct application of the T3 sandhi rule). Therefore, acoustic studies using a wug production test are needed to examine the effect of learning experience on L2 learners’ use of their underlying phonological knowledge of T3 sandhi in producing different types of Chinese words.

The wug test, which has been widely adopted to test the productivity of morpho/phonological alternations in non-Chinese languages (Hayes and Londe 2006; Hsieh 1976; Zuraw 2007), is also good at testing Mandarin speakers’ phonological knowledge of T3 sandhi (Zhang and Lai 2010). In a typical wug test, experimental participants are taught novel forms (pseudo words or novel words) in their native language (L1) and then asked to provide morphologically complex forms (e.g., Tagalog infixation; see Zuraw, 2007 for details), using the novel forms as the base. A better application of the morpho/phonological alternations in the novel forms

indicates a better phonological knowledge of the alternations. The T3 sandhi rule in Mandarin has been shown to be highly productive in novel forms for native Mandarin-speaking adult participants (Zhang and Lai 2010; Zhang, Xia, and Peng 2015) and child participants (Huang, Zhang, and Zhang 2018; Huang, Zuo, and Zhang 2019). For instance, Zhang and Lai (2010) used a wug test to examine adult native speakers' phonological knowledge of Mandarin tone sandhi patterns. The study aimed to test the synchronic relevance of phonetics by investigating native Mandarin speakers' applications of the T3 sandhi process from *real* disyllabic words, consisting of actual occurring (AO) morphemes, to two types of wug words: *pseudo* disyllabic words consisting of non-occurring sequences consisting of real morphemes, and *novel* words consisting of non-occurring sequences of non-occurring syllables of accidental gaps (AG). The results from Zhang and Lai (2010) showed that native speakers' production of pseudo and novel T3 sandhi words shared a greater similarity of pitch shape with the citation form (214) of T3 in having a lower and later turning point (i.e., a less rising slope) than did their production of real sandhi words (the sandhi form, 25). In pseudo and novel words, the T3 sandhi form, with a less rising slope, was produced more like an underlying T3 than it was in real words. This finding indicates an acoustically "incomplete application" of the T3 sandhi rule when generalizing the phonological knowledge to pseudo and novel words in adults. Zhang and Lai (2010) linked the "incomplete application" of the T3 sandhi rule by native speakers to a weaker phonetic motivation for this type of sandhi pattern in nature.

In a similar fashion, the wug test has also been used to test L2 learners' underlying phonological knowledge of T3 sandhi. For instance, Chen et al. (2019) adopted a similar paradigm to examine the ability to produce Mandarin T3 sandhi by two groups of L2 learners, that is, tonal (Cantonese) speakers and non-tonal (English) speakers. The functional data analysis of normalized pitch values revealed that compared with native speakers of Mandarin, L2 learners showed less accurate production of the T3 sandhi form (25) with lower pitch values and less rising slope, which was attributed to L2 learners' less (acoustically) detailed phonological representations of allophonic variants. However, Cantonese- and English-speaking L2 learners applied the T3 sandhi rule similarly for both real words and non-real (pseudo and novel) words, suggesting that their phonological knowledge of T3 sandhi was equally learned. While the learning experience of Mandarin was balanced between the English-speaking and Cantonese-speaking L2 learner groups, the L2 learners' learning experience was not manipulated. It remains unclear whether, and if so, how Mandarin learning experience modulates L2 learners' use of the phonological knowledge of T3 sandhi in producing real, pseudo, and novel words.

To fill this research gap regarding the effect of Mandarin learning experience and to complement findings of previous L2 studies (Zhang 2017) using an error-based approach, the current study adopted the wug test paradigm to investigate the effect of Mandarin learning experience on L2 learners' phonological knowledge of T3 sandhi. Specifically, experienced versus inexperienced (non-tonal) Korean-speaking adult L2 learners, who were relatively understudied in the previous literature, were compared in their productions of different types of Chinese words (real, pseudo,

and novel words). The productions were then acoustically analyzed and statistically modeled to examine their phonological knowledge of T3 sandhi and uncover the underlying mechanism involved.

2 Methods

2.1 Participants

Sixteen native Korean speakers (mean age: 24.6, SD: 4.3, nine females and seven males) who learned Mandarin as the L2 participated in this study. All the L2 learners reported that (Seoul) Korean was their L1, and that both their parents were native Korean speakers. They were all college students studying in Shanghai, China. Additionally, they reported having learned Mandarin after the age of 12 and not having been exposed to tone languages other than Mandarin. They reported normal hearing and no history of speech or language disorders. In compensation for their time, the participants each received the equivalent of ten US dollars.

Crucially, to test the effect of Mandarin learning experience on the phonological knowledge of T3 sandhi, eight L2 learners who had passed HSK 5 were recruited as experienced learners. Another eight L2 learners who had not passed HSK 5 were recruited as inexperienced learners. The L2 learners' Mandarin learning experience is summarized in Table 1. As can be seen from their biographical information, both the experienced and inexperienced learners started learning Mandarin at a similar age, that is, around 20 years old. However, the experienced learner group received Mandarin instruction for a longer time in the classroom and spent a longer time immersed in Mandarin-speaking areas than the inexperienced learner group.

Table 1 Biographical information of experienced and inexperienced Korean-speaking L2 learners of Mandarin

	AOE (year)	Years of Mandarin Instruction	LOR (month)
Experienced Learner Group (<i>n</i> = 8)	20 (3.4)	3.4 (2.1)	39.0 (24.4)
Inexperienced Learner Group (<i>n</i> = 8)	21 (3.6)	1.6 (1.0)	8.1 (5.9)

Mean (standard deviation), *AOE* age of first exposure to Mandarin, *LOR* length of residence in Mandarin-speaking areas

2.2 Materials

Following the experimental design of Zhang and Lai (2010) and Chen et al. (2019), a Chinese wug test was conducted with the Korean-speaking L2 learners of Mandarin. The items were disyllabic words consisting of either actual occurring (AO) morphemes such as *měi* with 美 as corresponding Chinese character or accidental gaps (AG) such as *hěi* with no corresponding Chinese character. To test the phonological knowledge of T3 sandhi, three sets of disyllabic test words which carry T3 in the first and second syllables, as listed in Table 2, were constructed as follows: (1) real words: real Chinese disyllabic words (AO-AO); (2) pseudo words: non-occurring sequences consisting of actual occurring morphemes (*AO-AO); (3) novel words: non-occurring sequences of non-occurring syllables of accidental gaps (AG-AG).

All the chosen real words and individual character/morpheme for pseudo and novel words were selected from the textbook used for the Mandarin class at Year 1. Thus, the L2 learners in this study were supposed to be familiar with the test items. In order not to reveal the purpose of the experiment, a total of 60 disyllabic filler words were used. The filler words included real words, pseudo words, and novel words in a similar design with the other 15 tonal combinations (T3 + T1; T3 + T2; T3 + T4; T1-T1; T1-T2; T1-T3; T1-T4; T2-T1; T2-T2; T2-T3; T2-T4; T4-T1; T4-T2; T4-T3; T4-T4). In total, there were 12 test words (4 items * 3 word types) and 60 filler words (4 items * 15 tone combinations) with all possible tonal combinations in the test. Before the experiment, each monosyllable used for the test words and filler words was recorded with three repetitions by a female native speaker of Beijing Mandarin. Recordings were conducted in a soundproof room using a microphone linked to a digital recorder. One token for each target monosyllable was chosen from three repetitions by the investigator based on its intelligibility and sound quality.

Table 2 Test words used in the condition of real words, pseudo words, and novel words

Word type	Chinese characters	Chinese pinyin
Real words (AO + AO)	美好 手表 整理 可以	měihǎo shǒubiǎo zhěnglǐ kěyǐ
Pseudo words (*AO + AO)	美朵 手怎 整早 可散	měiduǒ shǒuzěn zhěngzǎo kěsǎn
Novel words (AG + AG)	NA	hěidiǔ cǒusén sēnduǐ tēcǒu

2.3 Procedures

The task was conducted using the paradigm software (Perception Research Systems, Inc. <https://www.paradigmexperiments.com/>). In each trial, two monosyllables were presented in an auditory mode to the participants with 800 ms in between. Each monosyllable was also presented visually with their characters (if available) and phonetic symbols (pinyin) along with sounds. The participants were instructed to put the two monosyllables they heard together to verbally produce a disyllabic word in Mandarin. They were instructed to speak at a normal speaking rate and could self-correct when necessary. The experiment started with a demonstration session with the investigator demonstrating how to put the two monosyllables together by verbally producing a disyllabic word in Mandarin. The demonstration session was then followed by a practice session involving ten new practice trials, in which the participants were allowed to practice the task before the experimental session. In the experimental session, the order of trial presentation was randomized across participants within one block.

2.4 Measurements

Participants' verbal productions were acoustically coded in Praat (Boersma and Weenink 2018). The fundamental frequency (F0) was extracted at ten equidistant points within each annotated vowel using the ProsodyPro Praat script (Xu 2013). The extracted F0 values were then converted from Hz to semitones with a reference of 50 Hz. The F0 values of each token were z-score normalized against the mean pitch across all tokens for each individual speaker using the following formula: Normalized pitch = Observed pitch – Mean pitch / Standard deviation of pitch (mean pitch and standard deviation of pitch are the grand mean and standard deviations of all tokens per individual participant).

2.5 Data Analysis

Four tokens (one token from the experienced learner group, three tokens from the inexperienced learner group) for disyllabic test words were excluded from the analysis. Two tokens were produced as two isolated monosyllables (syllable interval longer than 300 ms) and the other two tokens were produced with errors (T3 mispronounced as T1 or T4). A total of 188 tokens for disyllabic test words were included in the data analysis (95 tokens from the experienced learner group; 93 tokens from the inexperienced learner group).

The dependent variable for the statistical analyses was normalized pitch values in semitone. The growth curve analysis (GCA) has the intercept for average pitch values, and it also uses the *poly* function to generate two other parameters for pitch shape, the

first-order linear polynomial, and the second-order quadratic polynomial. The two polynomials enable us to model participants' normalized pitch shape (curves) over time. According to Mirman (2014), the *intercept* captures the average pitch value with the higher the intercept, the higher the average pitch value; the *linear* polynomial captures the pitch slope with a positive (t) value indicating a rising pitch and a larger value indicating more steepness and vice versa; the *quadratic* polynomial indicates a single-inflection curve of pitch shape with a positive (t) value indicating a concave shape and a larger value indicating a more concave pitch shape and vice versa.

The GCAs were conducted with the *lme4* package in R (Bates et al. 2015). The analyses included the two polynomials (linear and quadratic) modeling pitch shape, condition (real words, pseudo words, and novel words), and group (experienced learners; inexperienced learners) as fixed effects. The effect of condition was dummy coded with real words as baseline, whereas group was contrast-coded (i.e., -0.5 and 0.5). A back-fitting function from the package *LMERConvenienceFunctions* in R (Tremblay and Ransijn 2015) was used to identify the model that accounted for significantly more of the variance than simpler models, as determined by log-likelihood ratio tests; only the results of the model with the best fit are presented, with p values being calculated using the *lmerTest* package in R (Kuznetsova, Brockhoff, and Christensen 2018). All analyses included participant as random intercept and the orthogonal polynomials as random slopes for the participant variable, which allowed the analysis to model a line of a different shape for each individual participant. A larger analysis that tested three-way interaction between the effects of polynomials, condition, and group was conducted to determine whether the two L2 groups differ in their tone production across the three word conditions and justify the main GCAs conducted separately to test the interactions between the effects of polynomials (also, intercept) and group in each condition. The GCAs were conducted separately for each condition with the alpha level being adjusted to 0.017 for each of the three models.

To conclude that the Mandarin learning experience influenced L2 learners' phonological knowledge of T3 sandhi (Chen et al. 2017), the GCA in each word condition must reveal either a main effect of group (interpreted on the intercept) or interactions between group and at least one polynomial (the linear or quadratic polynomial). Specially, a main effect of group indicates the participants' average pitch values are different between the two L2 groups, whereas the interaction between the effects of group and polynomials indicates that the shape (pitch slope indexed by the linear polynomial or concave shape indexed by the quadratic polynomial) of participants' pitch shape is different between the two L2 groups. Given the nature of the wug test, it is predicted that the two L2 groups will show differences in their production of pseudo and novel words, but not necessarily their production of real words which might not tap into L2 learners' underlying phonological knowledge given learners' familiarity with these words.

3 Results

Figure 1 shows normalized pitch values (semitones) of the first syllable in T3 sandhi disyllabic words produced by experienced and inexperienced Korean-speaking learners for real words, pseudo words, and novel words. As illustrated in Fig. 1, the rising slope carried by Syllable 1 indicates that both experienced and inexperienced Korean-speaking L2 learners had correctly applied the T3 sandhi rule to

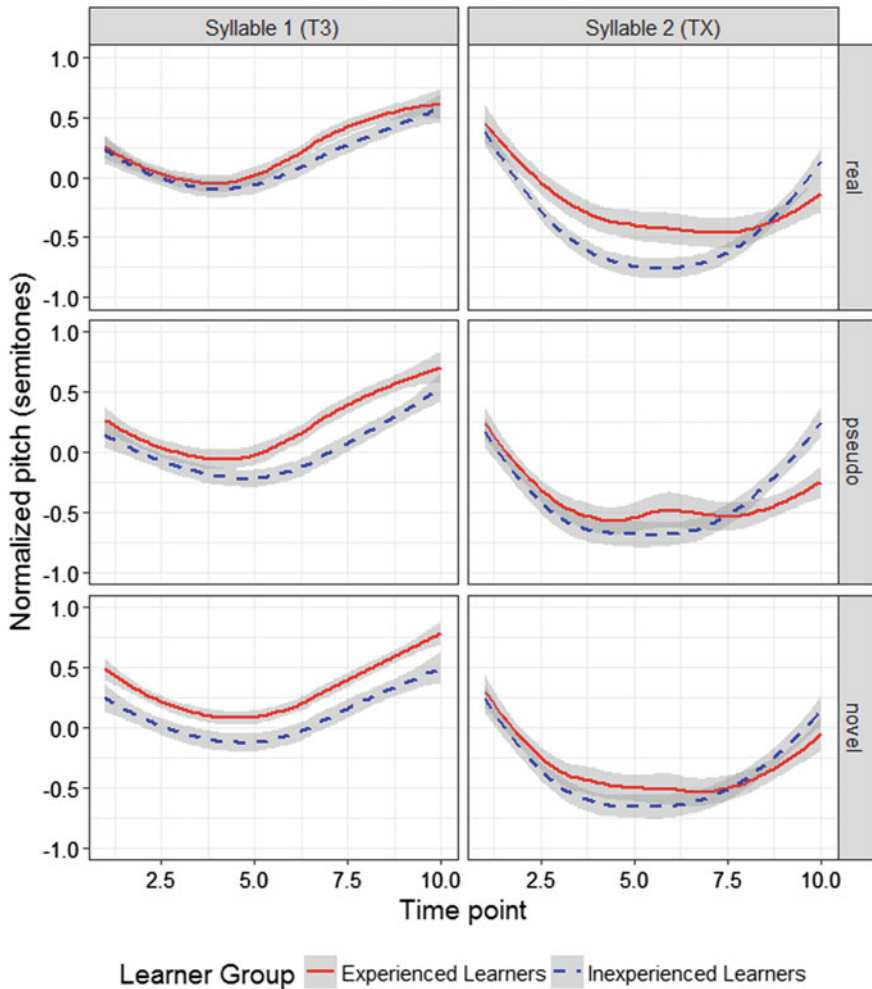


Fig. 1 Normalized pitch values in semitones of Syllable 1 in T3 sandhi words produced by experienced Korean-speaking learners (*red, solid*) and inexperienced learners (*blue, dashed*) in the condition of real words (*top*), pseudo words (*middle*), and novel words (*bottom*); the shaded area represents one standard error above and below the participant mean

real, pseudo, and novel disyllabic words. That is, they changed the tone of the first syllable (T3) to a rising tone (T2) in the T3 sandhi context. However, a visual inspection of Fig. 1 suggests that while the experienced and inexperienced learner groups had a similar pitch shape of Syllable 1 for real words, they had different acoustic realizations of Syllable 1 for pseudo words and novel words.

To determine whether the two L2 groups differ in their tone production across the three word conditions and justify the main analysis of three GCAs conducted in each condition as illustrated in Fig. 1, a larger GCA was performed on the normalized pitch values (semitone) of Syllable 1 in T3 sandhi disyllabic words with the effect of condition (real words, pseudo words, and novel words; baseline: real words), group (experienced learners; inexperienced learners; contrast-coded), and two polynomials (linear and quadratic; baseline: linear) modeling pitch shape as fixed factors. The results of this GCA with the best fit are presented in detail in Table 4 of the Appendix. The GCA with the best fit on the normalized pitch values included the linear and quadratic polynomials, condition, group, and the interactions between the linear polynomial and other factors. Importantly, there was a three-way interaction between the linear polynomial, condition, and group (for the pseudo words condition). The main GCAs were therefore performed on the effects of the linear polynomial and group separately for real words, pseudo words, and novel words, as illustrated in Fig. 1.

Table 3 presents the results of the main GCAs which included an effect of linear polynomial, group (contrast-coded), and their interaction in each condition. Of the results of the GCA on real words in Table 3, the significant positive t value for the linear polynomial indicates that L2 learners' normalized pitch for the real words had a rising pitch slope. There was no significant effect of group (experienced: 0.19 semitone; inexperienced: 0.16 semitone). Crucially, there was no significant interaction effect between the linear polynomial and group. The results of the model on real words suggest that the experienced and inexperienced learner groups had their T3 sandhi form of Syllable 1 produced as a rising tone with no difference in average pitch values and pitch slope.

For the results of the GCA on pseudo words, the linear polynomial was not significant. However, the significant positive t value for the group effect (experienced: 0.29 semitone; inexperienced: 0.05 semitone) indicates that the T3 sandhi form of the experienced learner group had higher pitch values than that of the inexperienced learner group for pseudo words. Crucially, the significant positive t value for the interaction between the linear polynomial and group indicates that the T3 sandhi form of the experienced learner group had a steeper rising slope than that of the inexperienced learner group for pseudo words. The results of the model on pseudo words suggest that the experienced learner group had their T3 sandhi form of Syllable 1 produced with higher pitch values and a steeper rising slope than the inexperienced learner group.

For the results of the GCA on novel words, the linear polynomial was not significant. However, the significant positive t value for the group effect (experienced: 0.29 semitone; inexperienced: 0.14 semitone) indicates that the T3 sandhi form of the experienced learner group had higher pitch values than that of the inexperienced

Table 3 Growth curve analyses on normalized pitch values (semitone) of syllable 1 in the T3 sandhi context for real words, pseudo words, and novel words

Condition	Effect	Estimate	<i>t</i>	<i>p</i>
Real words	(Intercept)	0.409	6.736	< 0.001
	Polynomial			
	Linear	0.380	3.211	< 0.01
	Group	0.010	0.226	0.821
	Polynomial × Group			
	Linear	0.322	2.280	0.023
Pseudo words	(Intercept)	0.276	4.485	< 0.001
	Polynomial			
	Linear	– 0.177	– 1.945	0.069
	Group	0.117	3.921	< 0.001
	Polynomial × Group			
	Linear	0.272	2.880	< 0.01
Novel words	(Intercept)	0.339	4.759	< 0.001
	Polynomial			
	Linear	– 0.088	– 0.874	0.389
	Group	0.111	2.612	< 0.01
	Polynomial × Group			
	Linear	– 0.222	– 1.650	0.099

$\alpha = 0.017$, significant results are in bold, real words: $n = 640$ observations, pseudo words: $n = 620$ observations, novel words: $n = 620$ observations

learner group for novel words. There was no significant interaction effect between the linear polynomial and group. The results of the model on novel words suggest that the experienced learner group had their T3 sandhi form of Syllable 1 produced with higher pitch values, without difference in pitch slope, than the inexperienced learner group.

To summarize the experienced and inexperienced groups of Korean-speaking L2 learners, both correctly pronounced a rising tone for Syllable 1 of the real words; however, their phonological knowledge of T3 sandhi in pseudo words and novel words differed. Specifically, the experienced learner group had higher pitch values and a steeper rising slope for pseudo words, and higher pitch values for novel words, than the learner group with less learning experience.

4 Discussion

The present study examined the effect of Mandarin learning experience on Korean-speaking L2 learners' phonological knowledge of T3 sandhi when producing real, pseudo, and novel Chinese words. The results of the acoustic analysis indicated that the Korean-speaking L2 learners with more learning experience were better at using

the phonological knowledge of T3 sandhi than less experienced learners in producing pseudo and novel words, but not real words. In the text below, we first discuss the effect of Mandarin learning experience on the L2 learners' better use of phonological knowledge of tones and then turn to the discussion of L2 pedagogical implications which could be applied to the Mandarin language classroom.

First, the finding regarding the Mandarin learning experience is consistent with previous studies on the L2 acquisition of Mandarin tones in isolation (Hao 2018; Sun 2012) and in tonal contexts (He and Wayland 2010; Yang 2011). While T3 was often mispronounced as T2 by L2 learners as the two tones are neutralized in the T3 sandhi context (Hao 2012; Yang 2011), an increased learning experience could help L2 learners reduce pronunciation errors of the T3 half-form (21) in the non-sandhi context (Zhang 2017). Complementing the existing studies of L2 tone acquisition, our findings supported with acoustic evidence further suggested that an increased learning experience also facilitated the L2 learners' use of their phonological knowledge in the sandhi context, with the experienced learners producing the T3 sandhi form (35) of pseudo and novel words more accurately than the inexperienced learners. Compared with inexperienced L2 learners of Mandarin, experienced L2 learners might have been more exposed to natural tonal variants in the Mandarin-speaking environment and thus should have more robust representations of lexical tones (Qin, Tremblay, and Zhang 2019).

Second, the results of T3 sandhi production across different word types found for the experienced and inexperienced L2 learner groups are also in line with developmental studies testing children's phonological knowledge of T3 sandhi (Huang, Zhang, and Zhang 2018; Huang, Zuo, and Zhang 2019). With an increased exposure to natural tonal variants of T3 and a stronger phonological/morphological awareness, a developmental trajectory emerged with older children's production of the T3 sandhi form becoming more adult-like than younger children's production. In a similar vein, a learning trajectory was found in this study with the experienced learners' production of the T3 sandhi form, specifically for pseudo and novel words, becoming more acoustically accurate than the inexperienced learners' production.

The effect of Mandarin learning experience revealed for pseudo and novel words, but not for real words, is presumably attributed to the nature of Mandarin T3 sandhi. As Zhang and Lai (2010) found, adult native Mandarin speakers differently pronounced the T3 sandhi form between real words and non-real words (including both pseudo and novel words), with the T3 sandhi form for non-real words having a less rising slope than that for real words. Thus, even native speakers may experience some difficulty using their phonological knowledge of T3 sandhi, which is not as phonetically motivated as other T3 variants (i.e., T3 half-form) (Zhang and Lai 2010). It is not surprising that, like adult native Mandarin speakers, L2 learners also had

difficulty applying the rule when producing the pseudo words and novel words with less experienced L2 learners having greater difficulty than their more experienced counterparts. Another possibility is that L2 learners' lexical knowledge, such as their familiarity with spoken forms of real words, may have assisted them in applying the T3 sandhi rule successfully for real words instead of other words regardless of their Mandarin learning experience. To tease apart the effect of tone sandhi nature from L2 learners' word familiarity, future studies may consider testing the (experienced and inexperienced) L2 learners' phonological knowledge of different types of tone sandhi (less phonetically motivated tone sandhi such as T3 sandhi variant vs. more phonetically motivated tone sandhi/coarticulation such as T3 half-form variant) using both real and non-real words (see Chen et al. 2017 for an example).

Interestingly, the experienced learner group had higher pitch values and a steeper rising slope than the inexperienced learner group for pseudo words, suggesting a better use of their phonological knowledge of T3 sandhi. However, they showed higher pitch values than their inexperienced counterparts only for novel words. One plausible explanation for the different effects of Mandarin learning experience between pseudo words and novel words is that both experienced and inexperienced learners shared a difficulty applying the T3 sandhi rule in novel words (not in pseudo words) given the novelty of the items. On the other hand, the experienced learners' increased exposure to Mandarin tonal input and larger vocabulary size might have resulted in a greater explicit awareness of novel words, which was presumably indexed by higher pitch values in their production of the T3 sandhi form (see Huang et al. 2019 for similar findings). To corroborate the plausibility, further studies are required to recruit native speakers as reference to investigate whether the T3 production of experienced learners is more acoustically native-like in terms of average pitch values and a rising slope than that of inexperienced learners (Chen et al. 2019).

This research not only deepens our understanding of the mechanism underlying L2 learners' production of the T3 sandhi form but also provides pedagogical implications for Mandarin language teachers. The results of our current research showed that inexperienced Korean-speaking L2 learners of Mandarin had greater difficulty than their experienced counterparts in generalizing their phonological knowledge of T3 sandhi from production of real words to that of pseudo and novel words. This difficulty for inexperienced L2 learners was possibly attributed to a less robust representation of T3, which has different tone variants. As Zhang (2017) pointed out, the citation form (214) of T3 is often taught to L2 learners first and thus might be treated as a default form of T3 by L2 learners. In contrast, both the sandhi form (35) and the T3 half-form (21) are not introduced in detail in the classroom setting. Accordingly, quite a few L2 Mandarin teaching practitioners also assume that the citation form (214) of T3 is the primary form and thus treats other variants of T3 as "unnatural" (Sun 1997). As a result of classroom instruction, L2 learners who have limited exposure to tonal variants will not have an explicit awareness of the

differences between T3 variants, and they will be less likely to apply the T3 sandhi rule to contexts other than familiar contexts (e.g., real words). Mandarin teaching practitioners are thus suggested to treat T3 as a special case given its difficulty for L2 learners and then integrate detailed pronunciations of its sandhi form and other variants in the L2 teaching curriculum (see the chapter by Jiang Liu).

To help L2 learners overcome their difficulties using their phonological knowledge of T3 sandhi, Mandarin language teachers are specifically encouraged to develop L2 teaching materials that focus on the application of the T3 sandhi rule, preferably using not only real words but also pseudo/novel words (Zhang 2017, 2018). One approach to deal with L2 learners' difficulty would be to have them complete intensive training in a laboratory setting, in which learners are required to learn to pay more attention to the differences between the T3 sandhi form and other variants in their production as well as perception of real and non-real (pseudo and novel) words (Li, Yang, and Chen 2018). Furthermore, to assist L2 learners in building a robust representation of Mandarin tones, acoustically variable tonal stimuli produced by different speakers (female and male), in different tonal contexts (sandhi and non-sandhi contexts), and in different types of words (real and non-real words), can be used in such a training paradigm (see the chapter by Yingjie Li). This high-variability training paradigm would initially improve L2 learners' ability in distinguishing tonal variants (the citation form vs. the sandhi form) in different contexts (Chang and Bowles 2015; Liu and Zhang 2016; Wang et al. 1999; Wang, Jongman, and Sereno 2003). And it would result in a more robust representation of tonal categories in the long term and ultimately a more efficient use of Mandarin tones for L2 learners.

To the best of our knowledge, the present study is one of the first to examine the effect of Mandarin learning experience on L2 learners' phonological knowledge of T3 sandhi in word production. The findings suggest that experienced Korean-speaking L2 learners were better in using their phonological knowledge of T3 sandhi than less experienced learners in producing pseudo and novel words, but not real words. These findings shed light on L2 learners' underlying mechanism of producing tones in the sandhi context and provide pedagogical implications for Mandarin teaching in the classroom setting. More importantly, the study sparks interest in questions regarding the different types of tone sandhi and the native likeness of the tone sandhi production for further research.

Acknowledgements We thank Dr. Haifeng Qi at the Shanghai International Studies University for her involvement in the earlier stages of the project and her assistance in data collection. We also thank the editor and the reviewer for their insightful comments on this research.

Appendix

(See Table 4).

Table 4 Growth curve analyses on normalized pitch values (semitone) of Syllable 1 in the T3 sandhi context

Effect	Estimate	<i>t</i>	<i>p</i>
(Intercept)	0.176	3.223	< 0.01
Polynomial			
Linear	0.750	6.026	< 0.001
Quadratic	0.626	8.664	< 0.001
Group	0.010	0.246	0.806
Condition (Pseudo)	– 0.046	– 0.256	0.798
Condition (Novel)	0.034	1.574	0.116
Polynomial × Group			
Linear	0.322	2.482	0.013
Polynomial × Condition (Pseudo)			
Linear	– 0.224	– 2.416	0.016
Polynomial × Condition (Novel)			
Linear	– 0.266	– 2.868	0.004
Group × Condition (Pseudo)	0.112	1.913	0.051
Group × Condition (Novel)	0.089	1.522	0.128
Polynomial × Group × Condition (Pseudo)			
Linear	– 0.548	– 2.961	0.003
Polynomial × Group × Condition (Novel)			
Linear	– 0.041	– 0.221	0.825

$\alpha = 0.05$, significant results are in bold, $n = 1880$ observations

References

- Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: linear mixed-effects models using Eigen and Eigen. *Journal of Statistical Software*, 67(1), 1–48. <http://dx.doi.org/https://doi.org/10.18637/jss.v067.i01>
- Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer [Computer program]. Version 6.0.43*. Retrieved 8 September 2018.
- Chang, C. B., & Bowles, A. R. (2015). Context effects on second-language learning of tonal contrasts. *Journal of the Acoustical Society of America*, 138(6), 3703–3716. <https://doi.org/10.1121/1.4937612>.
- Chen, S., He, Y., Wayland, R., Yang, Y., Li, B., & Yuen, C. W. (2019). Mechanisms of tone sandhi rule application by tonal and non-tonal non-native speakers. *Speech Communication*, 115, 67–77. <https://doi.org/10.1016/j.specom.2019.10.008>.
- Chen, S., He, Y., Yuen, C. W., Li, B., & Yang, Y. (2017). Mechanisms of tone sandhi rule application by non-native speakers. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Stockholm, Sweden*, 1760–1764. <https://doi.org/https://doi.org/10.21437/Interspeech.2017-143>
- Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, 40(2), 269–279. <https://doi.org/10.1016/j.wocn.2011.11.001>.

- Hao, Y. C. (2018). Second Language Perception of Mandarin Vowels and Tones. *Language and Speech*, 61(1), 135–152. <https://doi.org/10.1177/0023830917717759>.
- Hayes, B., & Londe, Z. C. (2006). Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology*, 23(1), 59–104. <https://doi.org/10.1017/S0952675706000765>.
- He, Y., & Wayland, R. (2010). The Production of Mandarin Coarticulated Tones by Inexperienced and Experienced English Speakers of Mandarin. *Proceedings of the 5th International Conference on Speech Prosody, Chicago, United States*, 1–4.
- Hsieh, H. I. (1976). On the unreality of some phonological rules. *Lingua*, 38(1), 1–19. [https://doi.org/10.1016/0024-3841\(76\)90038-3](https://doi.org/10.1016/0024-3841(76)90038-3).
- Huang, X., Zhang, G., & Zhang, C. (2018). A Preliminary Study on the Productivity of Mandarin T3 Sandhi in Mandarin-speaking Children. *Proceedings of the 6th International Symposium on Tonal Aspects of Languages, Berlin, Germany*, 88–92.
- Huang, X., Zuo, Y., & Zhang, C. (2019). Seven-year-olds reach an adult-like productivity in the application of Mandarin tone sandhi. *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia.*, 3125–3129.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2018). lmerTest Package: Tests in Linear Mixed Effects Models. R package version 3.0–1. *Journal of Statistical Software*. <https://doi.org/10.18637/jss.v082.i13>
- Li, B., Yang, Y., & Chen, S. (2018). Perceptual Evaluation of Mandarin Tone Sandhi Production by Cantonese Speakers before and after Perceptual Training. *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, Hong Kong*, 358–366. <https://www.aclweb.org/anthology/Y18-1041>
- Liu, J., & Zhang, J. (2016). The effects of talker variability and variances on incidental learning of lexical tones of Asian Languages and Literatures, University of Minnesota, USA 2. Incidental Learning — Video Game. *Proceedings of the 5th International Symposium on Tonal Aspects of Languages, Buffalo, United States*, 23–27.
- Mirman, D. (2014). *Growth curve analysis and visualization using R*. CRC Press.
- Pelzl, E. (2019). What makes second language perception of Mandarin tones hard? *Chinese as a Second Language*, 54(1), 51–78. <https://doi.org/10.1075/csl.18009.pel>.
- Qin, Z., Tremblay, A., & Zhang, J. (2019). Influence of within-category tonal information in the recognition of Mandarin-Chinese words by native and non-native listeners: An eye-tracking study. *Journal of Phonetics*, 73, 144–157. <https://doi.org/10.1016/j.wocn.2019.01.002>.
- Sun, K.-C. (2012). *The role of lexical tone in L2 Mandarin spoken word recognition*. Unpublished Ph.D. thesis. Department of Linguistics, the University at Buffalo, State University of New York.
- Sun, S. (1997). *The development of a lexical tone phonology in American adult learners of standard Mandarin Chinese*. University of Hawaii at Manoa.
- Tremblay, A., & Ransijn, J. (2015). lmerConvenienceFunctions: A suite of functions to back-fit fixed effects and forward-fit random effects, as well as other miscellaneous functions. R package version 2.1. *Comprehensive R Archive Network*.
- Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustical Society of America*, 113(2), 1033–1043. <https://doi.org/10.1121/1.1531176>.
- Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America*, 106(6), 3649–3658. <https://doi.org/10.1121/1.428217>.
- Wiener, S., Ito, K., & Speer, S. R. (2018). Early L2 Spoken Word Recognition Combines Input-Based and Knowledge-Based Processing. *Language and Speech*, 61(4), 632–656. <https://doi.org/10.1177/0023830918761762>.
- Xu, Y. (2013). ProsodyPro - A tool for large-scale systematic prosody analysis. *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, 7–10.
- Yang, C. (2011). *The acquisition of Mandarin prosody by American learners of Chinese as a foreign language (CFL)*. Unpublished Ph.D. Thesis. Graduate Program in East Asian Languages and Literatures, Ohio State University.

- Yang, C. (2019). The effect of L1 tonal status on the acquisition of L2 Mandarin tones. *International Journal of Applied Linguistics (United Kingdom)*, 29(1), 3–16. <https://doi.org/10.1111/ijal.12223>.
- Yip, M. (2006). Tone: Phonology. *Encyclopedia of Language & Linguistics*. <https://doi.org/10.1016/b0-08-044854-2/00054-7>.
- Zhang, C., Xia, Q., & Peng, G. (2015). Mandarin third tone sandhi requires more effortful phonological encoding in speech production: Evidence from an ERP study. *Journal of Neurolinguistics*, 33, 149–162. <https://doi.org/10.1016/j.jneuroling.2014.07.002>.
- Zhang, H. (2016). Dissimilation in the second language acquisition of Mandarin Chinese tones. *Second Language Research*, 32(3), 427–451. <https://doi.org/10.1177/0267658316644293>.
- Zhang, H. (2017). The effect of theoretical assumptions on pedagogical methods: A case study of second language Chinese tones. *International Journal of Applied Linguistics (United Kingdom)*, 27(2), 363–382. <https://doi.org/10.1111/ijal.12132>.
- Zhang, H. (2018). Current trends in research of Chinese sound acquisition. In *The Routledge handbook of Chinese second language acquisition* (pp. 217–233). Taylor and Francis. <https://doi.org/10.4324/9781315670706>.
- Zhang, H., & Xie, Y. (2020). Coarticulation effects of contour tones in second language Chinese. *Chinese as a Second Language Research*, 9(1), 1–30. <https://doi.org/10.1515/caslar-2020-0001>.
- Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, 27(1), 153–201. <https://doi.org/10.1017/S0952675710000060>.
- Zuraw, K. (2007). The role of phonetic knowledge in phonological patterning: Corpus and survey evidence from tagalog infixation. *Language*, 83(2), 277–316. <https://doi.org/10.1353/lan.2007.0105>.

Intelligibility, Comprehensibility, Accentedness, and Fluency

The Effect of Fundamental Frequency on Mandarin Intelligibility by L2 Learners in Quiet and Noise Environments: A Pilot Study



Kaidi Chen and Chunsheng Yang

Abstract Fundamental frequency (F0), listening environment, and semantic context are three important factors for both tonal and non-tonal language intelligibility by native speakers. However, it remains unclear how these factors affect second language (L2) learners of Mandarin Chinese and whether there are differences between native and L2 Mandarin speakers. Through speech re-synthesis and sentence counterbalancing, this study investigated the possible effects of F0 (i.e., natural F0 versus flattened F0) on the intelligibility of Mandarin speech by L2 Mandarin learners from different proficiency levels in quiet and white noise conditions when controlling for sentence context. A mixed-effect statistical model confirmed the main effects of F0 contour, listening environment, and proficiency level. That is to say, the lack of natural F0 contour, the presence of noise, and the lower proficiency level would predict the reduction in intelligibility when adjusting for the other two variables. However, no significant interactions were found. Specifically, the hypothesis that flattened sentences are as intelligible as natural sentences for more advanced learners was not supported due to the change of experimental subjects from native speakers to L2 speakers. It was proposed that compared to native speakers, L2 speakers' underdeveloped utilization of secondary cues and semantic contexts, due to a developing proficiency level, may lead to non-significant interactions. The finding of the effect of F0 on intelligibility also illustrates the importance of tone accuracy and diversifying L2 learners' linguistic input in Chinese pronunciation teaching and learning.

Keywords Intelligibility · Fundamental frequency · Listening environment · Proficiency

K. Chen (✉) · C. Yang
University of Connecticut, Storrs, CT, USA
e-mail: kaidi.chen@uconn.edu

© Springer Nature Singapore Pte Ltd. 2021
C. Yang (ed.), *The Acquisition of Chinese as a Second Language Pronunciation*,
Prosody, Phonology and Phonetics, https://doi.org/10.1007/978-981-15-3809-4_10

213

1 Introduction

1.1 *Intelligibility*

Intelligibility is one of the most important constructs in second language (L2) pronunciation research. However, there is no universally agreed definition and measure of intelligibility (Munro and Derwing 1999; Pickering 2006; Chen 2011, among others), likely due to its confusion with comprehensibility. Smith and Nelson (1985) defined intelligibility as listeners' ability to recognize individual words or utterances. They further pointed out that miscommunication occurs when people only recognize words and utterances but fail to understand the meaning (termed as comprehensibility), or the pragmatic meaning behind them (termed as interpretability). In Smith and Nelson's definition, intelligibility and comprehensibility are closely related to each other but refer to speech understanding at different levels.

In another line of literature, intelligibility was broadly defined as the extent to which a speaker's message is actually understood by a listener (Munro and Derwing 1999; Derwing and Munro 2005). Levis (2018) interpreted it as "the extent to which a speaker is understandable" in a "narrow sense," and "whether the particular words used by a speaker are successfully decoded (the lexical level intelligibility)" in a "broad sense." It is measured by orthographic transcription tasks, i.e., percentage of words correctly transcribed (Munro and Derwing 1999; Derwing and Munro 2005; Yang 2016, among others). Different from Smith and Nelson (1985), comprehensibility was defined by Derwing and Munro as listeners' perception of the degree of difficulty in understanding an utterance. Comprehensibility is usually measured by scalar judgment tasks, from "extremely easy to understand" to "extremely difficult to understand" (Derwing and Munro 2005). However, it is worth pointing out that even if we recognize every single word and utterance, it does not mean that we can understand it when listeners do not have enough background knowledge. Even if we understand the utterance in the context, it does not mean that we need to recognize every single word, and in many cases, we do not have to do so. Thus, definitions and measures of intelligibility in both narrow sense and broad sense were considered in this study.

1.2 *Factors Affecting Intelligibility*

1.2.1 **Fundamental Frequency**

Fundamental frequency, referred to as F0, is the lowest frequency of a complex periodic sound. F0 determines pitch contour generally and expresses intonation (broadly speaking) linguistically. Although F0 does not influence segmental parts (consonants and vowels) of the speech, its prosodic feature has many linguistic functions: distinguishing lexical meaning (only for tonal languages), discriminating declarative

and interrogative sentences, marking emphasis, and paralinguistic functions such as expressing emotions (such as F0 increase in anger or fear, and F0 decrease in grief, sorrow, or depression). According to Lehiste (1970, cited from Binns and Culling 2007), important content words will be accented in normal speech and the corresponding F0 tends to be above the average F0 of the sentence. In this sense, the content words will be acoustically clearer than surrounding words, together with the contributions from the factor that content words are often articulated louder and more slowly. However, when F0 is flattened, none of the words in the sentence are accented and all of them are at the same F0 (Binns and Culling 2007). Without F0 cues, it would be difficult to find where the content or important words are. In an inverted F0 contour, the accented content words will go to opposite directions: a fall will be a rise and vice versa; F0 above the average will be below the average F0 and vice versa. As a result, no F0 cues will highlight important words in monotonous sentence; the F0 cues in inverted sentences will be misleading and highlight words that are not important to the meaning of the sentence (Binns and Culling 2007).

Previous studies (Maassen and Povel 1984; Laures and Weismer 1999) have investigated the role of intact fundamental frequency (F0) contours on the intelligibility of non-tonal languages and indicated that lack of intact F0 will decrease intelligibility. Laures and Weismer (1999) tested a typical group who did not self-report hearing loss or professional training in speech science and experimental psychology. Their results showed that the intelligibility of English sentences in terms of both word transcription and interval scaling were significantly lower when F0 contour was flattened, as compared with naturally varying contours. Maassen and Povel (1984) explored the role of fundamental frequency on the intelligibility in atypical population, namely, deaf children who are frequently reported to have monotonous voice. The overall results showed that when the original F0 contour of Dutch sentences from the deaf utterances was replaced by artificial contours, the percentage of the identified words increased significantly (although the change is small). It led to the conclusion that intonation correction yields significant improvement of intelligibility.

1.2.2 Listening Environment

Many studies (Laures and Bunton 2003; Binns and Culling 2007; Watson and Schlauch 2008; Miller, Schlauch and Watson 2010) included listening environment when examining the role of F0 on intelligibility. They consistently demonstrated that dynamic F0 contours are significant to speech intelligibility when taking background noise into account. Results from Laures and Bunton (2003) showcased that the absence of fundamental frequency variation has a significant impact on overall speech intelligibility. A flattened fundamental frequency contour negatively influences intelligibility when taking account of the competing listening background (white noise and multi-talker babble noise). Watson and Schlauch (2008) had similar findings that sentences with flattened F0 yielded poorer intelligibility than the unmodified ones in white noise. Their study also tested the effects of resynthesized F0 that reflected the average low F0, the median F0 and the average high F0. Sentences

flattened at the average high F0 yielded poorer intelligibility than that at the median F0, and the average low F0 yielded better intelligibility than that at the median F0. Binns and Culling (2007) compared the effects of intact F0 contour on intelligibility with flattened F0 and inverted F0 in adverse listening conditions. They found that against speech-shaped noise, flattened F0 has no significant impact on speech reception thresholds (SRTs) while inverted F0 does increase SRTs significantly, compared to intact F0 contour; however, when against single-talker interferer, both flattened and inverted conditions have greater effects and significantly increase SRTs. Therefore, it was concluded that intact F0 actually improves the intelligibility in noise, as compared to monotone or inverted F0. Building upon research on flattened and inverted F0, Miller, Schlauch and Watson (2010) further investigated how the F0 manipulations affect intelligibility in background noise. They had unmodified F0, flattened F0 at the median, natural but exaggerated F0, inverted F0, and sinusoidally frequency modulated F0. The results showed that the last two F0s (which create misleading cues) have more detrimental effect on speech intelligibility than flattened F0 and intact F0 in background noise.

1.2.3 Semantic Context

Semantic context is a factor often considered to help listeners recognize and understand an utterance. For example, Cole and Perfetti (1980) used the task that children and adults listen to mispronunciations in a children's story to test the role of context on words recognition. It is suggested that children detected mispronunciations more accurately when they occurred in highly predictable context, and all age groups detected the mispronunciations more quickly in predictable words. Craig, Kim, Rhyner and Chirillo (1993) examined the interaction of acoustic information with contextual information during speech perception. The results showed that predictability-high (PH) words were recognized earlier and with greater confidence than predictability-low (PL) words for all ages ranging from 5 to 83. Later, researchers also started to combine listening conditions with context. Fernald (2001, cited from Zhou, Li, Liang, Guan, Zhang, Shu and Zhang 2017) claimed that previous work showed that children as early as two years old are able to use semantic context to assist speech recognition in quiet. Pichora-Fuller and Daneman's (1995) experiment illustrated that old adults derived more benefit from supportive context (with sentence-final words that were either predictable context or unpredictable context) than young adults in babble background. Sheldon, Pichora-Fuller and Schneider (2008) explored how younger and older adults benefited from context when identifying target words in noise-vocoded sentences. The first type of context is either highly predictable or not predictable sentence-final target words, and the second type of context is either with priming or not. The results indicated that younger and older adults benefited from each type of context and with the most benefit gained when both types were combined. Similarly, Dubno, Ahlstrom and Horwitz (2000) found that both older and younger adults with normal hearing derived equivalent benefit from context given equivalent speech audibility in noise. Benichov, Cox, Tun and

Wingfield (2012) included more factors than previous researches. They confirmed the robust role of linguistic context to aid spoken word recognition when taking age, hearing acuity, verbal ability, and cognitive function into consideration.

1.2.4 Intelligibility of Mandarin Chinese

As we can see from the above literature review, fundamental frequency, listening background, and semantic context are three important factors jointly affecting intelligibility. However, most of previous studies examined the intelligibility of non-tonal languages, primarily English. In tonal languages, such as in Mandarin Chinese, tones are lexically specified and lexical tones distinguish lexical meanings from otherwise identical strings of phonemes (Wang 1973; Wang, Shu, Zhang, Liu and Zhang 2013; Xu, Zhang, Shu, Wang and Li 2013). Different from lexical tones in tonal languages, F0 or intonation in non-tonal languages is mainly used for pragmatic purposes, such as sentence modality, emphasis, and emotion (Cutler, Dahan and Donselaar 1997). In this sense, it is expected that F0 may play a more important role in the intelligibility of tonal languages than that of non-tonal languages.

Only a very limited number of studies (Liu and Samuel 2004; Patel, Xu and Wang 2010; Wang et al. 2013; Xu et al. 2013; Chen, Wong and Hu, 2014; Zhou et al., 2017) have investigated the intelligibility of Mandarin Chinese. Liu and Samuel (2004) found that in whispered speech, identification of tonal patterns remains “surprisingly” good when the F0 information is neutralized. Native Mandarin listeners can use secondary cues (i.e., duration and amplitude) when the primary cue (F0) is unavailable. The prediction from the finding of whispered speech to flat-F0 speech was questioned by Patel et al. (2010) in that flat-F0 has voicing while whispered speech does not and F0 gives prominent cue for tone perception (Whalen and Xu 1992, cited from Patel et al. 2010). Patel et al. (2010) conducted their own experiment on the role of intact and flattened F0 when controlling for listening environments. They found that for native Mandarin listeners, monotonic speech is just as intelligible as natural speech in a quiet background, but the flat-F0 speech became substantially less intelligible than natural speech when noise was added. Their finding was corroborated by behavioral experiments by Xu et al. (2013) in which listeners (native Mandarin speakers with minimal music experience) rated monotone sentences as equally intelligible as normal sentences; it was also supported by Chen et al. (2014) which found that normal hearing listeners (native Mandarin speakers) perfectly recognized Mandarin sentences produced with modified tone contours (flat tone or tone randomly selected from the four Mandarin lexical tones) in a quiet environment, but their performance declined in noise. Furthermore, the fMRI result by Xu et al. (2013) provided an explanation for the equative intelligibility of flat F0 and natural F0 (regardless of listening background). Monotone sentences elicited greater activation in the left planum temporale (PT), demonstrating the automatic use of additional neural resources to recover the phonological loop from altered tonal patterns. However, the preceding studies did not explain what cues are utilized for comprehension when the sentence is flattened. Wang et al. (2013) investigated

the role of sentence context on intelligibility together with F0 contour and listening environment. It is revealed that for native Mandarin listeners, word list sentences with natural F0 contours were less intelligible than normal sentences counterparts in both quiet and noise conditions, indicating that sentence context improves speech intelligibility regardless of listening backgrounds; they also argued that sentence context partially explained the unchanged intelligibility of monotonous sentences in the quiet environment. Zhou et al. (2017) corroborated the influence of semantic context on intelligibility together with factors of F0 and listening environment by elementary and middle-school-aged children. Children of both age groups use semantic context to assist speech recognition; with flat F0 contours, younger children are worse in making use of context in recognizing speech than older children. Considering the interactions and joint impact of sentential semantic context, F0 contours and listening environments on Mandarin speech intelligibility by native Mandarin speakers, both children and adults, it would be interesting and worthwhile to examine how these factors affect L2 Mandarin speakers and whether there are any differences between native and L2 speakers.

This study attempts to investigate the effects of F0 (i.e., natural F0 versus flattened F0) on the intelligibility of Mandarin speech by L2 Mandarin learners in quiet and white noise conditions when controlling for sentence context. Intelligibility in present study is defined at two levels: it consists of both word and utterance recognition (Kirkparick et al., 2008), and to what extent a listener can understand the locutionary meaning of a message (Munro and Derwing, 1999). Previous studies have shown L2 Mandarin speakers' real-time perceptual development toward more native-like directions (in both reaction time and accuracy) on Mandarin tones AX-discrimination task (Wiener, 2017) and advanced L2 Mandarin learners' better perception of Mandarin intonation and better identification of intonation-superimposed tones as compared to the first- and second-year learners (Yang, 2016). Zhou et al. (2017) also showed the developmental changes of native Mandarin speakers' speech intelligibility (Zhou et al., 2017). To this end, we also want to examine how L2 learners of Mandarin at different proficiency levels differ in speech intelligibility.

We address the following questions in this study:

- (1) What are the effects of F0 (natural versus flat) and listening environment (quiet versus noise) on Chinese intelligibility when keeping semantic context constant?
- (2) How does proficiency level affect L2 Chinese listeners' intelligibility?
- (3) What are the interactions of F0, listening environment, and proficiency level in L2 Mandarin intelligibility?

Drawing upon the discussions above, we make the following predictions: noise, flat F0, and low proficiency will all reduce intelligibility when holding context constant. There are interactions among F0 variations, listening environment, and proficiency level. Specifically, in quiet background, flattened sentences are as intelligible as natural sentences for more advanced learners, but not for lower proficiency learners. When noise is added, the intelligibility of both flat and natural sentences will drop across all proficiency levels.

2 Methodology

2.1 Participants

Twenty L2 Mandarin learners, 4 at each of the 5 proficiency levels (level 1, 2, 2.5, 3, and 4), from an intensive summer program in the USA, were recruited for this study. At the beginning of the summer program, students were placed into these five levels according to their performance in the informal ACTFL standardized Oral Proficiency Interviews (OPI) conducted by the instructors of the summer school. They participated in the research at the end of the summer program.

2.2 Stimuli

18 Chinese sentences were created by the first author and read by a female Beijing Mandarin speaker in her 30s. All vocabulary and grammar were taken from the following Chinese textbook: *Integrated Chinese (volume 1 and 2)* (Liu, Yao, Bi, Ge and Shi 2016), *Basic Mandarin Chinese* (Kubler 2017) and *Intermediate Spoken Chinese* (Kubler 2013). Appendix 1 presents the whole list of these sentences in Chinese characters and their English translations. To help participants become familiar with the task, two practice sentences were prepared. Additionally, five filler sentences were inserted among the target sentences intermittently to alleviate the impact from cognitive confounding variables such as attention.

Praat (Boersma and Weenink 2018) and Praat vocal toolkit (Corrette 2012–2020) were used to manipulate the stimuli. Specifically, monotones were created by flattening the F0 contour of each sentence at the sentence's mean F0 (Fig. 1). In this sense, pitch-flattened sentence neutralized the intonations and lexical tones while keeping other syllabic and sub-syllabic acoustic information (such as intensity and duration) intact. White noise at + 65 SNR level was added. After manipulations, there were four conditions for each sentence: natural tone, natural tone + 65 dB noise, flat tone, and flat tone + 65 dB noise. Altogether there are 100 sentences (25 sentences \times 2 F0 conditions \times 2 noise conditions). All 100 sentences were amplitude normalized using Praat. Then the sentences were randomized into four blocks, equally distributed across the F0 conditions and noise conditions. Each block has all 25 sentences from the sentence list but in different F0 and noise conditions, all counterbalanced.

2.3 Procedure

Participants were recruited through the help of the instructors of various classes. When participants came to the study, they were given the consent forms first and were

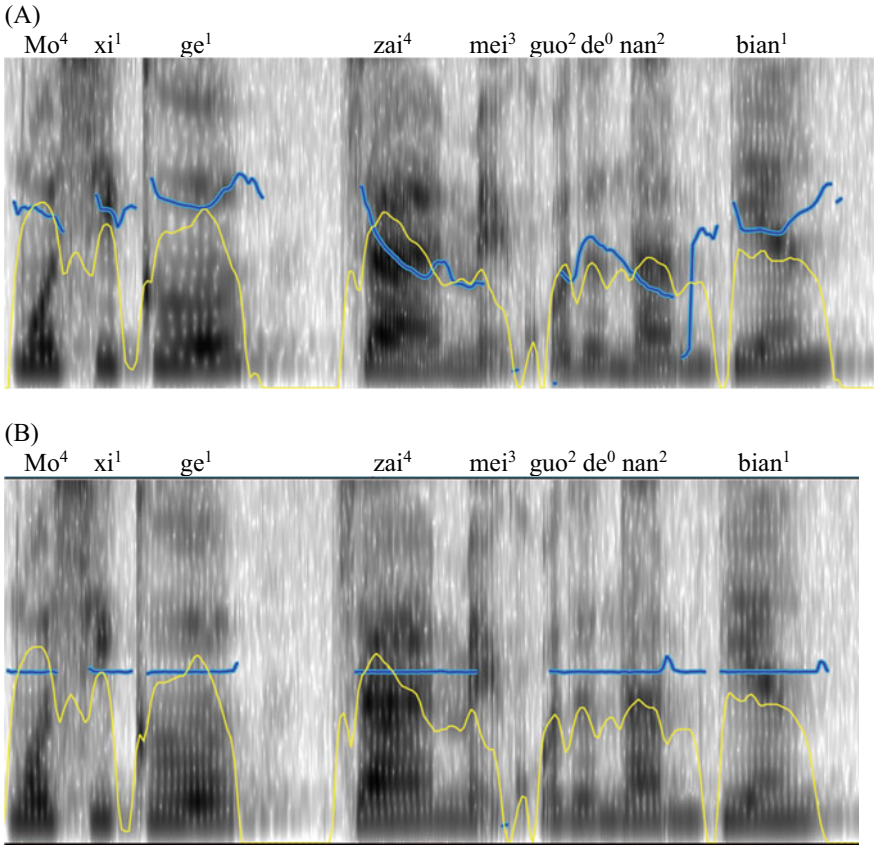


Fig. 1 Acoustic features of sample speech stimuli. Broadband spectrograms in black, intensity contours in yellow, and F0 contours in blue. Panel A: normal (natural F0) sentence. Panel B: F0-flattened counterpart

asked to read and sign before starting the task. Participants were tested individually in a quiet classroom while facing a Mac Pro. They heard each sentence from the speaker of the laptop at a comfortable level. The participants were asked to write down the sentences they heard in either Chinese characters or *pinyin* Romanization. The verbal instruction was in Chinese only, due to the “only Chinese” language pledge signed by all the students in the summer program. To ensure that the participants understood the Chinese instructions, written English instructions were also provided. The progression of the task was controlled by the first author. After listening to one sentence, the participants wrote down the sentence and then translated it into English on the answer sheet. Then the first author would proceed to the next sentence. Each participant listened to each stimulus only once.

The whole task took around 10–15 min. To avoid learning effect and confounding factors, such as attention and fatigue, each participant only listened to one block of

stimuli and the other three participants from the same level listened to the rest of the three different blocks.

2.4 The Measurement of Intelligibility

Following Lane (1963), Munro and Derwing (1999), Derwing and Munro (2005), and Yang (2016), intelligibility was measured by the proportions of correct syllables over the total syllable numbers in a sentence. Only when consonant(s) (if any), vowel, and tone were all correct, was a syllable considered to be correct. Because we adopted both narrow and broad measures of intelligibility, apart from word and utterance recognition, understanding of the sentence was also our concern. English translation was used to test whether participants understood the sentences correctly. If their translation was wrong, they did not really comprehend the meanings of the sentence. Therefore, as long as the participants did not translate the sentence correctly, even though they had correctly transcribed syllables, the syllables were not taken as correct. Correct intelligibility should include both correctly transcribed syllables and correct English translations. The intelligibility score was calculated for each sentence in different F0 and noise conditions.

Table 1 gives two examples of how intelligibility scores were calculated. This correct sentence, “Wǒ fēi cháng xǐ huān běi jīng dòng wù yuán,” was the baseline and we calculated how many syllables each participant transcribed and translated correctly. Participant A did not write the last five syllables (běi jīng dòng wù yuán) correctly, but instead wrote “zhōng guó rén” and accordingly translated it as “Chinese people” wrongly. This participant only transcribed and translated the first five syllabus correctly. Thus, $5/10 = 0.5$ is participant A’s intelligibility score for this sentence. Participant B listened to the same sentence but in different conditions (natural tone without noise). A mixture of Chinese characters and *pinyin* were given in the answer. This participant transcribed two syllables (dōng wú) wrongly. Thus, two correct syllables were missing in the transcription. For the syllable “yuán,”

Table 1 Samples of intelligibility scoring

Target sentence and its translation (condition)	Transcription and translation (participant)	Intelligibility score
wǒ fēi cháng xǐ huān běi jīng dòng wù yuán. (natural tone in noise) I really like Beijing Zoo	wǒ fēi cháng xǐ huān zhōng guó rén. (Participant A) I really like Chinese people	0.5
wǒ fēi cháng xǐ huān běi jīng dòng wù yuán. (natural tone without noise) I really like Beijing Zoo	我非常喜欢北京的 dōng wú yuán. (Participant B) I really like Beijing’s pet stores	0.7

although it was transcribed correctly, the translation was wrong. As a result, only the first 7 syllables got credits, and participant B’s intelligibility score for this sentence is “7/10 = 0.7”.

3 Data Analysis and Results

We used a mixed-effect model with proficiency level, flat tone, and noise as fixed variables and sentence number as a random variable. In this case, semantic context was held constant when testing other variables. The model can be written as:

$$y_i = \text{Noise} \times \beta_1 + \text{Proficiency Level} \times \beta_2 + \text{Flat Tone} \times \beta_3 + \text{Sentence Number} \times u + \epsilon$$

where y_i represents each intelligibility score.

First, we looked at the two-way and three-way interactions, and no significant interactions were found between/among any variables, as shown in Table 2.

Since there were no interactions, we excluded the interactions from our model. Table 3 presents the main effects of all three predictors, and tables in the appendix (see appendix 2) show the estimated marginal means for noise and flat tone from models without interactions. It can be seen that different proficiency levels predict

Table 2 Analysis of variance with interactions

Variable	F	Degrees of freedom	Degrees of freedom of the residues	P value	Eta_sq
ProficiencyLevel	32.11	4	323	< 0.0001	0.237
FlatTone	31.41	1	323.52	< 0.0001	0.059
Noise	47.62	1	323.51	< 0.0001	0.087
ProficiencyLevel:FlatTone	0.25	4	323.02	0.91	0.002
ProficiencyLevel:Noise	0.44	4	323.04	0.78	0.003
FlatTone:Noise	3.48	1	323.02	0.06	0.006
ProficiencyLevel:FlatTone:Noise	0.53	4	323.02	0.71	0.004

Table 3 Analysis of variance without interactions

Variable	F	Degrees of freedom	Degrees of freedom of the residues	P value	Eta_sq
ProficiencyLevel	32.57	4	336	< 0.0001	0.24
FlatTone	31.83	1	336.53	< 0.0001	0.06
Noise	48.24	1	336.5	< 0.0001	0.09

different intelligibility scores ($p < 0.0001, \eta^2 = 0.24$) when controlling for noise and flat tones; compared to flat tones ($M = 0.51$), natural tones ($M = 0.64$) predict a higher intelligibility score ($p < 0.0001, \eta^2 = 0.06$) when taking proficiency level and noise into account; compared to noise condition ($M = 0.49$), no noise condition ($M = 0.66$) predicts a higher intelligibility score ($p < 0.0001, \eta^2 = 0.09$) over proficiency level and flat tones. In addition, the effect size of proficiency level is large, accounting for 24% of the variance of sentence scores; the effect size of flat tone is medium, explaining 5.9% of the variance; the effect size of noise is medium, with 8.6% of variance in sentence scores explained. The main effects of the three variables can also be observed in Fig. 2.

Finally, to investigate how specific proficiency level predicts the intelligibility score, we compared each level to a reference level. The reference level here is proficiency level 1. Tables 4 and 5 present the estimated marginal means for each proficiency level and the pairwise differences across levels. We can see from the tables that there are no significant differences between proficiency Level 2 and Level 2.5 and between Level 3 and Level 4. For the rest of the comparisons, they are significantly different in predicting intelligibility scores when adjusted for noise and flat tone. Specifically, Level 2 ($M = 0.49$), Level 2.5 ($M = 0.56$), Level 3 ($M = 0.68$) and Level 4 ($M = 0.76$) significantly predict higher intelligibility scores than Level 1 ($M = 0.38$), $p_{2-1} < 0.05, p_{2.5-1} < 0.05, p_{3-1} < 0.05, p_{4-1} < 0.05$; Level 3 ($M = 0.68$)

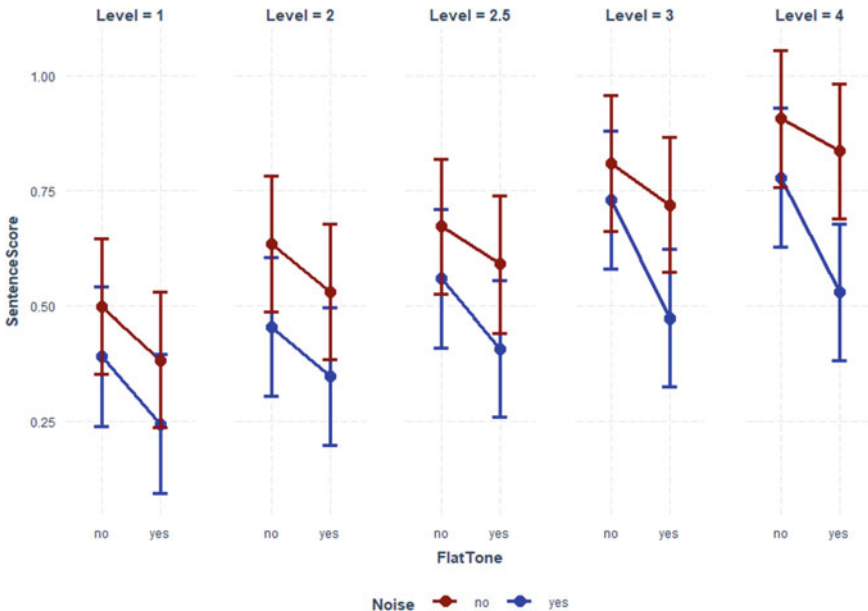


Fig. 2 Relationships of proficiency level, noise, flat tone, and sentence score

Table 4 Estimated marginal means for proficiency level

<i>Level</i>	<i>Emmean</i>	<i>SE</i>	<i>Df</i>	<i>Lower.CL</i>	<i>Upper.CL</i>
1	0.38	0.06	24.20	0.25	0.50
2	0.49	0.06	24.19	0.37	0.61
2.5	0.56	0.06	24.19	0.43	0.68
3	0.68	0.06	24.19	0.56	0.80
4	0.76	0.06	24.19	0.64	0.89

Table 5 Differences between estimated marginal means across proficiency level

<i>Contrast</i>	<i>Estimate</i>	<i>SE</i>	<i>Df</i>	<i>T.ratio</i>	<i>p.value</i>
1 - 2	-0.1153	0.0379	336.0000	-3.0448	0.0210
1 - 2.5	-0.1795	0.0379	336.0000	-4.7412	0.0000
1 - 3	-0.3056	0.0379	336.0000	-8.0699	0.0000
1 - 4	-0.3863	0.0379	336.0000	-10.2000	0.0000
2 - 2.5	-0.0642	0.0379	336.0000	-1.6963	0.4376
2 - 3	-0.1903	0.0379	336.0000	-5.0255	0.0000
2 - 4	-0.2710	0.0379	336.0000	-7.1558	0.0000
2.5 - 3	-0.1261	0.0379	336.0000	-3.3291	0.0085
2.5 - 4	-0.2067	0.0379	336.0000	-5.4592	0.0000
3 - 4	-0.0807	0.0379	336.0000	-2.1303	0.2097

and Level 4 ($M = 0.76$) significantly predict higher sentence scores than Level 2 ($M = 0.49$), $p_{3-2} < 0.05$, $p_{4-2} < 0.05$; Level 3 ($M = 0.68$) and Level 4 ($M = 0.76$) significantly predict higher sentence scores than Level 2.5 ($M = 0.56$), $p_{3-2.5} < 0.05$, $p_{4-2.5} < 0.05$.

4 Discussions

This study investigated the role of F0, listening environment (with or without noise), and proficiency level on the intelligibility of Mandarin Chinese by L2 Mandarin learners. The semantic context in target sentences was held constant (sentence number as a random factor) when testing other variables; in this case, the finding can be generalized to any Mandarin sentence in any semantic context. The three variables, F0 contour, listening environment, and proficiency level, were all found to affect Mandarin intelligibility by L2 Mandarin learners. That is to say, the lack of natural F0 contour, the presence of noise, and the lower proficiency level, would all predict reduction in intelligibility. The relationship of different proficiency levels and intelligibility was also confirmed.

Although the effects of F0 contour and noise are consistent with previous studies (Patel et al. 2010; Wang et al. 2013; Zhou et al. 2017), no significant interactions we hypothesized were found. Looking at Fig. 2, we can see that in a quiet environment, the intelligibility of flat-F0 sentences is lower than that of natural F0 sentences across all proficiency levels. In a noise condition, the pattern is similar, across all proficiency levels. The non-significant interactions of F0 contour and background noise are inconsistent with previous researches on native Mandarin speakers (Patel et al. 2010; Wang et al. 2013; Xu et al. 2013; Chen, et al. 2014; Zhou et al. 2017). These studies have found that the difference of intelligibility of flat F0 and natural F0 sentences depends on the listening environment; namely, flat F0 speech in a quiet environment is as intelligible as natural F0 speech, but in a noise environment, flat F0 dramatically reduced the intelligibility compared to the mild decrease for natural F0 sentences. It was argued that such finding highlighted “the importance of natural F0 contour for sentence intelligibility in noise” (Wang et al. 2013) and “the robustness and flexibility of spoken Mandarin comprehension” (Patel et al. 2010).

We argue that the inconsistent findings on the interactions of factors affecting Mandarin intelligibility are likely due to the change of subjects from native Mandarin listeners to L2 Mandarin listeners. Studies on L2 Mandarin suprasegmentals (Wiener 2017; Yang 2016) have showcased that there are either real-time developments on tone perceptions after classroom learning or various tone and intonation perceptions of L2 learners at different proficiency levels. For example, Yang (2016) found that with respect to the identification of intonation of statements, particularly for those ending with tone 2, native speakers were far more accurate than first-year L2 learners, second-year L2 learners and advanced L2 learners. Yang (2016) interpreted that as L2 learners’ proficiency improved over time, their perception of statement intonation also improved. Furthermore, Yang (2016) proposed that native and L2 listeners may be attending to different cues in perceiving intonation types: native listeners attend to both “global and localized F0 cues” in identifying intonations while L2 listeners primarily depend on “localized terminal F0 cues (mainly the tone of the last syllable).” The difference of mechanism in intonation identification of native and L2 listeners may help explain the different findings on Mandarin intelligibility to some extent. That is to say, L2 listeners tended to focus more on individual words when transcribing and translating, rather than focus on the entire sentence. Yang (2016) also discovered the difference in tone identification between native and L2 listeners: both native and advanced L2 listeners performed much better than first- and second-year L2 listeners. Results also showed a path of improvement from first year to advanced L2 learners in tone perception. Given the aforementioned findings of native and L2 difference in perception of tones and intonation, we assume that if participants in our current study only listen to natural F0 contour sentences, native and L2 listeners will perform differently in intelligibility task. However, for the flattened F0 contour sentences, it would be expected that L2 listeners would not be any worse than native speakers since neither group had tonal and intonational cues to rely on for intelligibility. In other words, L2 listeners in our study were supposed to be similar to native listeners in previous studies in terms of the intelligibility of flat-F0 sentences in a quiet environment. However, the interaction of F0 and the listening environment was

not borne out, implying that there are some other cues native listeners can access to assist intelligibility but L2 listeners cannot.

Besides F0, previous studies have shown that native listeners make use of secondary cues, such as duration, amplitude, or acoustic boundaries/landmarks, when F0 cues are not accessible (Liu and Samuel 2004; Li and Loizou 2008; Patel et al. 2010; Chen et al. 2014). Thus, due to their limited exposure to Mandarin Chinese, L2 learners are not as good as native speakers at making use of these secondary cues when tones and intonations were flattened. Thus, we propose that the constraints of proficiency, specifically the underdeveloped utilization of secondary cues other than tone contours, may lead to the non-significant interactions.

As we stated in the introduction, context is also a big factor influencing L1's intelligibility (Wang et al. 2013; Zhou et al. 2017). This could also be one aspect that L2 listeners lack. Since we have controlled semantic context to be constant in the present study, we could not know how different contexts affect L2's intelligibility. It is possible that L2 learners might still be in the process of developing the sensitivity to semantic context.

When we look at Fig. 2, we could clearly see that as proficiency level improves, the slopes of the red horizontal lines and the blue horizontal lines are progressing toward a converging point, showing their possible tendency to interact with each other and move off the parallels. We argue that two factors may be playing a role here. Firstly, the L2 learners in this study, including the Level 4 learners, are still in the process of developing their proficiency. This is due to their limited exposure to Mandarin Chinese, especially in terms of both phonetic/phonological variations often occurring in actual communication and the phonotactic constraints in the language. In this sense, the Level 4 participants are still not advanced enough, at least not native-like. On the other hand, the small sample size in our study is another factor which may prevent the occurrence of the interaction of flat F0 and noise. Future studies can be expanded to include more advanced L2 learners and increase the sample size of each level to 20 or 30.

Lastly, the measure of intelligibility in this study may lead to the inconsistent findings from previous studies. In this study, we adopted both narrow and broad definition of intelligibility and the measure we used included both the orthographic transcription and English translation. However, previous researchers used various measures of intelligibility, such as orthographic transcription (Patel et al. 2010; Wang et al. 2013), verbal repetition (Zhou et al. 2017), and scale ratings of comprehension (Xu et al. 2013). They are either mere recognition (Patel et al. 2010; Wang et al. 2013; Zhou et al. 2017) or comprehension (Xu et al. 2013). But none of these studies combined transcription/recognition and translation/comprehension in their measurement of intelligibility.

5 Pedagogical Implications, Limitations and Future Studies

This study has significant pedagogical implications. The finding of the effect of F0 on intelligibility highlights the importance of tone accuracy in L2 Mandarin teaching and learning. Although monotone sentences can be as equally intelligible as natural F0 sentences for native speakers in a quiet environment, this unfortunately does not apply to L2 learners. L2 listeners' ability to use secondary cues, such as duration and amplitude, is still developing, and their limited experience and exposure do not provide them with phonetic/phonological variations and the (implicit) knowledge of Mandarin phonotactic constraints. Thus, they do not have the resources to rely on to recognize and comprehend utterances when F0 is not available in both quiet and noise environments. To help L2 learners become better listeners, tone accuracy should be emphasized in L2 Chinese classes, not only at the beginning level, but also at the intermediate and advanced levels. More importantly, tone training should be incorporated in meaningful communicative activities or focus-on-form tasks in addition to mechanical drillings (Yang 2016 and 2020). To help L2 learners understand well in undesirable environments, such as in a noisy listening condition, they should be provided with access to different types of linguistic input. For example, L2 learners should listen to both slow speech and fast speech, both standard speech and non-standard, or even accented speech, and both speech by native speakers and speech by non-native speakers. By exposing L2 learners to a diversity of linguistic input and integrating tone and pronunciation training in task-based pronunciation activities in various listening environments, L2 learners will acquire allophonic/allotonic knowledge of Mandarin tones and learn to use secondary acoustic cues (i.e., duration and amplitude).

One limitation of current study is that we did not consider individual difference. Cognitive variables, such as attention and working memory, vary from person to person. Fatigue can also be a confounding variable as the first author has witnessed some participants saying "very tired" when coming to testing venue for this study right after their immersion class. The alternative choice of either transcribing in Chinese characters or *Pinyin* is also a limitation in manipulating individual difference. If participants have not formed automatic connection between meaning, sound and form yet, it would cost them more cognitive resources to write characters, which may lower their intelligibility scores compared to *pinyin* users. The first author witnessed a participant stuck with a character and miss the remaining part of a complete sentence. Additionally, individual's attitudes and strategies are different. After missing some words from the recording, some were more "risky" and would try their best to recall and guess what it might be and wrote them down, while others may be very "conservative" (frustrated as well) and gave up the whole sentences. Future studies are expected to take all these individual differences into consideration.

Another limitation or a confounding factor is the way intelligibility is measured. The orthographic transcription measure was carried out in such a way that if the answers were in *pinyin*, only by transcribing all segmental (consonants and vowels) and suprasegmental (tones) components of a syllable correctly, can they be treated

as correct. We observed that some participants did not write tone marks, but all consonants, vowels, and translations were correct. They lose that intelligibility score for doing so. However, we do not know whether it was because they just forgot the tone marks or they did not recognize the tones. Since it is common to see L2 Mandarin learners ignore the tone marks when writing Chinese *pinyin* because of the lack of suprasegmental counterparts in their native language English, it is possible that in this study, they already recognized the tones and understood the sentences, but just forgot to write down tones. If it was the case, could the incorrect tones only be treated as typos, like misspellings in English, and credits should not be deducted. Unfortunately, we had no idea of which scenario out of the two lead to the lack of tone marks in some sentences. As a result, we adopted a more stringent and consistent measure and deducted points for those cases without tone marks. Future studies may require the testers to monitor participants' response and remind them to always include tone marks when transcribing in *pinyin* to avoid the potential ambiguity in intelligibility measurement.

This study expands previous studies on Mandarin intelligibility by focusing on L2 Mandarin learners across proficiency levels. Future studies are warranted to further examine the possible interaction of flat F0 and noise, and the chance of achieving closer intelligibility to native speakers, by including L2 learners of various proficiency levels and increasing the sample size. We could further explore at what advanced proficiency level or threshold L2 learners can recognize and understand the flattened sentences in the quiet environment as native speakers do, namely the issue of ultimate attainment in L2 intelligibility.

As argued in the discussion part, secondary cues like amplitude, duration, and acoustic boundaries may assist listener's intelligibility when sentences are flattened, especially in quiet environment. We have yet to know to what extent L2 learners may utilize these cues and what are their relationships with intelligibility. More studies are needed to explore L2 learners' developing competence of using secondary cues. In terms of semantic context, although we controlled sentence semantic variations and make it constant by statistical measures to reduce total errors, we still do not know how it impacts L2 learners' speech intelligibility in different semantic context. Future studies can examine whether normal sentences and wordlist sentences make a difference to intelligibility judgment.

6 Concluding Remarks

This study examined the effects of fundamental frequency, listening environment, and proficiency levels on the intelligibility of Mandarin Chinese by L2 learners. The findings revealed that flattened F0, background noise, and lower proficiency levels all lead to the decrease in intelligibility when holding semantic context constant. However, no interactions were found among the three factors, which is not consistent with previous finding on native Mandarin speakers. The hypothesis on the difference

of the intelligibility of flat F0 speech and natural F0 speech in quiet and noise environments for advanced learners were not borne out. Different from native speakers, L2 Mandarin learners did not understand the flat F0 and natural F0 sentences equally well in the quiet environment. As a matter of fact, the intelligibility of flat F0 sentences was lower for L2 learners across proficiency levels. Several accounts were proposed for the non-significant interactions and discrepancy between native speakers and L2 learners, such as the underdeveloped capability for the utilization of semantic contexts, lack of knowledge of phonetic/phonological variations and phonotactic constraints, and not attending to secondary cues, such as amplitude, duration, and acoustic boundaries.

This study contributes to our understanding of intelligibility from the perspective of second language learners of a tonal language and supports the importance of tone accuracy and diversifying L2 learners' linguistic input in Chinese pronunciation teaching and learning. Future studies should incorporate larger sample size and more advanced L2 Mandarin learners to explore the possibility of ultimate attainment in L2 intelligibility.

Acknowledgements We acknowledge the help of Shuang Yin and Timothy E. Moore of UConn's Statistical Consulting Services.

Appendix 1. Sentences List

Sentence type	Chinese sentences	English translations
Practice 1	他是美国人。	He is American
Practice 2	我爸爸是律师。	My father is a lawyer
Filler 1	时间过得太快了!	Time flies!
Filler 2	我上个星期到加拿大去了。	I went to Canada last week
Filler 3	谢谢你百忙之中还抽空来看我。	Thank you for visiting me even when you are very busy
Filler 4	打太极拳的都是中老年人。	Those who play Tai-Chi are all old people
Filler 5	瑞士是个富有的国家。	Switzerland is a weather country
Target 1	我非常喜欢北京动物园。	I really like Beijing Zoo
Target 2	这是很久以前的事了。	This is the issue long time ago
Target 3	中国总共有几百种方言。	China has hundreds of dialects
Target 4	他对民国时期的文学特别感兴趣。	He is very interested in the literature of the Republic China
Target 5	爱情要紧还是面包要紧?	Love matters or bread matters?
Target 6	爱笑的人活得更长。	Who loves laughing lives longer

(continued)

(continued)

Sentence type	Chinese sentences	English translations
Target 7	很多事情说起来容易做起来难。	Many things are easy to say but hard to do
Target 8	请把今天的功课交给老师!	Please give today's homework to the teacher!
Target 9	他的护照被偷走了。	His passport was stolen
Target 10	红烧牛肉很好吃。	Braised beef is very delicious
Target 11	香港和澳门使用繁体字。	Hong Kong and Macau use traditional characters
Target 12	墨西哥在美国的南边。	Mexico is to the south of America
Target 13	法国是世界上最好的香水。	French has world's best perfume
Target 14	马友友是一位非常著名的音乐家。	Yoyo Ma is a well-known musician
Target 15	美国老一代的华人,大部分是从广东来的。	Old generation Chinese American mostly come from Guangdong
Target 16	在中国,孩子一定要听父母的话。	In China, children must heed what their parents say
Target 17	我有很多朋友。	I have a lot of friends
Target 18	今天天气很糟糕。	Today's weather is very terrible

Appendix 2

Estimated marginal means for noise from model without interactions.

Noise	Emmean	SE	Df	Lower.CL	Upper.CL
No	0.66	0.06	18.60	0.54	0.77
Yes	0.49	0.06	18.83	0.37	0.61

Estimated marginal means for flat tone from model without interactions.

Flat Tone	Emmean	SE	Df	Lower.CL	Upper.CL
No	0.64	0.06	18.77	0.53	0.76
Yes	0.51	0.06	18.65	0.39	0.62

References

- Benichov, J. C., Cox, L. A., Tun, P., & Wingfield, A. (2012). Word Recognition Within a Linguistic Context: Effects of Age, Hearing Acuity, Verbal Ability, and Cognitive Function. *Ear and Hearing, 33*(2), 250–256.
- Binns, C., & Culling, J. (2007). The role of fundamental frequency contours in the perception of speech against interfering speech. *the Journal of the Acoustical Society of America, 122*(3), 1765–1776.
- Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 14 March 2018 from <https://www.praat.org>
- Chen, F., Wong, L., & Hu, Y. (2014). Effects of lexical tone contour on Mandarin sentence intelligibility. *Journal of Speech, Language, and Hearing Research : JSLHR, 57*(1), 338–345.
- Chen, H. C. (2011). Judgments of intelligibility and foreign accent by listeners of different language backgrounds. *Journal of Asia TEFL., 8*, 61–83.
- Cole, R. A., & Perfetti, C. A. (1980). Listening for mispronunciations in a children's story: The use of context by children and adults. *J. Verbal Learn. Verbal Behav., 19*, 297–315.
- Corrette, R. (2012–2020). Praat Vocal Toolkit. <https://www.praatvocaltoolkit.com>
- Craig, C. H., Kim, B. W., Rhyner, P. M., & Chirillo, T. K. (1993). Effects of Word Predictability, Child Development, and Aging on Time-Gated Speech Recognition Performance. *Journal of Speech and Hearing Research, 36*(4), 832–841.
- Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics, 20*(1), 55–60.
- Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech, 40*(2), 141–201.
- Derwing, T., & Munro, M. (2005). Second Language Accent and Pronunciation Teaching: A Research-Based Approach. *TESOL Quarterly, 39*(3), 379–397.
- Dubno, J., Ahlstrom, J., & Horwitz, A. (2000). Use of context by young and aged adults with normal hearing. *the Journal of the Acoustical Society of America, 107*(1), 538–546.
- Fernald, A. (2001). *Making use of semantic context in early language understanding*. Minneapolis, MN: Paper Presented at the Society for Research in Child Development.
- Jenkins, J. (2000). *The Phonology of English as an International Language: New models, New Norms, New Goals*. Oxford: OUP.
- Kirkpatrick, A., Deterding, D., & Wong, J. (2008). The international intelligibility of Hong Kong English. *World Englishes, 27*(3–4), 359–377.
- Kubler, C. (2013). *Intermediate Spoken Chinese: A Practical Approach to Fluency in Spoken Mandarin*. Vermont: Tuttle Publishing.
- Kubler, C. (2017). *Basic Mandarin Chinese: Speaking & listening*. Vermont: Tuttle Publishing.
- Lane, H. (1963). Foreign accent and speech distortion. *Journal of the Acoustical Society of America, 35*(4), 451–453.
- Laures, J., & Bunton, K. (2003). Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions. *Journal of Communication Disorders, 36*(6), 449–464.
- Laures, J., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech, Language, and Hearing Research: JSLHR, 42*(5), 1148–1156.
- Lehiste, I. (1970). *Suprasegmentals*. Cambridge, Mass.: M.I.T. Press.
- Levis, J. (2018). *Intelligibility, Oral Communication, and the Teaching of Pronunciation (Cambridge Applied Linguistics) (Cambridge Applied Linguistics)*. Cambridge: Cambridge University Press.
- Li, N., & Loizou, P. (2008). The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise. *the Journal of the Acoustical Society of America, 124*, 3947–3958.
- Liu, S., & Samuel, A. G. (2004). Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech, 47*(2), 109–138.

- Liu, Y., Yao, D., Bi, N., Ge, L., & Shi, Y. (2016). *Integrated Chinese: Zhong wen ting shuo du xie (Volume 1 & 2)*. Boston, MA: Cheng & Tsui.
- Maassen, B., & Povel, D. (1984). The effect of correcting fundamental frequency on the intelligibility of deaf speech and its interaction with temporal aspects. *the Journal of the Acoustical Society of America*, 76(6), 1673–1681.
- Miller, S., Schlauch, R., & Watson, P. (2010). The effects of fundamental frequency contour manipulations on speech intelligibility in background noise. *the Journal of the Acoustical Society of America*, 128(1), 435–443.
- Munro, M., & Derwing, T. (1999). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, 49, 285–310.
- Patel, A. D., Xu, Y., & Wang, B. (2010). The role of F0 variation in the intelligibility of Mandarin sentences. *Proceedings of Speech Prosody 2010*, Chicago, IL
- Pichora-Fuller, M., Schneider, B., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *the Journal of the Acoustical Society of America*, 97(1), 593–608.
- Pickering, L. (2006). Current Research on intelligibility in English as a lingua franca. *Annual Review of Applied Linguistics.*, 26, 219–233.
- Sheldon, S., Pichora-Fuller, M., & Schneider, B. (2008). Priming and sentence context support listening to noise-vocoded speech by younger and older adults. *the Journal of the Acoustical Society of America.*, 123(1), 489–499.
- Smith, L., & Nelson, C. L. (1985). International intelligibility of English: Directions and resources. *World English*, 4, 33–342.
- Wang, J., Shu, H., Zhang, L., Liu, Z., & Zhang, Y. (2013). The roles of fundamental frequency contours and sentence context in mandarin Chinese speech intelligibility. *the Journal of the Acoustical Society of America*, 134(1), 91–97.
- Wang, W.S.-Y. (1973). The Chinese Language. *Scientific American*, 228(2), 50–60.
- Watson, P., & Schlauch, R. (2008). The Effect of Fundamental Frequency on the Intelligibility of Speech with Flattened Intonation Contours. *American Journal of Speech-Language Pathology*, 17(4), 348–355.
- Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, 49(1), 25–47.
- Wiener, S. (2017). Changes in Early L2 Cue-Weighting of Non-Native Speech: Evidence from Learners of Mandarin Chinese. *INTERSPEECH*.
- Xu, G., Zhang, L., Shu, H., Wang, X., & Li, P. (2013). Access to lexical meaning in pitch-flattened Chinese sentences: An fMRI study. *Neuropsychologia*, 51(3), 550–556.
- Yang, C. (2016). *The acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice* (Bilingual processing and acquisition; v. 1). Amsterdam: John Benjamins Publishing Company.
- Yang, C. (2020). Teaching Chinese Intonation and Rhythm. In C. Shei, M. E. McLellan Zikpi, & D. Chao (Eds.), *The Routledge handbook of Chinese language teaching* (pp. 180–194). London: Routledge.
- Zhou, H., Li, Y., Liang, M., Guan, C., Zhang, L., Shu, H., & Zhang, Y. (2017). Mandarin-Speaking Children's Speech Recognition: Developmental Changes in the Influence of Semantic context and F0 Contours. *Front. Psychol.*, 8, 1090.

Effects of Segments, Intonation and Rhythm on the Perception of L2 Accentedness and Comprehensibility



Chunsheng Yang, Jing Chu, Si Chen, and Yi Xu

Abstract This study examines the effects of segments, intonation and rhythm on the perception of second language (L2) accentedness and comprehensibility by focusing on a tone language, Mandarin Chinese. Fifteen Chinese sentences were manipulated by transferring the segments, intonation and rhythm between native and L2 speakers. 64 Chinese judges listened to the original and the manipulated sentences and were asked to rate the accentedness and comprehensibility of these sentences. Results of the Chinese native judges' ratings showed that segments contribute more to the perception of L2 accentedness and comprehensibility than intonation and rhythm, and that intonation contributed more to L2 perception than rhythm. It was also found that accentedness ratings highly correlated with comprehensibility judgment. The findings of this study confirm what some recent studies have found regarding the contribution of segments and prosody to L2 perception, but differ from some previous studies in regards to the relationship between L2 accentedness and comprehensibility. This study has both theoretical and pedagogical implications.

Keywords Segments · Intonation · Rhythm · Accentedness · Comprehensibility · Mandarin · L2 · Tones

C. Yang (✉)

Department of Literatures, Cultures and Languages, University of Connecticut, Storrs, CT 06269, USA

e-mail: chunsheng.yang@uconn.edu

J. Chu

College of Literature, Bohai University, Jinzhou, Liaoning, China

S. Chen

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China

Y. Xu

Division of Psychology and Language Sciences, University College London, London, UK

1 Introduction

Second language (L2) speech learning entails the learning of both segments and suprasegmentals or prosody (i.e., tones, intonation, rhythm, stress, prosodic phrasing). Due to phonetic and phonological difference in segmental and suprasegmental aspects between L2 and the learners' native language (referred to as L1), L2 learners may have difficulty in acquiring some L2 sounds. These difficulties have been accounted for in different theoretical frameworks, such as contrastive analysis hypothesis (CAH, Lado 1957), speech learning model (SLM, Flege 1995), perceptual assimilation model (PAM, Best 1995), and PAM-L2 (Best and Tyler 2007) and so on. More often than not the difficulties in L2 sounds render L2 speech different from L1 speech, hence the perception of "foreign accent" in L2 speech. While many studies have shed light on the difficulty in acquiring L2 sounds (Bradlow 1995; Iverson and Evans 2007, 2009; Flege et al. 1995; Elvin et al. 2014; Best 1995; Best and Tyler 2007, among many others), they also show that it is almost impossible to achieve native-like pronunciation without any foreign accent for most adult L2 learners (Flege et al. 1995).

Besides foreign accent/accentedness, intelligibility and comprehensibility are two important constructs in L2 pronunciation. Intelligibility refers to the extent to which the speaker's message is understood by the listener and is usually measured by asking listeners to orthographically transcribe what they hear (Kirkpatrick et al. 2008; Munro and Derwing 1999). Comprehensibility refers to the degree of easiness or difficulty in understanding an utterance. Similar to accentedness, comprehensibility is often measured by asking listeners to rate the degree of easiness/difficulty in understanding L2 speech on a scale. Thomson (2018) in his thorough review of previous studies on these three constructs showed that intelligibility, comprehensibility, and accentedness correspond to different linguistic features at various levels: intelligibility is closely related to phonemic (but not phonetic deviations) and word stress errors, comprehensibility is affected not only by phonology, but also by lexical and grammatical errors, and oral fluency, and accent is related to a variety of factors, such as phonetic deviations, syllable-level errors and non-native rhythm, lexical stress, slow speech rate and pausing, and reduced pitch range (see Thomson 2018 for more details). Due to the different correspondence between intelligibility, comprehensibility and accentedness, and linguistic features in speech, studies have shown inconsistent findings in terms of their relationship. For example, while Munro and Derwing (1995, 1997) and Derwing and Munro (2015: p. 5–6) found that strong accent does not necessarily impact intelligibility or comprehensibility, Julkowska and Cebrian (2015) found that accent and intelligibility are weakly correlated, but accent and comprehensibility are moderately or strongly correlated.

Considering the overarching goal of learning an L2 (i.e., to communicate in L2), the field of L2 pronunciation research has somewhat reached a consensus on the goal of L2 pronunciation teaching and learning, namely to make L2 speech as intelligible as possible, instead of free of foreign accent (Levis 2005; Munro and Derwing 2011; Thomson 2018; Levis 2018). It is worth noting that the setup of this goal in L2

pronunciation is more of a compromise when confronting so many tasks in learning L2, rather than the ideal goal (Thomson 2018). However, the often-used measurement of intelligibility is different from what people encounter in actual communication in that word-by-word transcription of an utterance (i.e., intelligibility) does not equal to understanding of the message in the utterance and people usually do not need to recognize every single word in an utterance in order to understand the message. For these reasons, some recent studies only examined comprehensibility and did not include intelligibility task (Trofimovich and Isaacs 2012; Kang 2010; Munro and Derwing 2001; Saito et al. 2016). In the same vein, we focus on comprehensibility as measured by listener judges' ratings in this study.

While intelligibility or comprehensibility is considered to the first priority of L2 pronunciation teaching, we argue that L2 foreign accent, "a deeply personal and inherently social phenomenon" (Levis and Moyer 2014), should not be dismissed as unimportant for several important reasons. First and foremost, accent, both L1 and L2, is likely to influence normal communication, at least among some speakers (Munro and Derwing 1995, 1997) cause negative perception of the speakers (Harrison 2014; Kang and Rubin 2014) and even impede one's career development (Lacey 2011, cited in Trofimovich and Isaacs 2012). Meanwhile, even if accent does not affect interpersonal communication, it may lead to social isolation and even psychological insecurity (Levis 2016). Chun (2002: p. 83–84) argues that if L2 speakers need to develop comprehensive language competence and proficiency and to produce "socially acceptable" speech, learning appropriate pronunciation registers, including accent, should not be considered "icing on the cake." For these reasons, accentedness is included in this study.

While previous studies (see Thomson 2018 for a review) have somewhat identified the linguistic factors affecting comprehensibility and accentedness in L2 speech, the effects of segments, intonation and rhythm on L2 perception are still unclear and even controversial among some studies. Therefore, this study investigates the effects of segments, intonation and rhythm on the perception of L2 accentedness and comprehensibility by focusing on a tone language, Mandarin Chinese. Since intonation and rhythm are operationalized differently in previous studies, it is necessary to define them first. By intonation, we refer to the overall F0 or pitch patterns of an utterance, instead of the pitch accents or peak alignment, and as a result, tones on each and every syllable are part of intonation in this study. By rhythm, we refer to the duration patterns of the syllables in a sentence and accordingly use duration and rhythm interchangeably throughout the paper, although they may be interpreted differently in other contexts.

2 Background

2.1 *Segments and Prosody in L2 Pronunciation Teaching and Research*

Historically, both segments (i.e., consonants and vowels) and prosody (intonation, stress, and other suprasegmental aspects) have been included in L2 pronunciation teaching. For example, in earlier textbooks and handbooks on English pronunciation teaching (Lado and Fries 1984; Nisen and Nisen 1971; Prator 1951, cited in Derwing and Munro 2015; Prator and Robinett 1984, cited in Derwing and Munro 2015), both segments and prosody were included. However, as correctly pointed out by Derwing and Munro (2015: p. 22–23), prosody may be marginalized and even not be touched upon in actual teaching practice.

In contrast to the marginalized role of prosody over segments in L2 pronunciation teaching, researchers have paid more attention to prosody than segments in their research (Avery and Ehrlich 1992; Chun 2002; Derwing et al. 2004; Hahn 2004; Isaacs 2008; Levis and Pickering 2004; Morley 1991; Munro and Derwing 2001; Kang 2010, among others). Of course, there is also a wealth of research on the acquisition of L2 consonants and vowels, such as Flege (1995) and Best (1995), and many studies mentioned therein. The focus on prosody in L2 pronunciation research seems to correlate with the relative more contribution of prosody to L2 accentedness and comprehensibility than segments in some studies. For example, Anderson-Hsieh, Johnson and Koehler (1992), Anderson-Hsieh and Koehler (1988), Holm (2009), Munro and Derwing (1999), and Quene and van Delft (2010) showed that prosody (intonation and duration) correlated more with accentedness and comprehensibility than segments. Other studies, however, revealed different findings. Jilka (2000) showed that segments contribute more to L2 accentedness than intonation, and Winter and O'Brien (2013) found that segments correlate more with accentedness than prosody, although prosody does affect L2 intelligibility. Furthermore, Sereno et al. (2016) showed segments contribute substantially to the perception of foreign accentedness than intonation, and native speakers rely mainly on segments when determining accentedness. It seems that further studies are required to further test the relative contribution of segments and prosody to L2 accentedness and comprehensibility and involve languages that have not been investigated previously, as recommended in Winter and O'Brien (2013) and Trofimovich and Isaacs (2012).

In terms of the contribution of intonation and duration to the perception of L2 accentedness and comprehensibility, Winter and O'Brien (2013) found that non-native duration and intonation cues contribute equally to perceived accentedness, even though non-native intonation patterns reduce intelligibility more than non-native duration cues do. Many studies only examined the effect of one or the other prosodic factor and seldom include both intonation and duration in one study. For example, Sereno et al. (2016) did not include duration. Quene and van Delft (2010) only examined the relationship of duration and intelligibility and it is difficult to tease apart the effects of intonation and duration in Holm (2009). Worth pointing out is

that Trofimovich and Baker (2006) found that while prosody overall contributes to foreign accent, some aspects, such as pause duration and speech rate, are more likely to do so than others, such as stress timing and peak alignment. However, we will leave it for future studies to differentiate the effects of the various aspects of prosody on L2 perception.

2.2 *Research Gaps*

As the discussion of previous studies shows, there are some research gaps. Firstly, although there are studies attempting to tease apart the contributions of segments and prosody (stress, pitch range, or peak alignment) to L2 accentedness or comprehensibility, not every study included segments, intonation and duration in their design, rendering the research findings difficult to compare and contrast. Secondly, previous studies have included various linguistic features, such as segmental/phonemic errors (Anderson-Hsieh et al. 1992; Andersen-Hsieh and Koehler 1988; Munro and Derwing 1995; Munro and Derwing 1997; Saito et al. 2016; Trofimovich and Baker 2006; Trofimovich and Isaacs 2010, among others), duration/temporal variable (Tajima, Port, and Dalby 1994; Winter and O'Brien 2013; Sereno et al. 2016; Quene and Van Delft 2010), speech rate including pause (Kang 2010; Trofimovich and Baker 2016; Saito et al. 2016; Derwing et al. 2004), syllable structure (errors) (Anderson-Hsieh and Koehler 1988; Anderson-Hsieh et al. 1992; Trofimovich and Isaacs 2012), word stress (errors) (Saito et al. 2016; Kang 2010; Trofimovich and Isaacs 2012), rhythm (vowel reduction ratio in Trofimovich and Isaacs 2012), and intonation goodness (Derwing and Munro 1997; Munro and Derwing 1995; Saito et al. 2016; Sereno et al. 2016; Winter and O'Brien 2013), as well as more accurate prosodic details, such as stress timing and peak alignment (Trofimovich and Baker 2016), and pitch range and pitch contour (Kang 2010; Trofimovich and Baker 2016). While it is good to pin down specific segmental and prosodic deviations which are related to L2 accentedness and comprehensibility ratings, it is almost impossible to exhaust these errors, that is to say, it is difficult to include all the possible linguistic errors/deviations. For this purpose, study design involving the manipulation of intonation or duration between two sentences through speech (re)synthesis, such as Jilka (2000), Winter and O'Brien (2013), and Sereno et al. (2016), may be able to avoid the exhaustivity issue. Thirdly, since previous studies have mainly focused on European languages, such as English, French, German, more typologically different languages, such as African or Asian languages (Yoruba or Mandarin Chinese), should be used to test the generalizability of the findings of previous studies. As Yang (2016) showed, the inclusion of such European language might unveil issues that are not readily available when only European languages are researched upon.

In our study, we attempt to tease apart the effects of segments and prosody (intonation and rhythm) on L2 accentedness and comprehensibility by focusing on Mandarin Chinese, a language that has not been involved in similar studies, and by carefully

manipulating the segmental and prosodic information between L1 and L2 utterances. The current study is guided by the following questions:

- (1) Do segments contribute to the perception of L2 accentedness and comprehensibility in the same way as prosody (rhythm and intonation)?
- (2) Do rhythm and intonation contribute to L2 accentedness and comprehensibility judgment equally?
- (3) How does L2 accentedness correlate with comprehensibility?

Worth noting is that some recent studies (Saito et al. 2016; Trofimovich and Isaacs 2012) have found that while accentedness is closely tied with phonology, including rhythm, segmental and syllable structure accuracy (Trofimovich and Isaacs 2012; Saito et al. 2016), comprehensibility is related to both phonology and grammatical accuracy and lexical richness (Trofimovich and Isaacs 2012; Saito et al. 2016). However, we only focus on the phonological factors in this study by controlling for the lexical and grammatical parameters.

2.3 Predictions

Based upon the discussion in previous sections, the following predictions are proposed.

- (1) Drawing upon the findings in Jilka (2000), Winter and O'Brien (2013) and Sereno et al. (2016), we predict that segments will contribute more to the perception of L2 accentedness and comprehensibility than intonation and rhythm.
- (2) Previous studies did not show a clear picture on the effects of rhythm and intonation on L2 accentedness and comprehensibility. However, considering the fact that Chinese is a tone language and both tones and intonation are represented by fundamental frequency (F0), we predict that intonation will contribute more to the perception of L2 accentedness and comprehensibility than rhythm.
- (3) Trofimovich and Isaacs (2012) found one linguistic feature, word stress in L2 English spoken by native French speakers, to be a common contributor to accentedness and comprehensibility ratings, which they attributed to the French and English typological difference, namely syllable-timing versus stress timing. Similarly due to the difference in tonal status between Mandarin Chinese and English, it is expected that tones may play a similar role to stress as in Trofimovich and Isaacs (2012). It is expected that tone errors in L2 Chinese, as included in intonation, affect both accentedness and comprehensibility in L2 Chinese. As a result, it is predicted that accentedness will highly correlated with comprehensibility in L2 Chinese.

3 Methodology

3.1 Material Preparation

Speakers

In order to examine the effects of segments, rhythm and intonation on L2 comprehensibility and accentedness, L2 Chinese learners who have strong accent should be recruited. For that purpose, three male intermediate-high/advanced-low American learners of Chinese were recruited at a mid-western public university in the US. Three of the co-authors who are all native speakers of Chinese agreed that the three L2 learners had a strong foreign accent in their L2 Chinese (a mean of 4.7, on a five point scale 1–5 in which 1 indicates little or no foreign accent and 5 indicates very strong accent). In addition, two male native speakers of Beijing Mandarin Chinese were recruited as the control group. Both Beijing Mandarin speakers were born and grew up in Beijing before coming to pursuing their master or doctoral degree in the US. Table 1 presents the demographic information of the five speakers.

Recording materials

To elicit speech as natural as possible, two short passages were used for recording, a fairy tale *The Sun and the North Wind*, and a short reading paragraph from Lesson 1 of the second-year Chinese textbook for L2 learners widely used in American colleges and universities.

Recording procedure

The recording took place in a recording studio at the mid-western university. The two reading passages were presented on the screen of a computer. The five speakers were told to read the passages as naturally as possible and in their normal speech rate. For the L2 speakers, they could ask for help if there were any characters if they did not recognize. Actually, none of the characters posed difficulty for the L2 speakers. The readings of the two passages were recorded with Audacity in a computer and

Table 1 Information of the five speakers

No	Native language	Age	Age of onset	Duration of Chinese learning (yrs)	Duration of study abroad (yrs)
1	English	24	20	4	0.5*
2	English	22	18	4	0.75
3	English	25	22	3	0.25
4	Mandarin	25	n/a	native	1
5	Mandarin	29	n/a	native	4.5

(*Duration of study abroad for L2 learners refer to the duration of study in abroad in China or Taiwan, whereas for native Mandarin speakers, it refers to the duration of study in the US)

then were saved as .wav files for further manipulations. The five speakers received \$10 dollars for their participation.

Manipulations

15 sentences were chosen from the two reading passages. Since the story of The Sun and the North Wind is familiar to some, the sentences which contain some words, such as “the sun” and “the north wind,” may imply to the listeners about the story and influence the comprehensibility judgment and therefore were not chosen. As for the other passage, one proper noun 王国明 Wang Guoming “a personal name” was repeated several times. In order to avoid the potential effect of such repeated nouns on the judgment of accentedness and comprehensibility, the proper noun was removed. Even so, the treated sentences still sounded natural for three of the co-authors. See the Appendix for the list of the 15 sentences used in this study.

In order to differentiate the effects of segments, rhythm and intonation on L2 perception, the 75 sentences (15 sentences * 5 speakers) were manipulated by transferring intonation and rhythm between the native and L2 speakers. Following Sereno et al. (2016), the choice of which native (two speakers) and L2 (three speakers) sentences were manipulated was random. However, efforts were made to ensure that the sentences by the three L2 speakers and the two native speakers were equally used in the manipulations. Altogether there were 120 sentences (the original and derive ones) used in this study: 15 target sentences * 2 groups (native vs. L2) * 4 versions (a. the original sentence; b. a derived sentence with different duration; c. a derived sentence with different intonation; and d. a derived sentence with different intonation and duration). For each target sentence, the following eight versions were obtained:

CsCiCr, CsCiEr, CsEiEr, CsEiCr.

EsEiEr, EsEiCr, EsCiCr, EsCiEr.

(C: Chinese; E: English; s: segment; i: intonation; r: rhythm. For example, EsEiCr means that this utterance has English segments (Es), English intonation (Ei) and Chinese rhythm (Cr)).

Below is the procedure of rhythm and intonation transfer.

Step 1: Rhythm transfer: The program PENTrainer (Xu and Prom-on 2014), a semi-automatic software package written as Praat scripts integrated with Java programs, was used in this step. Based on the Parallel Encoding and Target Approximation (PENTA) framework (Xu 2005), the quantitative Target Approximation (qTA) model (Prom-on et al. 2009), and the simulated annealing optimization (Kirkpatrick et al. 1983), PENTrainer can automatically learn the optimal parameters of all possible functional combinations that users have annotated and the learned parameters can be used to synthesize F0 contours according to any given communicative functions. Since Chinese is a monosyllabic language, the transfer of rhythm can be taken as the transfer of the syllable-by-syllable duration between two sentences. In order to transfer the rhythm of two sentences, the two sentences were first transcribed syllable by syllable by running PENTrainer in Praat. After obtaining the individual syllable duration data in both sentences, the duration data of one sentence were replaced with those of the other one manually on the syllable duration tier. Then

resyntheses were implemented for both sentences to derive a new utterance with the duration pattern of the other sentence in the pair.

Step 2: Intonation and rhythm transfer: The Praat Vocal Toolkit, a free plugin for Praat with automated scripts for voice processing, was used in this step (Corrette 2012). The toolkit was first installed in Praat. With the toolkit, it is easy to transfer the pitch contour from other utterance to another. Note that this step was based upon the previous one. That is to say, the intonation contour of one original sentence was transferred to the same derived sentence to which the rhythm of the original was transferred. After this step, both intonation and rhythm were transferred between the two sentences.

Step 3: Intonation only transfer: This step was based upon the previous two steps and the program PENTrainer (Xu and Prom-on 2014) was used again. To transfer intonation only, the duration of the various syllables in the derived sentence whose rhythm and intonation had been replaced by another sentence was restored to that of the original sentence, following the same procedures as in Step 1.

Following the manipulations, all sentences were amplitude normalized to 65 dB.

3.2 *Chinese Native Judges*

64 Chinese native judges were recruited at Bohai University, Liaoning, China. They were all undergraduate students majoring in Teaching Chinese as a Foreign Language and were in their early 20 s (mean: 19.5, SD = 0.7) at the time of the study. The gender ratio (F:M) was 3.57:1. All the Chinese native judges were native speakers of northern Mandarin. They participated in the study for course credits. All judges reported no speech or hearing problem.

3.3 *Procedure*

This study was conducted online (qualtrics.uconn.edu). In order to offset the possible effect of a particular order on the listeners' perception, eight randomized orders of the 120 utterances were created, with the comprehensibility and accentedness rating tasks counterbalanced. Every set of sentences was listened to and judged by eight Chinese native judges. The recruitment of participants and the running of the experiment were coordinated by one of the co-authors.

The online experiment consisted of three sections. The first section was to elicit the participants' demographic information. The second and third sections were to elicit the participants' ratings on the utterances' comprehensibility and accentedness. In the second and third sections, the participants needed to click to listen to the sentence only once and then rate the degree of comprehensibility (namely, how easy

to understand the utterance, the higher the rating, the more easily to comprehend) and accentedness (namely, how foreign does the utterance sound? the higher the rating, the more foreign) on 1–5 Likert scale.

4 Results

In this section, we begin with the descriptive statistics of the comprehensibility and accentedness judgements, followed by the inferential statistics.

Table 2 presents the means and standard deviations of the comprehensibility and accentedness ratings by the Chinese native judges. As can be seen from the table, the mean comprehensibility ratings for sentences containing Chinese segments (Cs) are overall higher than those with English segments (Es), except for the all-English sentences (EsEiEr), whereas the sentences containing English segments do not differ dramatically, regardless of intonation and rhythm. For the ratings on the accentedness, the all-Chinese sentences have the lowest accentedness ratings, and as the components of English increase, the accentedness ratings start to increase. It seems that, while the ratings on comprehensibility and accentedness are related to each other, the English components seem to have greater impact on accentedness ratings than comprehensibility ratings.

Comprehensibility

Considering that the ratings of comprehensibility are ordinal, not continuous, we fitted the cumulative link mixed model (CLMM) with the R package “ordinal” (R core team 2014). Figure 1 plots the listener judge effects for comprehensibility across subjects. As shown in Fig. 1, the 43rd subject gave the lowest ratings of comprehensibility while the 44th judge gave the highest. The judge effect indicates that subjects have different standard for comprehensibility. Therefore, we modeled subjects as random effects.

Table 2 Descriptive statistics of the comprehensibility and accentedness judgment

	Comprehensibility		Accentedness	
	Mean	Standard deviation	Mean	Standard deviation
CsCiCr	4.59	0.68	1.63	0.95
CsCiEr	4.31	0.77	2.08	1.20
CsEiEr	3.84	0.85	2.88	1.06
CsEiCr	4.11	0.83	2.51	1.23
EsCiCr	3.78	0.93	3.01	1.06
EsCiEr	3.72	0.93	3.29	1.09
EsEiCr	3.79	0.91	3.21	1.05
EsEiEr	3.84	0.92	3.63	0.96

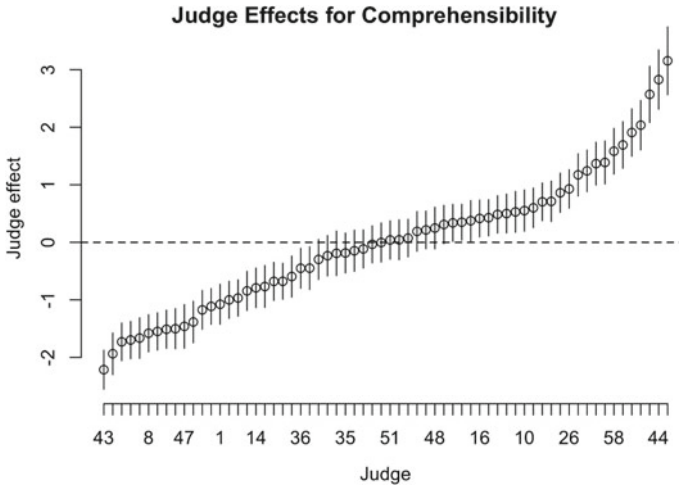


Fig. 1 Judge effects for comprehensibility by different subjects

The CLMM coefficients for English segments, intonation and rhythm (-1.146 , -0.611 , -0.431) were negative, which indicated that including more English components in the stimuli would decrease comprehensibility. Odds ratios, a measure of association between an exposure and an outcome, namely the odds that an outcome will occur given a particular exposure as compared to the odds of the outcome occurring in the absence of that exposure (Szumilas 2010), were also calculated for segments, intonation and rhythm. The results showed that segments were related to the lowest odds ratio (0.3178513), followed by intonation (0.543) and rhythm (0.65), showing their different effects on comprehensibility. That is to say, segments had the most dramatic effects on comprehensibility, and the more English segments there are, the lower the odds of comprehensibility. Using likelihood ratio tests, the effects from segment, intonation and rhythm were all significant on comprehensibility rating ($\chi^2(1) = 618.77$, $p < 0.001$; $\chi^2(1) = 181.51$, $p < 0.001$; $\chi^2(1) = 90.76$, $p < 0.001$).

We also computed the probabilities of comprehensibility rating for average, 5th and 95th percentile judges at the eight experimental conditions (different combinations of stimuli). Figure 2 plots these probabilities. In Fig. 2, the solid line represents average judges, whereas the dashed and dotted lines stand for the 5th and 95th percentile judges. From Fig. 2, it can be seen that, if all the components were Chinese (for the panel where segment = C, intonation = C, rhythm = C), it was very likely to receive a rating of 5 (probability ≈ 0.6) for average judges. Bringing in English rhythm in the stimuli (for the panel where segment = C, intonation = C, rhythm = E) decreases the probability of a rating of 5 to around 0.5 and bringing in English intonation decreases the probability to around 0.4 (for the panel where segment = C, intonation = E, rhythm = C).

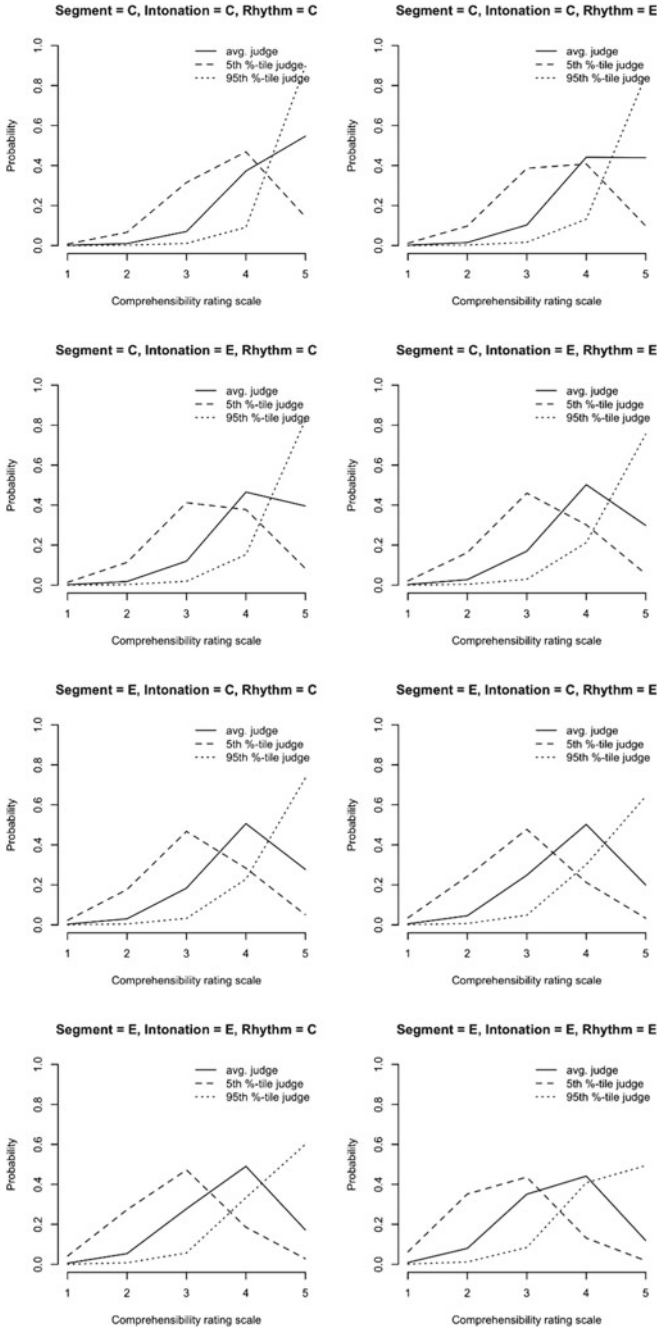


Fig. 2 Comprehensibility rating probabilities for average and extreme judges of different stimuli

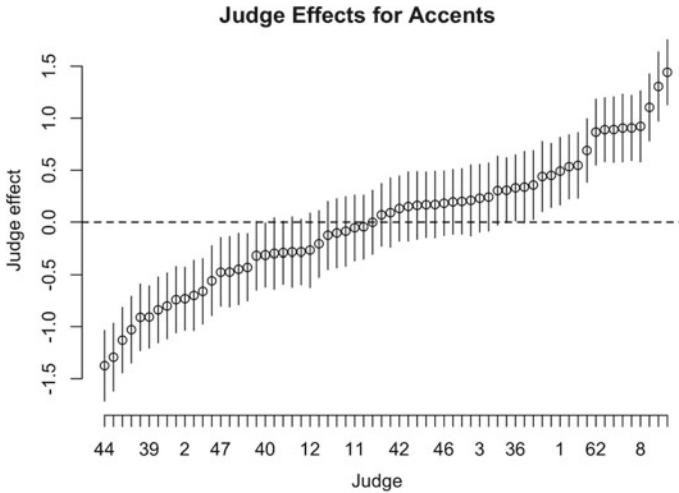


Fig. 3 Judge effects for accents ratings by different subjects

Accentedness

We fitted the cumulative link mixed model (CLMM) with the R package “ordinal” to examine the relationship between the three factors segments, intonation and rhythm and foreign accents ratings. As shown in Fig. 3, the 44th subject gave the lowest ratings of accents while the 8th judge gave the highest. The judge effect indicated that subjects had different standard for comprehensibility. Therefore, we also modeled subjects as random effects in our model.

The coefficients for English segments, intonation and rhythm (1.82613, 1.03224, 0.69031) were positive, indicating that more English components increased perception of accents. Odds ratios were also calculated. The odds ratios indicated that more English components were related with higher odds of perceived accents and that segments were related with the highest odds (6.209808), followed by intonation (2.807338) and then rhythm (1.994337). Using the likelihood ratio tests, segment, intonation and rhythm were all significant in accentedness rating ($\chi^2(1) = 1678.46, p < 0.001; \chi^2(1) = 575.31, p < 0.001; \chi^2(1) = 260.85, p < 0.001$).

We also computed the probabilities of accent rating for average, 5th and 95th percentile judges at the eight experimental conditions (different combinations of stimuli). Figure 4 plots these probabilities. In Fig. 4, the solid line represents average judges, whereas the dashed and dotted lines stand for 5th and 95th percentile judges. From Fig. 4, it can be seen that, if all the components were English (for the panel where segment = E, intonation = E, rhythm = E), it was likely to receive a rating of 4 or 5 (probability ≈ 0.4 and 0.3 respectively), namely strong accent. Bringing in Chinese intonation (for the panel where segment = E, intonation = C, rhythm = E) decreases the probability of the accent rating of 4 and 5 to around 0.3 and 0.1 , respectively, and bringing Chinese segments decreases the probability of the accent

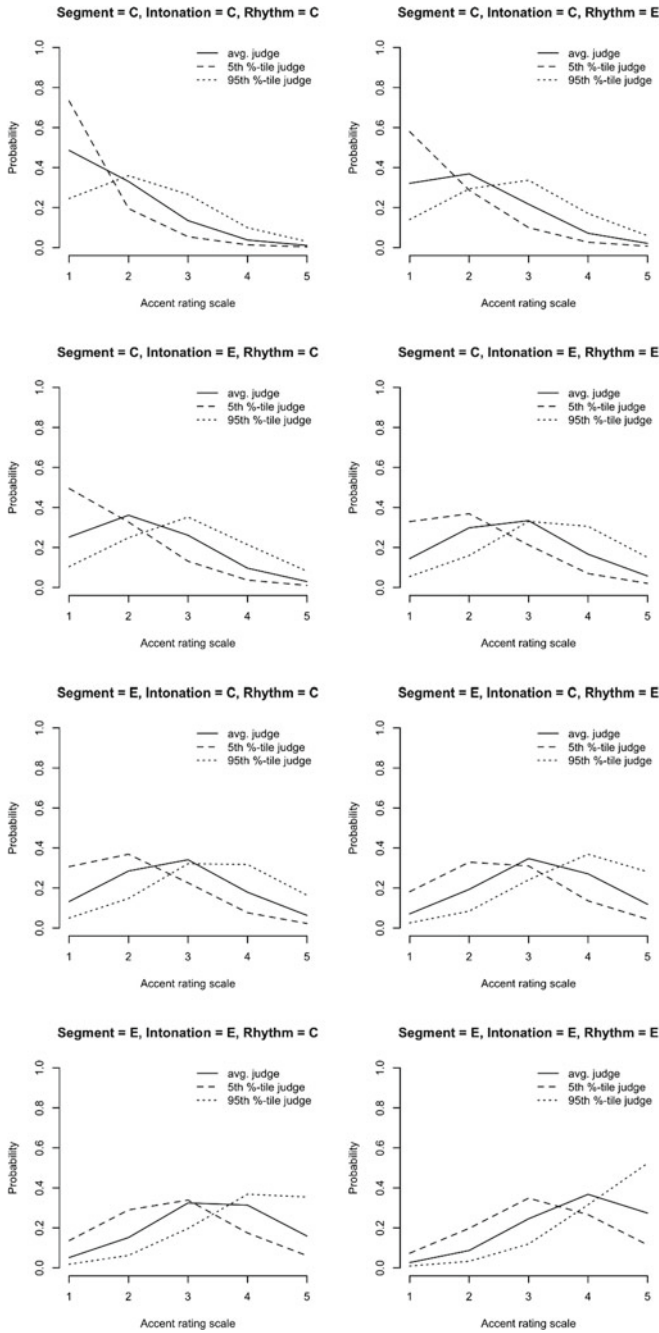


Fig. 4 Accentedness rating probabilities for average and extreme judges of different stimuli

rating of 4 and 5 even lower (0.2 and 0.1 respectively), showing the importance of segments in accent rating.

Relationship between Comprehensibility and Accentedness

To examine the relationship between comprehensibility and accentedness, we first fitted the cumulative link mixed model (CLMM) with subjects as a random effect and accentedness rating as a predictor to predict comprehensibility. The coefficient is negative (-0.6421), indicating that higher accentedness rating leads to decrement in comprehensibility. The odds ratio (0.4258086) also indicates that higher accentedness rating is related with lower odds of comprehensibility. Using a likelihood ratio test, the factor comprehensibility is significant ($\chi^2(1) = 1038.9$, $p < 0.001$).

Then we fitted the cumulative link mixed model (CLMM) with subjects as a random effect and comprehensibility as a predictor to predict accents. The coefficient is negative (-0.85377), indicating that higher comprehensibility rating leads to decrement in accent rating. The odds ratio (0.5261859) also indicates that higher comprehensibility rating is related with lower odds of accentedness. Using a likelihood ratio test, the factor comprehensibility is significant ($\chi^2(1) = 983.5$, $p < 0.001$).

5 Summary and Discussions

5.1 Summary

In this study, we meticulously transferred rhythm and intonation between sentences produced by Chinese native speakers and L2 speakers in order to tease apart the effects of segments, rhythm and intonation on the perception of L2 accentedness and comprehensibility. The ratings of the Chinese native judges showed that they relied more upon segments in their comprehensibility and accentedness judgment than upon intonation and rhythm, which confirms our first prediction. This finding accords with the findings of some recent studies on similar topics (Jilka 2000; Sereno et al. 2016; Winter and O'Brien 2014). As for the effect of intonation and rhythm on L2 perception, our finding showed that intonation contributed more to L2 comprehensibility and accentedness than rhythm, confirming our second prediction. However, it should be noted that both intonation and rhythm had much lower odd ratios, as compared to segments, in the perception of accentedness and comprehensibility. The analysis of the relationship between comprehensibility and accentedness showed that accentedness highly correlates with comprehensibility in L2 Chinese. That is to say, as the accentedness of L2 speech increases, the comprehensibility drops and vice versa. The relationship between comprehensibility and accentedness supports the finding in Yang (2016), but differs from Derwing and Munro (1997) and Munro and Derwing (1998). Thus, our third prediction was also borne out.

5.2 *Discussions*

5.2.1 **Roles of Segments and Prosody on L2 Perception**

Previous studies have revealed indefinite conclusions on the role of segments and prosody (i.e., intonation and rhythm) on the perception of L2 accentedness and comprehensibility (Anderson-Hsieh and Koehler 1988; Anderson-Hsieh et al. 1992; Chun 2002; Hahn 2004; Holm 2009; Levis and Picking 2004; Morely 1991; Munro and Derwing 1999; Munro and Derwing 2001; Kang 2010; Quene and van Delft 2010; Sereno et al. 2016; Trofimovich and Isaacs 2012; Winter and O'Brien 2013, among many others). There are two factors that may explain the indefinite roles of segments and prosody. On the one hand, as Sereno et al. (2016) pointed out, some studies compared the perception of the natural speech with that of the resynthesized speech, which may be problematic. Secondly, while some studies focus on the roles of different aspects of prosody on L2 perception (Chun 2002; Kang 2010), the non-inclusion of segments does not mean that segments play a lesser role than prosody. The findings in this study and some other studies (Jilka 2000 and Sereno et al. 2016), which involved intonation and rhythm manipulations, however, seem to highlight the more significant role of segments in L2 perception. However, the importance of segments over intonation and rhythm in L2 perception does not seem to be something surprising. In non-tone languages, such as English or German, segments play a relatively more important role than intonation and rhythm in that segmental deviations may lead to misinterpretation of the message being conveyed. For example, if someone produces “think” as “sink,” it surely will lead to the perception of strong foreign accent and cause miscommunication on English native speakers, unless the syntax or the context favors the semantic meaning of “think.” In the same vein, stress in English can differentiate lexical meanings as well, such as “PROduce” versus “pro-DUCE” (capitalization indicates lexical stress), and the wrong placement of lexical stress can cause difficulty or misleading, although it should be noted that there are not many such minimal pairs of lexical stress in English and the wrong place of lexical stress do not necessarily influence L2 comprehension. When it comes to tone languages, such as Chinese, tones are as important as segments because tones can differentiate lexical meanings, in the same way as segments. Interestingly, Patel et al. (2010) found that intelligibility, not comprehensibility though, of Mandarin Chinese sentences with natural F0 contours, was comparable to the intelligibility of the monotone (flat-F0) sentences created via speech resynthesis for Chinese native speakers in ideal listening environment (i.e., no noise or little noise), that is to say, flat tones or tonal deviations do not influence the intelligibility of the speech by native speakers. Does it mean that tone deviations do not matter? On the one hand, the non-difference between the intelligibility of the sentences with natural intonation and those with flat F0 may be due to the familiarity of the topics in those sentences. Further study needs to examine the effect of the familiarity of topics on the intelligibility of flat-tone sentences. On the other hand, as Patel et al. (2010) showed, when noise was added, the intelligibility of the monotone sentences worsens. Since most normal

communication takes places with some background noise, the intelligibility and comprehensibility of the flat-tone sentences is likely to pose difficulty for listeners. For L2 listeners, they may not be familiar with the topics involved and, more importantly, the co-articulations of tones, and their listening ability is adversely affected by the environment (Bradlow and Bent 2002; Cooke et al. 2008; Cutler et al. 2008; Zhou et al. 2017; Zhang et al. 2016). Therefore, it is expected that the flat-intonation sentence will very likely pose even greater challenge for L2 listeners. Furthermore, while previous studies on L2 English showed that L2 judges rate the accentedness and comprehensibility similarly to native speaker judges (Derwing and Munro 2013; Flege 1988; MacKay et al. 2006), similar studies on L2 Chinese speakers may reveal different findings, again due to the role of tones, in that native speakers are more tolerant of tone deviations and more used to tonal co-articulations in context due to L1 experience, whereas L2 learners, especially those who learn Chinese in a foreign context, may not have developed such capabilities. There are many anecdotes of misunderstanding caused by L2 tonal deviations among Chinese practitioners (e.g., *shuǐjiào* “to sleep” was produced as *shuǐjiǎo* “dumpling,” or *sǎobǎn* “boss” was produced as *sǎobàn* “old partner or spouse”).

Other than the role of segments, the relatively more important role of intonation than rhythm to the ratings of accentedness and comprehensibility in L2 Chinese confirmed the important role of tones in Chinese. Tones and intonation are represented by the same acoustic parameter, F0, and when transferring intonation from one sentence to another, tones were transferred as well. In previous studies, only Winter and O’Brien (2013) attempted to separate the effects of intonation and duration in the perception of L2 English and German, with different effects of intonation and duration found. Winter and O’Brien’s findings were partially confirmed in this study, namely intonation contributes to intelligibility/comprehensibility more than duration (they found that intonation and duration contributed to accentedness equally though). On the other hand, this relative more important role of intonation as opposed to duration shows that the F0 pattern in Chinese, such as the various types of intonation (i.e., statement versus question intonation, focus, tone co-articulation), is very important and as a result should be incorporated in Chinese language teaching (Yang 2019).

5.2.2 Implications for L2 Pronunciation Teaching and Research

The consensus on the goal of L2 pronunciation, namely, the intelligibility principle (Levis 2005; Munro and Derwing 2011), suggests that the most important construct in L2 pronunciation teaching is L2 intelligibility or comprehensibility and that L2 accent does not matter that much as long as it does not affect L2 comprehension. The finding of this study that accentedness highly correlates with comprehensibility seems to suggest that accent is important, at least in such a tone language as Chinese. Note that the correlation between accentedness and comprehensibility does not mean that accentedness leads to comprehensibility difficulty or the other way around. Considering the different findings on L2 accent in this study and some

studies on L2 English (Derwing and Munro 1997; Munro and Derwing 1998, and so on), we speculate that the accent deriving from tone deviations may cause or lead to comprehensibility difficulty, due to the mediating role of tones (Yang 2016). If this speculation turns out to be correct, the role of accent in L2 pronunciation, at least in tone languages, should be re-examined, instead of being dismissed as something secondary or unimportant.

The findings of this study also have important pedagogical implications. For one thing, the more important role of segments than intonation and rhythm suggests that, although tones are important, segmental accuracy is of great importance in L2 Chinese teaching. For another, the different contributions of intonation and rhythm to L2 Chinese perception have important implications for setting up the agenda of teaching L2 Chinese pronunciation. As mentioned above, further study should examine the role of intonation and rhythm in non-tone languages, as well as other tone languages, to see whether the findings in this study are applicable to other tone languages or to L2 in general in order to set up the teaching agenda or priority of L2 pronunciation across languages. If the relative weightings of intonation and rhythm are borne out across languages, it will be reasonable to prioritize intonation over rhythm in L2 pronunciation teaching research. For the teaching of Chinese as a second language, at least, this study has shown that intonation is more important than rhythm, although there is clearly rhythmic difference between L1 and L2 Chinese and, therefore, should be incorporated in the teaching practice.

5.2.3 Limitations and Directions for Further Studies

This study has several limitations. For example, while the stimulus manipulations in this study helped tease apart the effects of segments and prosody on the perception of L2 comprehensibility and accentedness, these manipulations may create one issue. In order to transfer the intonation from one utterance to another, the duration of the two utterances should be the same, namely to transfer duration first. After the intonation was transferred, the duration of the utterance concerned was changed back to its original syllable-by-syllable duration. However, in so doing, the intonation would be changed slightly. Figure 5 shows the pitch contour of one utterance produced by a Chinese native speaker and the pitch contour of the same utterance produced by an L2 speaker but with intonation transferred from the Chinese native speaker. As seen in the two panels of Fig. 5, the two utterances have different rhythm (or duration); the pitch contours of them are roughly the same although there are minor differences. In the upper panel, there is a portion of missing pitch contour (the syllable 早 zao3 “early”), due to the creaky voice of the tone of this syllable, a low tone, in this Chinese speaker. When transferring the intonation from the Chinese native speaker to the L2 speaker, such phonation does not transfer. As a result, there might be some minor difference in the intonation transferred to the L2 speaker from that in the original utterance by the native speaker.

The findings of this study point to some new directions of further studies in order to capture a more thorough and accurate picture of the effects of segments

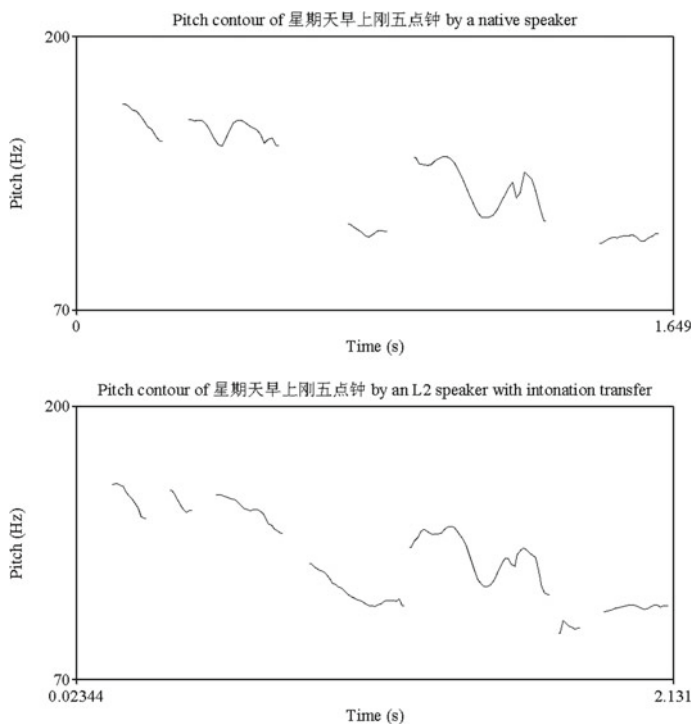


Fig. 5 Pitch contours of one utterance by a native speaker and the same utterance by an L2 speaker but with intonation transferred

and prosody on L2 comprehensibility and accentedness. Firstly, while segments contributes more to L2 perception than prosody, it is still unknown in L2 Chinese what segments are more likely to compromise L2 comprehensibility, namely the error gravity of Chinese segments (Derwing and Munro 2015: p. 74). Therefore, further studies should examine the error gravity of different segments (i.e., segments and vowels) and then prioritize the teaching of those consonants and vowels with greater error gravity. Secondly, studies along the same line focusing on other typologically different languages, both tone and non-tone languages, should be conducted to differentiate the roles of intonation and rhythm on L2 perception. Thirdly, the role of tones on comprehensibility in Chinese and other tone languages should be investigated, by utilizing sentences with unfamiliar topics or even something entirely nonsensical to examine whether native and L2 listeners can understand and transcribe the sentences with flat F0. The findings of such studies may shed new lights upon the differences of language processing by native and L2 speakers.

6 Conclusions

This study aims to contribute to the debate on the effects of segments and prosody on L2 comprehensibility and accentedness by focusing on a tone language, Mandarin Chinese. The findings of the study showed that segments contribute more to L2 perception than prosody, both intonation and rhythm, and that intonation contributed more than rhythm. Meanwhile, comprehensibility was found to highly correlate with accentedness in L2 Chinese. While accent may not be as important as comprehensibility in L2 speech, as shown in the consensus of the goal of L2 pronunciation teaching, accent in L2, at least in a tone language, such as Chinese, should not be dismissed as unimportant, because L2 accent does affect comprehensibility adversely, due to the mediating role of tones. The findings of this study have important theoretical and pedagogical implications for L2 pronunciation. While further studies are required to test whether the findings of this study are applicable to other languages, the findings of this study highlight the importance of extending L2 pronunciation studies to non-European and/or tone languages in order to test the generalizability of the findings in studies that focus on European languages only.

Appendix: Sentences Used

星期天早上刚五点钟 (xingqi tian zaoshang gang wudian zhong).

家里人都已经起来了 (jiali ren dou yijing qilai le).

要坐早上八点钟的火车到北京去 (yao zuo zaoshang badian zhong de huochedao Beijing qu).

王先生帮着小王收拾行李 (wang xiansheng bangzhe xiaowang shoushi xingli).

王太太特别给王国明做了很多吃的东西 (wang taitai tebie gei wang guoming zuole henduo chi de dongxi).

这两天天气热 (zhe liangtian tianqi re).

车上的东西恐怕不干净 (cheshang de dongxi kongpa bu ganjing).

他们把行李收拾好了的时候 (tamen ba xingli shoushi hao le de shihou).

已经七点钟了 (yijing qidian zhong le).

别的同学都在那儿等着他呢 (bie de tongxue dou zai nar dengzhe ta ne).

身上穿了一件大衣 (shenshang chuanle yijian dayi).

他们俩就商量好了 (tamen liang jiu shangliang hao le).

太阳很快地发出他所有的热量 (taiyang henkuai de fachu ta suoyou de reliang).

热得受不了了 (rede shou buliao le).

便将衣服一件件脱下 (bian jiang yifu yijianjian tuoxia).

References

- Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, 42, 529–555.
- Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning*, 38, 561–613.
- Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). Baltimore, MD: York Press.
- Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). Amsterdam: John Benjamins.
- Bradlow, A. (1995). A comparative acoustic study of English and Spanish vowels. *Journal of the Acoustical Society of America*, 97(3), 1916–1924.
- Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America*, 112(1), 272–284.
- Chun, D. M. (2002). Discourse in intonation in L2: From theory and research to practice. John Benjamins Publishing Company.
- Cooke, M., Lecumberri, M. L. G., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America*, 123(1), 414–427.
- Corrette, R. (2012). Praat Vocal Toolkit. <https://www.praatvocaltoolkit.com>.
- Cutler, A., Lecumberri, M. L. G., & Cooke, M. (2008). Consonant identification in noise by native and non-native listeners: Effects of local context. *Journal of the Acoustical Society of America*, 124(2), 1264–1268.
- Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 20, 1–16.
- Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, 63, 163–185.
- Derwing, T.M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins Publishing Company.
- Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thompson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, 54, 655–679.
- Elvin, J., Escudero, P., & Vasiliev, P. (2014). Spanish is better than English for discriminating Portuguese vowels: Acoustic similarity versus vowel inventory size. *Frontiers in Psychology*, 5, 1188.
- Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *The Journal of the Acoustical Society of America*, 84, 70–79.
- Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistics experience: issues in cross-language research* (pp. 233–272). Baltimore, MD: York Press.
- Flege, J. E., Munro, M. J., & MacKay, I. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, 97, 3125–3134.
- Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38(2), 201–223.
- Harrison, G. (2014). Accent and “othering” in the workplace. In J. Levis, & A. Moyer, (Eds.), *Social dynamics in second language accent* (pp. 255–272). De Gruyter Mouton.
- Holm, S. (2009). Intonational and durational contributions to the perception of foreign-accented Norwegian. Unpublished doctoral dissertation, Norwegian University of Science and Technology.
- Kostin, I. (2004). *Exploring item characteristics that are related to the difficulty of TOEFL dialogue items*. (TOEFL Research Report RR-79). Princeton, NJ: Educational Testing Service.

- Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *The Canadian Modern Language Review*, 64, 555–580.
- Iverson, P., & Evans, B. G. (2007). Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration. *Journal of Acoustical Society of America*, 122, 2842–2854.
- Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *Journal of Acoustical Society of America*, 126, 866–877.
- Jilka, M. (2000). *The contribution of intonation to the perception of foreign accent*. Unpublished doctoral dissertation, University of Stuttgart.
- Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, 38, 301–315.
- Kang, O., & Rubin, D. (2014). Listener expectations, reverse linguistic stereotyping, and individual background factors in social judgments and oral performance assessment. In J. Levis, & A. Moyer, (Eds.), *Social Dynamics in Second Language Accent* (pp. 239–253). De Gruyter Mouton.
- Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671–680.
- Kirkpatrick, A., Deterding, D., & Wong, J. (2008). The international intelligibility of Hong Kong English. *World Englishes*, 27(3–4), 359–377.
- Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. Ann Arbor: University of Michigan Press.
- Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3), 369–377.
- Levis, J. (2016). Accent in second language pronunciation research and teaching. *Journal of Second Language Pronunciation*, 2(2), 153–159.
- Levis, J. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press.
- Levis, J., & Moyer, A. (Eds.). (2014). *Social dynamics in second language accent*. Boston and Berlin: Mouton de Gruyter.
- Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System*, 32, 505–524.
- MacKay, I. R., Flege, J. E., & Imai, S. (2006). Evaluating the effects of chronological age and sentence duration on degree of perceived foreign accent. *Applied Psycholinguistics*, 27, 157–183.
- Morley, J. (1991). The pronunciation component of teaching English to speakers of other languages. *TESOL Quarterly*, 25, 481–520.
- Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, 38, 289–306.
- Munro, M. J., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning*, 48, 159–182.
- Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 49, 285–310.
- Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: the role of speaking rate. *Studies in Second Language Acquisition*, 23(4), 451–468.
- Munro, J. M., & Derwing, T. M. (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching*, 44(3), 316–327.
- Patel, A. D., Xu, Y., & Wang, B. (2010). *The role of F0 variation in the intelligibility of Mandarin sentences*. Speech Prosody 2010, Chicago.
- Prom-on, S., Xu, Y., & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America*, 125, 405–424.
- Quene, H., & Van Delft, L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Communication*, 52, 911–918.

- R Core Team. (2014). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, 37, 217–240.
- Sereno, J., Lammers, L., & Jongman, A. (2016). The relative contribution of segments and intonation to the perception of foreign-accented speech. *Applied Psycholinguistics*, 37, 303–322.
- Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 19(3), 227–229.
- Tajima, K., Port, R., & Dalby, J. (1994). Influence of timing on intelligibility of foreign-accented English. *The Journal of the Acoustical Society of America*, 95(5), 3009.
- Thomson, R. (2018). Measurement of accentedness, intelligibility, and comprehensibility. In O. Kang, & A. Ginther, (Eds.), *Assessment in Second Language Pronunciation* (pp. 11–29). Routledge.
- Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, 28, 1–30.
- Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, 15(4), 905–916.
- Willems, N. (1982). *English intonation from a Dutch point of view*. Dordrecht: Foris.
- Winters, S., & O'Brien, M. G. (2013). Perceived accentedness and intelligibility: The relative contributions of F0 and duration. *Speech Communication*, 55, 486–507.
- Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, 46, 220–251.
- Xu, Y., & Prom-on, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication*, 57, 181–208.
- Yang, C. (2016). *The Acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice*. John Benjamins Publishing Company.
- Yang, C. (2019). Teaching Chinese intonation and rhythm. In C. Shei, M. E. M. Zikpi, & D. Chao (Eds.), *The routledge handbook of Chinese language teaching* (pp. 180–194). Abingdon and New York: Routledge.
- Zhang, L., Li, Y., Wu, H., Li, X., Shu, H., Zhang, Y., & Yi, P. (2016). Effects of semantic context and fundamental frequency contours on Mandarin speech recognition by second language learners. *Frontiers in Psychology*, 7, 908.
- Zhou, H., Li, Y., Liang, M., Guan, C. Q., Zhang, L., Shu, H., & Zhang, Y. (2017). Mandarin-speaking Children's speech recognition: Developmental changes in the influences of semantic context and F0 contours. *Frontiers in Psychology*, 8, 1090.

Foreign Accent in Second Language Mandarin Chinese



Eric Pelzl

Abstract This chapter discusses second language pronunciation of Mandarin from the perspective of the native Mandarin speakers who listen to it. For such listeners, second language Mandarin often bears a noticeable foreign accent. I will provide a framework for defining foreign accent and for distinguishing accented pronunciation from pronunciation errors. I will then review the results of research related to foreign-accented Mandarin and how it affects listeners' judgments, comprehension, and the efficiency with which they process second language Mandarin speech. Naturally, lexical tones will receive special attention in this discussion.

Keywords Mandarin · Second language pronunciation · Foreign accent · Pronunciation error · Tones

1 Introduction

Mandarin Chinese (*Pǔtōnghuà*) speakers often use the phrase *yáng qiāng yáng diào* (洋腔洋调) to describe the speech produced by second language (L2) Mandarin speakers. Ignoring, for the moment, that this phrase may come with some social baggage, its existence shows that native Chinese listeners hear something different in L2 or foreign-accented speech. Even though listeners are familiar with Cantonese, Shanghaiese, Taiwanese, and other native Mandarin accents, in some perceptible way, foreign-accented speech is different. I know what you're thinking—*it's the tones!* That may be correct, but in this chapter, we aren't going to rush to any conclusions. We will take our time considering the many ways that foreign accent might be apparent in L2 Mandarin speech and how this impacts listeners.

We will start by defining some of the important qualities of L2 pronunciation, but overall our focus will be on how foreign-accented speech affects native Chinese interlocutors—the common conversation partners of L2 speakers. By approaching L2 pronunciation from the perspective of listeners, we can gain insight into which

E. Pelzl (✉)

The Pennsylvania State University, University Park, PA, USA

e-mail: ezp218@psu.edu

© Springer Nature Singapore Pte Ltd. 2021

C. Yang (ed.), *The Acquisition of Chinese as a Second Language Pronunciation*, Prosody, Phonology and Phonetics, https://doi.org/10.1007/978-981-15-3809-4_12

257

aspects of pronunciation ought to be prioritized in learning and teaching. In this chapter, I won't attempt to explain *why* L2 accent happens, but interested readers can refer to key theoretical studies considering L2 pronunciation (e.g., Best and Tyler 2007; Escudero and Boersma 2004; Flege 1995; Major 2001).

As a review of research, this one comes with a big caveat—there is not much to review that is specific to L2 Mandarin. Research on accented speech perception and comprehension is only just beginning and, by my count there, are only five existing studies that have directly addressed foreign-accented Mandarin (not including other studies that might appear in this volume). I will review them all in some detail, focusing on ways that we might build on them for future research, but I will also draw heavily on other lines of research on native (L1) and L2 Mandarin speech. I will also draw connections to the much more extensive work that has examined foreign-accented speech in other languages, especially English.

One last note before we get going in earnest. Describing the people who speak with accents in this research is not always straightforward. In places where Mandarin is recognized as an official language, there are many who identify as minority language speakers. For them, Mandarin is also a second language that they may only ever master imperfectly. Additionally, given the diversity of Chinese regional languages (Norman 1988), many who identify as L1 Mandarin speakers, also produce the language with “non-standard” accents and, technically, might be called L2 speakers. For the purposes of this chapter, the L2 speakers we will be thinking about are primarily those who have lived most of their lives outside of Chinese language communities and have learned Mandarin largely as adults. It is this type of learner that we know is very likely to speak with a noticeable *foreign* accent (Flege et al. 1995). Differences among L2 speakers' native language backgrounds will certainly lead to different qualities of foreign accent. However, the few currently existing studies on foreign-accented Mandarin include L2 participants from a mix of L1 backgrounds, so we will not narrow in on any specific L1 in this review.

We begin our discussion with an attempt to more clearly define foreign accent.

2 What is Foreign-Accented Mandarin Like?

Everyone who speaks a language has a sense of what is typical and atypical in the pronunciation of their language. This sensitivity reflects their broad experience of the language. For instance, they may notice that their local speech community sounds somewhat different from that in another area, and perhaps none of these local speech varieties sound like the “standard” TV news anchor. Still, all of these groups are recognized as native speakers and their different pronunciations are within the realm of what is typical. Very loosely then, a foreign accent is pronunciation that is outside of the typical range, not just of the local speech community, but of the broader community recognized as native speakers of the language.

There are many ways pronunciation can differ across accents. Pronunciation that directly affects words in Mandarin includes segments (vowels and consonants) and

suprasegments (tones and perhaps stress). Other aspects of accent create impressions across phrases or longer stretches of speech. These include the rhythm, intonation, speech rate, and pauses that speakers produce. For the moment, we will focus specifically on segmental and tonal speech sounds.

2.1 The Speech Sound Distributions of a Language

When we think about vowels, consonants, and tones, we usually have a specific list—or inventory—of sounds in mind. This inventory includes all the sound categories that make up our words and sentences. Although we can give these categories labels (for example, the /m/ and /a/ sounds in *ma*), the truth is that whenever we produce one of these sounds, it’s never exactly the same as the time before. This is true for a single speaker and is certainly the case across speakers. Our different body shapes and sizes and our different linguistic experiences all lead to large variability in the sounds we produce. Although we recognize patterns in the pronunciation of our language, there is actually great variability under the surface.

This is illustrated in Fig. 1, where we can picture each individual utterance of a sound as a single point in space. The dimensions of that space (*x* and *y*) will be measurable physical properties of the sound. For example, its duration and fundamental frequency (F0, which we perceive as pitch), or vowel formants (F1, F2, F3—the energy of vibrations in the air within certain frequency ranges). If we measured many instances of the same sound being uttered, we could form a distribution for that sound category (i.e., what listeners perceive as being the same sound). This distribution will look rather circular, with the most typical instances of the sound accumulating at the center of the shape, and less typical instances spreading out toward or beyond the edges. With enough instances and enough different speakers, our circular shape will be a reasonable representation of the typical values of that speech sound.

Fig. 1 Visualization of speech sound distributions. The *x* and *y*-axis represent two separate acoustic measurements such as two vowel formants or pitch and duration

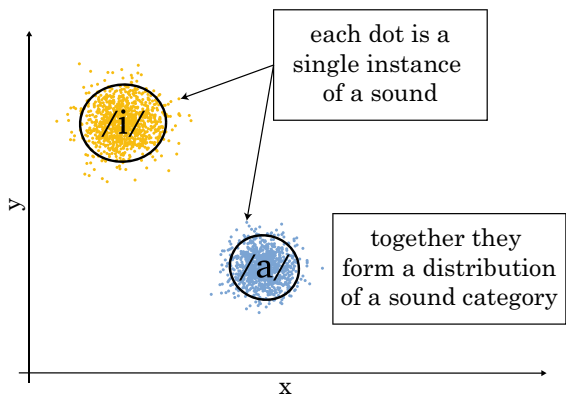
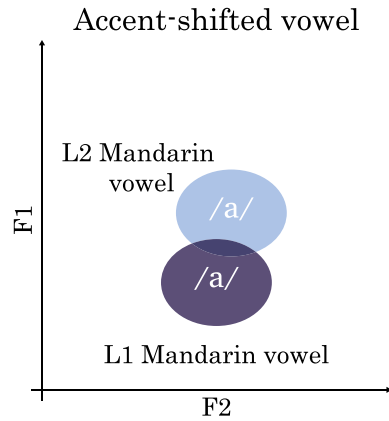


Fig. 2 Illustration of accent-shifted L1 Mandarin vowel



This visualization can help us think about accented pronunciation. In simple terms, when sounds fall outside the distribution of typical values, they are accented. If most or all of the sounds a speaker produces fall outside the typical range that speaker will be perceived to have an accent. (For a much more thorough and technical description of these issues, see Kleinschmidt and Jaeger 2015).

To make this description more concrete, let's consider an example from L2 Mandarin. Figure 2 shows what a hypothetical American English speaker's /a/ sound might look like when they produce Mandarin. By comparing values of the vowel formants (F1 and F2), we can see how similar or different the distribution of the Mandarin /a/ sound (Pinyin *a*) is when produced by our imaginary L1 and L2 speakers. Notice, the L2 vowel distribution slightly overlaps with the L1 distribution, indicating that sometimes the L2 vowel sounds nativelike.

2.2 Accent-Shifted Pronunciation and Pronunciation Errors

This way of thinking about accented speech gives us the ability to highlight some specific phenomena that often occur in L2 pronunciation. I will describe them as *accent-shifted pronunciation* and *pronunciation errors* and these are illustrated in Fig. 3. (This presentation expands on ideas laid out in Pelzl et al. 2020).

2.2.1 Accent-Shifted Pronunciation

The left panel in Fig. 3 shows the distribution of an accent-shifted pronunciation for speech category A. For now, this could be any sound. The L2 speaker produces their own distribution of the sound (A'), and some instances of it fall within the range of the L1 category distribution, but most do not. The result is an accent-shifted sound that will often be recognizably different than the typical L1 sound. Importantly,

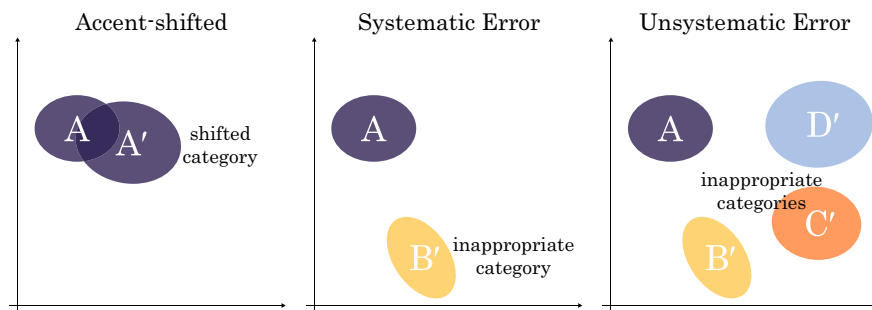


Fig. 3 Illustration of types of accent and error. Accent-shifted pronunciation occurs when a speaker produces the intended sound (A) as a shifted (A') version. Systematic pronunciation error occurs when an inappropriate category (B') is regularly substituted for the appropriate category (A). Unsystematic pronunciation error occurs when multiple inappropriate categories (B', C', D') are substituted for the appropriate one (A). Figure adapted with permission from Pelzl et al. (2020)

however, the L2 version of the sound is not *randomly* different. There's a pattern that will become clear with enough experience. Presumably, then, listeners will be able to adapt to this type of accented pronunciation. Adaptation in this case means that listeners can learn the new sound pattern and quickly recognize accented words containing that sound as being those that the L2 speaker intended. Research with foreign-accented English has shown exactly this type of adaptation (e.g., Baese-Berk et al. 2013; Bradlow and Bent 2008; Clarke and Garrett 2004; Xie et al. 2018). This does not necessarily mean that listening to foreign-accented speech becomes effortless (McLaughlin and Van Engen 2020), but perhaps with enough experience, it would be.

When it comes to L2 Mandarin, most researchers have not discussed pronunciation in terms of foreign accent, but the type of accent-shifted pronunciation pattern described above is nevertheless documented for a variety of L2 Mandarin consonants and vowels (*consonants*: Hao 2012; Lai 2009; Liu and Jongman 2013; Shi 2008; Wang and Chen 2020; Yang and Yu 2019; *vowels*: Hao 2012; Shi 2009; Wu 2011; Wu and Shih 2012).

Similarly, L2 tones are described in ways that I would call accent-shifted. As this may be a novel way to think about tones, we can consider a few examples. L2 tones have been described as often having an overall F0 range that is constrained compared to L1 tones (Chen 1974; Shen 1989; Yang 2015: Chap. 7). Beginning learners have been reported to produce the high Tone 1 as either too high or too low (Miracle 1989; Shen 1989; Wang et al. 2003), and the pitch onset of Tone 4 has been described as too low relative to the speaker's overall F0 range (Shen 1989; Wang et al. 2003; Zhang 2010). Yang (2015: Chap. 4) also discusses patterns that may be influenced by prosodic factors (intonation, phrasing), observing some L2 speakers to consistently over- or under-shoot tones in certain prosodic locations.

According to the analysis presented here, all of these could be considered accent-shifted versions of tones—rather than tone errors. The accent-shifted L2 tones are

different from native patterns, but hypothetically should be recognizable to listeners after they gain some experience with the L2 speaker. However, no research has yet tested this hypothesis.

2.2.2 Pronunciation Errors

In addition to having a foreign accent, another reality for L2 speakers is that they often produce pronunciation errors. In many discussions of foreign accent, errors are simply ignored, or accent-shifted features are described as errors. This is understandable. For listeners, both accents and errors are all wrapped up in the same speech signal and it may not be obvious which sounds are intentional and which accidental. It's also true that a pronunciation error can at the same time be an accent-shifted sound. Still, by drawing sharper distinctions, we can appreciate ways that accent and error from each other, both in terms of why the L2 speaker produces them, and how they might impact listeners.

For L2 speakers, pronunciation errors can be caused by many factors, including inability to hear or form the sounds, insufficient motor muscle control to consistently produce the sounds, or even not knowing what sound is supposed to belong to a given word. Depending on the specific mix of factors, we can outline two broad types of errors: systematic errors that occur with a regular pattern and unsystematic errors that have no clear pattern.

2.2.3 Systematic Pronunciation Errors

The middle panel in Fig. 3 depicts systematic pronunciation errors. In this case, the L2 speaker produces a sound (B') that, for the listener, is categorically different from the typical occurrence of sound A. From the listener's point of view, this is a pronunciation error because it is not the sound they expected to hear. For the speaker, it may well be that they are trying to produce the correct sound, but failing. However, as in the case of accent-shifted sounds, systematic pronunciation errors occur with a pattern. The L2 speaker regularly (if not always) swaps the intended sound with their L2 version of it. In the end then, even though the pronunciation error may be odd, with sufficient experience, a listener could learn the pattern behind it, and adapt so that they more easily and quickly understand the speaker.

As a specific illustration, an L2 speaker of Mandarin may regularly produce the vowel /y/ (as in *lǜ* “green”) as something closer to what the L1 listener expects to be /u/ (as in *lù* “road”). The result would be that words like *lǜ* and *lù* sound the same or much more similar than they should. The pattern does not have to result in another word. For example, an L2 speaker's /p/ (as in *bà* “dad”) could sound like an English speaker's /b/ (as in “bee”). This would not sound quite right, and some listeners might judge it to be an error—but it would also not sound like Mandarin /pʰ/ (as in *pà* “be afraid”). Anecdotally, these examples are actually sounds that English speaking learners of Mandarin struggle to get right. So, though we lack empirical

studies about segmental pronunciation errors in L2 Mandarin, we do have reason to suspect that this type of systematic error pattern will often apply to consonants and vowels (for L2 *perception* of /y/, see Hao 2018).

Systematic errors are also possible, and likely, for tones. It has been suggested that L2 speakers often produce Tone 1 as a falling tone (Miracle 1989; Shen 1989; Wang et al. 2003; Yang 2016), and there may also be positional influences that regularly result in tone swaps or distortions of a certain type (*in disyllabic words*: Zhang and Xie 2020; *in phrases*: Yang 2016). The pattern behind these positional errors might be more difficult for listeners to learn, but as it is a pattern, there is still a chance they will. For speakers of other tonal languages who learn Mandarin as a second tonal language, there may also be consistent tonal errors that happen due to the influence of the tone categories in their L1. For example, Hao (2012) found Cantonese speakers often swapping the high Tones 1 and falling Tone 4 in their L2 Mandarin productions.

2.2.4 Unsystematic Pronunciation Errors

The final, right-most panel in Fig. 3 depicts unsystematic errors. Here, the production of the L2 speaker varies so that multiple inappropriate sound categories are produced for what ought to be a single category. There are a few common causes for unsystematic errors. First, the L2 speaker may not be able to perceive the target speech sound, leading to uncertainty about what it ought to sound like. When they need to produce that sound, they simply make a guess or follow some mistaken intuition about what ought to be produced. In this case, the problem is their knowledge of the sound itself. A related problem is that they may forget what sound a word should have, or be mistaken about what they remember. When several words ought to have the same sound, that sound may instead be different for each word. Sometimes, the L2 speaker swaps sound A with sound B and sometimes with sound C or D. Finally, the error may be driven by a physical lack of control. This might occur with sounds in certain positions in a phrase, or due to emotions, or perhaps nervousness. In all cases, the result for the listener is similar—there is an error, but the cause and direction of the error are not clear.

Unlike accent-shifted pronunciation or systematic pronunciation errors, even with extensive experience listeners will not be able to learn the pattern of unsystematic errors, because there is none. If the unsystematic errors happen with enough frequency, listeners may “adapt” in the sense of learning to ignore pronunciation errors. But whereas adaptation to systematic features of L2 speech improves the speed and ease of understanding the speaker, this type of negative adaptation would only serve to remove a source of interference, pushing the listener to rely more heavily on other contextual cues. This might not actually lead to more efficient or easier comprehension of the L2 speaker. Given that a lifetime of experience has taught listeners to automatically use pronunciation for word recognition, it may be the case that they cannot actually learn to ignore unsystematic pronunciation errors.

Unsystematic errors affecting consonants and vowels may not be common. This is partly because these sounds tend to have simple two-way distinctions, so any category

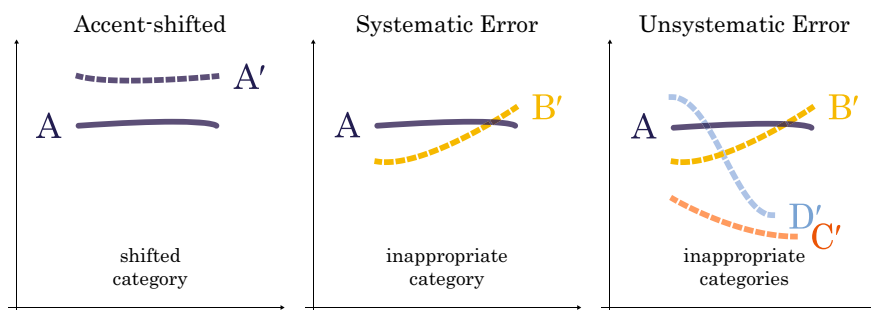


Fig. 4 Illustration of instances of tone accent and tone error. Accent-shifted tones occur when a speaker produces the intended tone (A) as a shifted (A') version. Systematic tone error occurs when an inappropriate tone category (B') is regularly substituted for the appropriate category (A). Unsystematic tone error occurs when multiple inappropriate categories (B', C', D') are substituted for the appropriate one (A)

swaps would naturally lead to a discernable pattern. A speaker who mispronounces the Mandarin /p/ is likely to waver between /b/ and /p^h/, but not to produce /k/. One instance where it may apply in L2 Mandarin is with the high-rounded front vowel /y/ (Pinyin *ü*), mentioned above. The systematic swapping between /y/ and /u/ could be further complicated if the speaker sometimes also produced the sound as /i/ (as in *lì* “force”). If this happened with no discernable pattern, it would be an unsystematic error.

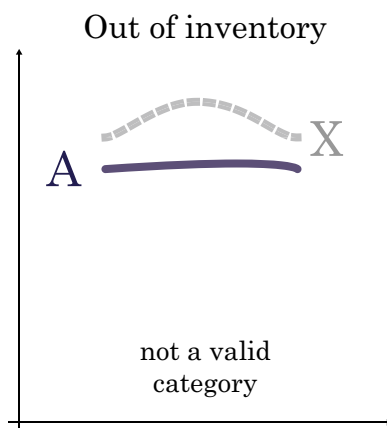
What about tones? Here, it is not only conceivable, but likely quite common for unsystematic errors to occur. Figure 4 recasts the earlier figure to depict how accent-shifted pronunciation and pronunciation error types might apply to tones. In just a moment, we will consider this in much greater detail.

2.2.5 Out of Inventory Errors

One additional error pattern that is worth highlighting, especially when thinking about tones, is the *out of inventory* error (Fig. 5). That is, an L2 tone category that simply doesn't exist in the language. This type of error could be either systematic or unsystematic, and in some cases, it may be just an extreme form of an accent-shifted tone. For example, a beginning L2 speaker might sometimes produce a high tone that is shifted so high as to be judged no longer within the conceivable boundaries of a well-formed high tone.

Zhang (2010) reports that approximately 14% of tone errors made by her L2 participants were judged as out of inventory by raters, and that these tones were mainly realized as a mid-tone or a low-falling tone. Zhang (2010) is somewhat unique in providing this type of analysis. Most studies have not commented on whether errors are in or out of the Mandarin tone inventory.

Fig. 5 Illustration of a tone error that is not just an inappropriate category, but a non-existent category (X)



2.3 How Frequent Are Tone Errors in L2 Speech?

Returning to unsystematic pronunciation errors, the importance of this distinction for tones will depend heavily on whether this type of error is frequent in L2 speech. We do not yet have a clear answer, but there are reasons to suspect they are quite frequent. In research with my colleagues (Pelzl et al. in press), we found that advanced L2 speakers often have gaps in tone knowledge for about 20% of the words they otherwise know confidently. That is, they know the meanings, but not the tones. As these learners know thousands of words, this suggests they will make tone errors for hundreds or even thousands of specific words. From what we can see so far, there appears to be little pattern to what words L2 speakers do or do not know the tones for. It is not the case, for example, that one specific tone is always the culprit, or that errors are always a switch of the same two tones. This lack of clear patterns would seem to make the occurrence and direction of these lexically-based L2 tone errors largely unpredictable for listeners.

Unfortunately, we can't do much more than speculate at this point. As far as I know, the distinctions drawn here (and in Pelzl et al. 2020) are novel, and so no studies have attempted to characterize the accentedness of L2 tones, or to diagnose whether tone errors are systematic or unsystematic. Still, it may be useful to do a short survey of L2 tone production studies to get a sense for how common tone errors (of any type) are, and why we may or may not have noticed the presence of unsystematic errors in earlier studies.

Among beginning L2 Mandarin speakers, research suggests tone errors may be *very* frequent. Chen et al. (2016) created a large corpus of beginning L2 Mandarin speech, with speakers from a wide variety of L1s. They report that tone errors occurred on 32% of all syllables produced. It's worth stressing that this was for the reading of isolated syllables with tones explicitly marked in Pinyin. Explicit notation of tones with the Pinyin diacritics (ā á ǎ à) eliminates the memory component from

tone production and provides an iconic cue to the pitch contour. In other words, elicitation using Pinyin is likely to decrease the occurrence of unsystematic errors.

For more experienced speakers, we certainly expect that the frequency of tone errors will be lower. Estimating based on information available in Yang (2016: Chap. 3), third and fourth year students seemed—on average—to make errors on about 10% of syllables in a reading passage of about 200 characters. Hao (2012) doesn't provide an overall error rate, but it can be seen that for a reading task, errors of some types (e.g., swapping T3 with T2) occurred nearly 30% of the time. Once again, both of these results are for reading with tones explicitly provided.

For spontaneous L2 Mandarin speech, error frequency may be more difficult to judge. Two studies used relatively unscripted responses to question prompts (Kim et al. 2015; Winke 2007). Both report greater than 90% overall tone accuracy. Considering results from more controlled elicitation methods, this is a rather striking finding. Though both studies report a high degree of consistency between raters, it still may be the case that different approaches to training raters would have increased the detection of errors. The fact that the spontaneously elicited speech in these studies was contextualized may also have reduced raters' sensitivity to pronunciation errors. In any case, even 90% accuracy would still mean a speaker makes an error on one in ten syllables.

2.4 How to Investigate Tone Errors in Future Research?

While there will never be a single answer as to the best approach to eliciting L2 speech, the elicitation method is never neutral and will directly impact what we find (e.g., Hao and de Jong 2016). Reading tasks with tones annotations are often favored because they give us a large degree of control over the specific tonal patterns that speakers (attempt to) produce. They may also be a good method for determining how much control L2 speakers have in ideal circumstances. However, these reading tasks will not tell us about a speaker's knowledge of tones for words. Free or planned responses to question prompts may give a better sense of the frequency and type of tone errors that occur, but it can be very difficult to elicit specific words or tone sequences in such tasks.

To date, most L2 tone production research has been framed around questions of the relative difficulty of the different tone categories. Future work might also attempt to analyze the systematic or unsystematic nature of those errors. This will require the use of elicitation methods other than reading tasks in order to give unsystematic errors a chance to occur. Some general approaches might include describing pictures or using question–answer pairs that strongly guide the form of the elicited speech. L2 spoken language corpora would be a potentially invaluable resource for understanding broad trends across L2 speakers. In the opposite direction, targeted studies of individual learners who, impressionistically, produce many or few tone errors could also provide insight into the individual differences that lead to L2 tone errors.

What applies to tone errors is also true for segmental speech errors, which have rarely been examined in L2 Mandarin (but see Chen et al. 2016). Though the unsystematic error type is less likely to occur for segments, it may be that systematic features of segmental L2 speech could influence or be influenced by the frequency and type of errors that occur for L2 tones. For example, perhaps unsystematic tone errors will force listeners to rely more heavily on segmental aspects of L2 speech, thus, increasing the importance of clear pronunciation for those segmental features.

In summary, foreign-accented Mandarin includes accent-shifted pronunciation and pronunciation errors. These features are typical of L2 speech, though their frequency will vary from speaker to speaker. A key question for teachers is: How important is it for L2 learners to overcome accent and reduce errors? The next section begins to address this question.

3 How Does Foreign-Accentedness Affect the Comprehensibility of L2 Mandarin?

People who learn a new language as adults overwhelmingly speak with some degree of foreign accent (e.g., Flege et al. 1995). This does not mean that they cannot improve their pronunciation, but it does raise an important question. *Is a foreign accent a barrier to communication?* Without even conducting any research, we know that extreme answers will not be correct. Widely shared experience tells us that it is not necessary for an L2 speaker to sound exactly like a native speaker in order to communicate effectively. On the other hand, there are certainly cases where a foreign accent can create communication difficulties.

Research in L2 pronunciation has built on these intuitions by trying to measure the relationship between an L2 speaker's accentedness and the comprehensibility of their speech. A highly cited study by Munro and Derwing (1995) suggests the relationship may not be particularly strong. When asked to rate L2 speech samples for accentedness (from weak to strong) and comprehensibility (how easy or difficult a listener finds the speech to understand), they found that even speech rated as strongly accented could still also be rated as highly comprehensible.

These results focused on English. So, as we turn our gaze to Chinese and other tonal languages, a reasonable first question is whether this key finding—that heavily accented speech can also be highly comprehensible—holds for foreign-accented Mandarin speech? Unsurprisingly, from the very start researchers have also wanted to know how tones fit into this relationship.

3.1 *The Relationship Between Accentedness and Comprehensibility in L2 Mandarin*

Lee and Xing (2012) were the first to directly investigate these questions in Mandarin. To explore how tones and segments impacted accentedness and comprehensibility ratings, they made recordings of native Korean L2 speakers of Mandarin reading five simple sentences (e.g., *Jīnwǎn kěnéng huì xiàyǔ*. “It might rain tonight.”). Native Mandarin speakers also produced the same five sentences. Lee and Xing then synthesized versions of the sentences with the prosody (i.e., intonation and tones) and segmental features swapped, so that there were sentences with Korean L2 segments and L1 Mandarin prosody, as well as sentences with Korean L2 prosody and L1 Mandarin segments. These manipulated sentences were then rated by a group of native Chinese listeners. Results showed a clear difference in the perceived accentedness of the manipulated sentences. When L2 segments were present (with L1 prosody), accent was rated more strongly than when L2 prosody was present (with L1 segments). The authors interpret this as evidence that L2 segmental features in Mandarin are more important in conveying accentedness than are the prosodic features—which, of course, includes tones. Like Munro and Derwing (1995), comprehensibility ratings failed to show a strong relationship with accentedness ratings. However, this could be because the sentences were very simple and repeated many times over the course of the study, so that comprehensibility was never a serious issue for listeners after they had heard the sentences a few times. It should also be noted that a single statistical significance test cannot provide support for the absence of an effect.

Lee and Xing’s study is the only one to date that has attempted to make a direct comparison between segmental and tonal (prosodic) features of foreign-accented Mandarin. The result is striking and might suggest tones are not as important as segmental pronunciation in L2 Mandarin. Unfortunately, there are some missing details that make it difficult to fully evaluate the outcomes. Specifically, we do not know what the L2 speakers’ tones were like in the recorded stimuli. Were they accent-shifted tones? Did they include tone errors? Accented but otherwise accurate tones might not be expected to have much impact on ratings, whereas outright tone errors would be expected to have much stronger impacts. The small number of very simple stimulus sentences also raises some questions about the generalizability of results to more complicated and varied L2 speech. What happens when vocabulary is not so frequent and predictable? Nevertheless, Lee and Xing applied an interesting approach that might be worth pursuing further in future work.

Working with native English speaking Mandarin learners, Yang also evaluated the role of tones and prosody in foreign-accented speech (Yang 2016: Chap. 8). A group of native Chinese raters listened to a small number of short sentences read by either L1 or L2 Mandarin speakers. The raters transcribed the sentences, rated the comprehensibility and accentedness of the speaker, and provided some indication of what they had based their ratings on. Results suggested strong correlations between the

accuracy of transcriptions and the ratings of comprehensibility and accentedness—that is, the stronger a speaker’s accent, the less comprehensible listeners thought that speaker was.

On its face, this contrasts with the results in English (Munro and Derwing 1995). However, Yang’s stimulus sentences were quite different from those used in previous accent studies. Whereas those studies typically had people describe pictures or read narrative passages, Yang’s sentences were crafted with much more specific features in mind. Each sentence was exactly six syllables long, had tightly controlled tone patterns, avoided many of the Mandarin consonants, and always contained a rather tricky word-boundary ambiguity (Yang 2016: Chap. 4, pp. 60–61). For example, the sentence “*Wū Ānyīng xiū fēijī.*” (“Wu Anying repairs planes.”) has only the high Tone 1 and requires (like all sentences did) a subtle difference in prosodic phrasing in order to disambiguate whether the proper name is two or three syllables long. With the change of just one written character (and slightly different phrasing), the sentence becomes “*Wū Ān yīngxiū fēijī.*” (“Wu An should repair planes.”). These tricky sentences resulted in a rather large number of transcription errors even when they were produced by native speakers. These challenging stimuli contributed heavily to the outcomes. We can certainly conclude that accent *can* contribute to comprehensibility—and likely will when prosodic or tone ambiguities are present. We cannot conclude that it *usually* does so, because spoken language typically occurs in context and quite rarely has either the tonal or prosodic features seen in these stimuli.

Freeborn and Rogers (2019) also carried out a rating study with foreign-accented Mandarin, though their aims were a bit different. They wanted to establish whether individual differences among learners would relate to ratings of accentedness. Four L1 and seventy L2 Mandarin speakers—from a wide variety of language backgrounds—read a passage in Chinese with Pinyin annotations. Fifteen L1 Mandarin listeners rated the accentedness of the first two sentences produced by each speaker. Using these ratings, Freeborn and Rogers explored how a large set of seventeen different speaker variables were related to the ratings. Variables included things like current age, age when the speaker began learning, musical training, and so on. The strongest relationship to L1 ratings turned out to be the L2 speaker’s own rating of their personal accentedness, with proficiency level (participant’s level on the *Hanyu Shuiping Kaoshi*, a standard test of Chinese proficiency used in the PRC), and motivation as the second and third most related factors.

The authors argue that these results show the importance of tones for L2 accentedness. However, this interpretation is not very convincing. Their study had no objective measure of tones at all. Their arguments are based on speculation that L2 learners’ ratings of their own accentedness depended on their experience of having conversational breakdowns caused by poor control of tones. This chain of logic might be correct, but they provide little evidence to support it. Additionally, there are reasons to be skeptical of the statistical outcomes in the study given the large number of variables and relatively small number of ratings.

Though not a full-blown rating study, a follow-up question for participants in a study I conducted with my colleagues may also shed some light on the question of tone and accentedness (Pelzl et al. 2020). As shown in Fig. 6, we found that L1 Chinese

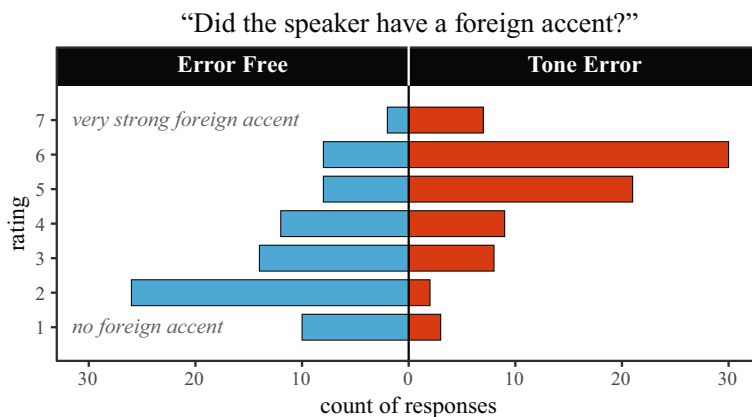


Fig. 6 Listener ratings of foreign accent strength when the same L2 speakers either produced Mandarin without tone errors (error free) or with frequent tone errors (tone error). Figure adapted with permission from Pelzl et al. (2020)

listeners consistently judged an L2 speaker as more accented when that speaker produced tone errors compared to when that same speaker did not produce tone errors. This suggests that tone errors do play some role in producing impressions of a foreign accent. However, this result does not necessarily show special importance for tones over other aspects of L2 pronunciation. Our study had only two L2 speakers, they produced only isolated disyllabic words, and the study design specifically contrasted speakers with respect to their control of tones. Just like in Yang’s study (2016), these factors were likely to make tones (and tone errors) highly salient. (See also the chapter by Chen and Yang in this volume.)

Though not specifically investigating foreign-accented Mandarin, researchers who work with hearing or speech impaired populations also want to understand how tones affect the comprehensibility of Mandarin speech. A number of studies have examined the role of tones by flattening or otherwise manipulating the F0 contours of words and sentences, and then having listeners perform transcription or rating tasks with those sentences. Patel et al. (2010) presented sentences with either their original tones, or a monotone across the whole sentence. In quiet background, the monotone sentences did not cause difficulty for listeners. However, when multi-speaker babble noise was added, listeners were less accurate in transcribing the monotone sentences compared to sentences with tones intact. Further research has shown that flattened tones may have even stronger impacts on elderly or hearing impaired Mandarin listeners (Jiang et al. 2017). These lines of work suggest that similar difficulties would be likely for L2 speech, where tones are not just flattened, but often misleading. Speech-in-noise research could be valuable for understanding how tones in foreign-accented speech might interact with natural (noisy) environments.

Once again, we should exercise some caution when interpreting these studies. On the one hand, a simplistic interpretation of results could lead us to underestimate the value of tones. In these studies, no direct contrast was made with segmental

features, so results only speak to the impact of tones when segmental pronunciation is accurate. Listeners might rely on tones more heavily when segmental pronunciation is less clear. At the same time, a simplistic reading of results could exaggerate the importance of tones. The stimuli sentences were designed to be challenging and to test comprehension. For example, Patel and colleagues used relatively formal news language, which may present different lexical challenges than most typical L2 speech. Other studies have used word lists or nonsense sentences, specifically aiming to remove the benefits of meaningful context. Of course, context matters (Wang et al. 2013). When listeners can rely on context, they may be able to easily overcome some of the challenges that misleading tones (or foreign-accented speech) might otherwise present.

In summary, current research clearly shows that tones *can* be a marker of L2 accent, and that in adverse listening conditions or when words are ambiguous because of tones, they can contribute to difficulties in comprehension. However, if pitted against segmental features, it remains unclear whether tones play an equal, greater, or lesser general role in creating the impression of accent or interfering in smooth comprehension.

3.2 Can Listeners Adapt to Foreign-Accented Mandarin?

Even when listeners initially find foreign-accented speech difficult to comprehend, we know they can often adapt. People can improve in word recognition for specific accented sounds after hearing just a handful of sentences (Clarke and Garrett 2004; Xie et al. 2018). They get better at the transcription of foreign-accented speech over time, regardless of the strength of a speaker's accent (Bradlow and Bent 2008; for a review, see Baese-Berk et al. 2020).

These positive trends are encouraging. However, this is an instance where the differences between accent-shifted pronunciation and pronunciation errors may become quite important. When accented speech has a pattern, listeners should be able to adapt. When errors undermine the presence of an obvious pattern, listeners may be unable to adapt or perhaps will adapt by ignoring the errors and looking elsewhere to guide comprehension. This latter outcome is one possible interpretation for a number of neuro-imaging studies that have found listeners displaying different brain responses to grammatical and lexical errors if those errors are produced by foreign-accented speakers rather than native speakers (e.g., Caffarra and Martin 2019; Grey and van Hell 2017; Hanulíková et al. 2012; Romero-Rivas et al. 2015).

Taking cues from such studies, my colleagues and I used behavioral and neural measures to test how L1 Mandarin listeners responded to pronunciation errors that occurred in spoken sentences (Pelzl et al. in press). Two speakers read a large number of sentences. One was a native speaker with a typical (Beijing) Mandarin accent and the other was an American L2 speaker of Mandarin. Listeners heard the sentences while their electroencephalogram (EEG, “brainwaves”) was recorded, and for each sentence they judged whether or not they had heard a word or pronunciation error.

We wanted to know whether the listeners would respond differently to tonal and segmental pronunciation errors depending on which speaker produced them. The behavioral judgments of listeners made it clear that they responded differently to the foreign-accented speaker—some listeners seemed to find errors even in his “good” sentences. At the same time, as a group, listeners were more likely to judge sentences with tone errors as acceptable if they were produced by the L2 speaker. This may mean they ignored or did not notice some of the L2 tone errors, but it could also indicate they had more difficulty judging tone errors in foreign-accented speech compared to native speech.

Listeners’ neural responses did not show any major differences between the two speakers. There was, however, an overall trend that fits with previous accent studies, indicating that perhaps listeners are less likely to be surprised or even to notice fine-grained pronunciation errors from a foreign-accented speaker. This trend was similar for tonal and segmental pronunciation errors.

Unfortunately, we did not gather more information about *why* listeners made the judgments they did. It could be that some had more or less experience hearing foreign-accented Mandarin (e.g., on TV, among friends), had different levels of strictness in deciding what an error was, or focused on different aims during the task (i.e., comprehending the message vs. judging pronunciation). We also did not get ratings of accentedness or comprehensibility during this study.

Finally, even though one of our goals was to investigate adaptation to foreign-accented Mandarin, we did not find any evidence of changes in listener responses over the course of the study. Failure to find adaptation, however, does not mean adaptation did not occur. Perhaps a different task or response, or simply a larger-scale study (more participants) would find evidence of adaptation. It is also possible that the way we manipulated pronunciation (both tonal and segmental) resulted in arbitrary and unsystematic errors. As argued above, it may be impossible for listeners to adapt to this type of error.

In another study (Pelzl et al. 2020), we focused in more narrowly on tones, specifically aiming to examine the effects of unsystematic tone errors. Two L2 speakers—we’ll call them speaker A and speaker B—produced isolated two-syllable words. On each trial in the study, native Mandarin listeners heard an L2 speaker produce a word and then saw a written Chinese word. In some cases, the written words matched what was spoken; in others, the written word was different. This was meant to create a *priming* effect so that responses would be faster when words were the same in both spoken and written form. Native Chinese participants all heard both L2 speakers, but for half of the participants speaker A made tone errors on 50% of filler words, while speaker B made no tone errors. For the other half of participants, this was reversed: speaker B made 50% tone errors, speaker A made none. Our question was whether the frequency of tone errors would slow down listeners’ recognition of words, even when the words were produced correctly. If so, this would be a strong argument for the negative effects of unsystematic tone errors on L2 comprehension.

The answer from this single study was negative. It didn’t matter whether or not the L2 speaker made tone errors, listeners always responded equally fast when words were spoken correctly. At the same time, when tone errors did occur, listeners were a

bit slower to recognize the words. For example, if they heard the incorrectly produced *nènglì* and then saw the “matching” written word 能力, they were slightly slower to recognize it as a real word. In short, we found that tone errors have a direct impact on the speed with which listeners recognize words, but we did not find any evidence that listeners adapted to a speaker who made frequent tone errors.

Practically speaking, the results of our two studies show that, for two-syllable words, tone errors do impact the speed and efficiency with which listeners recognize words both in isolation and in context. At a minimum then, tone errors seem likely to increase the effort needed to understand foreign-accented Mandarin. We cannot be sure the same patterns would apply for single syllable words, where tone errors are much more likely to result in a completely different words, rather than merely mispronounced ones. This would suggest single syllable words will lead to more confusion—but it has to be balanced against the fact that many single syllable words are extremely frequent in conversation (Tao 2015), and likely to be easily inferred in context.

4 Future Directions for Foreign-Accented Mandarin Research

As research on foreign-accented Mandarin is just beginning, there are many basic questions that can be asked. For those interested in research with practical applications to classrooms, I will take a moment to consider three of the main questions whose answers might provide significant guidance for teaching practices.

4.1 *What Specific Sounds May Be Most Important to Target in Pronunciation Teaching?*

Given the major role tones play in L2 pedagogy and the challenge they present to many learners, the focus on tones in current research is understandable. Another reason that tones may be a popular topic of study is that, with only a handful of them, it is much more tractable to target them all at once, compared to consonants and vowels. Still, whatever the ultimate findings are for the importance of tones, it will not mean that consonants and vowels don't matter.

Future research might try to find a route into segments by evaluating whether some consonants or vowels are more important than others. In research on English, one interesting approach to this question has been through the lens of *functional load* (Kang and Moran 2014; Munro and Derwing 2006; Suzukida and Saito 2019). Essentially, the idea is that some sets of contrastive sounds may be more important than others, because—across the spoken vocabulary—they serve to distinguish more words. For example, /b/ and /p/ in English differentiate many words (bit/pit,

back/pack, bat/pat, etc.) and thus have a high functional load. In contrast, the sounds /θ/ and /ð/ (as in “**th**igh” and “**th**y”) differentiate very few words and so have a low functional load. Though so far somewhat exploratory, the studies that have investigated these issues in English seem promising. For work along these lines in Mandarin, guidance can be sought from a very active line of research addressing the informational and statistical properties of consonants, vowel, and tones (Tong et al. 2008; Wiener 2020; Wiener and Ito 2015, 2016; Wiener et al. 2019; Wiener and Turnbull 2015; Yao and Sharma 2017).

Additionally, existing studies on L2 tone production can guide explorations about how specific tonal features impact listeners’ perceptions of accentedness or the actual comprehension of L2 speech. For example, recent discussions about the best approach to teaching Tone 3 might gain further clarity by gathering listener responses to L2 speech (e.g., He et al. 2016; Shi 2007; Sparvoli 2017; Wen and Yan 2015; Zhang 2014).

4.2 How Do Prosodic Features of Foreign-Accented Speech Impact Comprehensibility?

Tones and segments are not the only important aspects of pronunciation. In English language research, suprasegmental aspects of foreign-accented speech—intonation, stress, speech rate—have received quite a bit of attention (Kang 2010; Munro 1995). Some studies have suggested training on those features does more to increase L2 comprehensibility than training only on segmental features (Derwing et al. 1998; Derwing and Rossiter 2003). Future work in Mandarin would do well to also consider these prosodic features of foreign-accented speech. As mentioned above, this was one part of Yang’s (2016) study, and Lee and Xing (2012) also describe their study in terms of prosody, rather than just tones. By expanding from this work, and also incorporating insights from other L2 research, we can begin to test whether broader prosodic trends might deserve more attention in Chinese classroom teaching.

4.3 What Are the Social Implications of Foreign-Accented Mandarin?

Even when foreign-accented pronunciation does not impede comprehensibility, it often comes with social costs. I began this chapter by referencing the phrase *yáng qiāng yáng diào*, which is used to refer to the speech of foreign-accented Mandarin speakers. Though the specific implications of the phrase can be shaped by many contextual factors, it often bears a negative connotation (DeFrancis 2003). So then, whether we like it or not, it is worth understanding the social costs associated with

foreign-accented speech, as well as what L2 speakers can or cannot do to mitigate those costs.

As mentioned briefly above, one type of social cost comes from the increased effort foreign-accented speech sometimes requires of listeners (McLaughlin and Van Engen 2020). Not every person will have the same amount of patience and determination when communicating with an L2 speaker. Insofar as L2 speakers can improve their pronunciation, they may be able to lessen the burden on their listeners.

Unfortunately, not every social cost can be mitigated by improved L2 pronunciation. Social psychologists have found bias toward or against individuals based on their appearance, such that the same vocal recordings presented with different faces resulted in different judgments of accentedness—a phenomenon that has come to be called “reverse linguistic stereotyping” (Kang and Rubin 2009, 2014; Rubin 1992). Undoubtedly, similar things occur among Chinese listeners who may be biased to expect foreign-accented Mandarin from those who fit their expectations of what L2 speakers look like (i.e., non-Chinese), or alternatively, biased to expect nativelikeness from those who look like L1 speakers (i.e., appear Chinese). Research in these areas should be conducted with due sensitivity, but could be very useful for understanding what is and isn’t in the control of the L2 speaker.

Relatedly, additional work could be conducted looking at the role of non-standard (regional) Chinese accents when produced by L2 speakers. Diao (2017) has conducted one interesting study along these lines, considering L2 speakers who chose to retain regional features in their Mandarin speech.

For all research into foreign-accent, it will of course be important to determine what results are broadly generalizable across different native language groups, and what results are more dependent on the L2 speaker’s specific linguistic experience.

5 Conclusion

Every speaker has an accent. L2 speakers of Mandarin are no different. By studying the ways that foreign-accented speech affects listeners, we can slowly build toward a more empirically driven understanding of what needs to happen for learners to communicate effectively in Mandarin. This work is just beginning, I hope that in another ten years, a review like this will have more numerous and more concrete results to share.

Acknowledgements This research was supported in part by the National Science Foundation (SBE 2004279).

References

- Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, *133*(3), EL174–EL180. <https://doi.org/10.1121/1.4789864>.
- Baese-Berk, M. M., McLaughlin, D. J., & McGowan, K. B. (2020). Perception of non-native speech. *Language and Linguistics Compass*, *14*(7). <https://doi.org/10.1111/lnc3.12375>.
- Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). Amsterdam: John Benjamins.
- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. <https://doi.org/10.1016/j.cognition.2007.04.005>.
- Caffarra, S., & Martin, C. D. (2019). Not all errors are the same: ERP sensitivity to error typicality in foreign accented speech perception. *Cortex*, *116*, 308–320. <https://doi.org/10.1016/j.cortex.2018.03.007>.
- Chen, G. (1974). The pitch range of English and Chinese speakers. *Journal of Chinese Linguistics*, *2*, 159–171.
- Chen, N. F., Wee, D., Tong, R., Ma, B., & Li, H. (2016). Large-scale characterization of non-native Mandarin Chinese spoken by speakers of European origin: Analysis on iCALL. *Speech Communication*, *84*, 46–56. <https://doi.org/10.1016/j.specom.2016.07.005>.
- Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647. <https://doi.org/10.1121/1.1815131>.
- DeFrancis, J. (2003). *ABC Chinese-English comprehensive dictionary: Alphabetically based computerized*. Honolulu: University of Hawai'i Press.
- Derwing, T. M., & Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning*, *13*(1), 1–17.
- Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, *48*(3), 393–410. <https://doi.org/10.1111/0023-8333.00047>.
- Diao, W. (2017). Between the standard and non-standard: Accent and identity among transnational Mandarin speakers studying abroad in China. *System*, *71*, 87–101. <https://doi.org/10.1016/j.system.2017.09.013>.
- Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*(04). <https://doi.org/10.1017/S0272263104040021>.
- Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). Timonium, MD: York.
- Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, *97*(5), 3125–3134.
- Freeborn, L., & Rogers, J. (2019). Nonlinguistic factors that affect the degree of foreign accent in second language Mandarin. *Studies in Chinese Linguistics*, *40*(1), 75–99. <https://doi.org/10.2478/sci-2019-0003>.
- Grey, S., & van Hell, J. G. (2017). Foreign-accented speaker identity affects neural correlates of language comprehension. *Journal of Neurolinguistics*, *42*, 93–108. <https://doi.org/10.1016/j.jneuroling.2016.12.001>.
- Hanulíková, A., van Alphen, P. M., van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, *24*(4), 878–887.
- Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*(2), 269–279. <https://doi.org/10.1016/j.wocn.2011.11.001>.

- Hao, Y.-C. (2018). Second language perception of Mandarin vowels and tones. *Language and Speech*, 61(1), 135–152. <https://doi.org/10.1177/0023830917717759>.
- Hao, Y.-C., & de Jong, K. (2016). Imitation of second language sounds in relation to L2 perception and production. *Journal of Phonetics*, 54, 151–168. <https://doi.org/10.1016/j.wocn.2015.10.003>.
- He, Y., Wang, Q., & Wayland, R. (2016). Effects of different teaching methods on the production of Mandarin tone 3 by English speaking learners. *Chinese as a Second Language*, 51(3), 252–265.
- Jiang, W., Li, Y., Shu, H., Zhang, L., & Zhang, Y. (2017). Use of semantic context and F0 contours by older listeners during Mandarin speech recognition in quiet and single-talker interference conditions. *The Journal of the Acoustical Society of America*, 141(4), EL338–EL344. <https://doi.org/10.1121/1.4979565>.
- Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, 38(2), 301–315. <https://doi.org/10.1016/j.system.2010.01.005>.
- Kang, O., & Moran, M. (2014). Functional loads of pronunciation features in nonnative speakers' oral assessment. *TESOL Quarterly*, 48(1), 176–187. <https://doi.org/10.1002/tesq.152>.
- Kang, O., & Rubin, D. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, 28(4), 441–456.
- Kang, O., & Rubin, D. (2014). Reverse linguistics stereotyping. In J. Levis & A. Moyer (Eds.), *Social dynamics in second language acquisition* (pp. 239–253). Berlin: DeGruyter Mouton.
- Kim, J., Dewey, D. P., Baker-Smemoe, W., Ring, S., Westover, A., & Eggett, D. L. (2015). L2 development during study abroad in China. *System*, 55, 123–133. <https://doi.org/10.1016/j.system.2015.10.005>.
- Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148–203. <https://doi.org/10.1037/a0038695>.
- Lai, Y. (2009). Asymmetry in Mandarin affricate perception by learners of Mandarin Chinese. *Language and Cognitive Processes*, 24(7–8), 1265–1285. <https://doi.org/10.1080/01690960802113850>.
- Lee, J.-K., & Xing, L. (2012). The relative weight of prosody and segment in the perception of Korean accented Chinese speech. *Language and Linguistics (Korean Journal)*, 54, 263–293.
- Liu, J., & Jongman, A. (2013). *American Chinese learners' acquisition of L2 Chinese affricates /ts/ and /tsh/* (p. 060005). Kansas City. <https://doi.org/10.1121/1.4798223>.
- Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. New York, NY: Routledge.
- McLaughlin, D. J., & Van Engen, K. J. (2020). Task-evoked pupil response for accurately recognized accented speech. *The Journal of the Acoustical Society of America*, 147(2), EL151–EL156. <https://doi.org/10.1121/10.0000718>.
- Miracle, W. C. (1989). Tone production of American students of Chinese: A preliminary acoustic study. *Journal of the Chinese Language Teachers Association*, 24(3), 49–65.
- Munro, M. J. (1995). Nonsegmental factors in foreign accent: Ratings of filtered speech. *Studies in Second Language Acquisition*, 17, 17–34.
- Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech and second language learners. *Language Learning*, 45(1), 73–97.
- Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, 34(4), 520–531. <https://doi.org/10.1016/j.system.2006.09.004>.
- Norman, J. (1988). *Chinese*. New York, NY: Cambridge University Press.
- Patel, A. D., Xu, Y., & Wang, B. (2010). The role of F0 variation in the intelligibility of Mandarin sentences. *Speech Prosody*, 4.
- Pelzl, E., Carlson, M. T., Guo, T., Jackson, C. N., & van Hell, J. G. (2020). Tuning out tone errors? Native listeners do not down-weight tones when hearing unsystematic tone errors in foreign-accented Mandarin. *Bilingualism: Language and Cognition*, 1–8. <https://doi.org/10.1017/S1366728920000280>.

- Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (in press). Even in the best-case scenario L2 learners have persistent difficulty perceiving and utilizing tones in Mandarin: Findings from behavioral and event-related potentials experiments. *Studies in Second Language Acquisition*. <https://doi.org/10.1017/S027226312000039X>.
- Pelzl, E., Lau, E. F., Guo, T., Jackson, S. R., & Gor, K. (in press). Behavioral and neural responses to tone errors in foreign-accented Mandarin. *Language Learning*. <https://doi.org/10.1111/lang.12438>.
- Romero-Rivas, C., Martin, C. D., & Costa, A. (2015). Processing changes when listening to foreign-accented speech. *Frontiers in Human Neuroscience*, 9. <https://doi.org/10.3389/fnhum.2015.00167>.
- Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, 33(4), 511–531. <https://doi.org/10.1007/BF00973770>.
- Shen, X. S. (1989). Toward a register approach in teaching Mandarin tones. *Journal of the Chinese Language Teachers Association*, 24(3), 27–47.
- Shi, J. (2007). On teaching tone three in Mandarin. *Journal of the Chinese Language Teachers Association*, 42(2), 1–10.
- Shi, F. (2008). *The structure of speech sounds: The intersection of phonetics and phonology* [语音格局—语音学与音系学的交汇点]. Beijing: Shangwu Yin Shu Guan.
- Shi, F. (2009). *Shiyan Yinxue Tansuo* [Exploration of experimental phonology]. Beijing: Peking University Press.
- Sparvoli, C. (2017). From phonological studies to teaching Mandarin tone: Some perspectives on the revision of the tonal inventory. In I. Kecskes & C. Sun (Eds.), *Key issues in Chinese as a second language research* (pp. 81–100). New York, NY: Routledge.
- Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research*, 136216881985824. <https://doi.org/10.1177/1362168819858246>.
- Tao, H. (2015). Profiling the Mandarin spoken vocabulary based on corpora. In W. S.-Y. Wang & C. Sun (Eds.), *The Oxford handbook of Chinese linguistics*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199856336.013.0031>.
- Tong, Y., Francis, A. L., & Gandour, J. T. (2008). Processing dependencies between segmental and suprasegmental features in Mandarin Chinese. *Language and Cognitive Processes*, 23(5), 689–708. <https://doi.org/10.1080/01690960701728261>.
- Wang, X., & Chen, J. (2020). The acquisition of Mandarin consonants by English learners: The relationship between perception and production. *Languages*, 5(2), 20. <https://doi.org/10.3390/languages5020020>.
- Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, 113(2), 1033. <https://doi.org/10.1121/1.1531176>.
- Wang, J., Shu, H., Zhang, L., Liu, Z., & Zhang, Y. (2013). The roles of fundamental frequency contours and sentence context in Mandarin Chinese speech intelligibility. *The Journal of the Acoustical Society of America*, 134(1), EL91–EL97. <https://doi.org/10.1121/1.4811159>.
- Wen, B., & Yan, F. (2015). Liuxuesheng Hanyu shengdiao xide zhong yang-shang heliude shiyan yanjiu. *Journal of the Chinese Language Teachers Association*, 50(1), 19–41.
- Wiener, S. (2020). Second language learners develop non-native lexical processing biases. *Bilingualism: Language and Cognition*, 23(1), 119–130. <https://doi.org/10.1017/S1366728918001165>.
- Wiener, S., & Ito, K. (2015). Do syllable-specific tonal probabilities guide lexical access? Evidence from Mandarin, Shanghai and Cantonese speakers. *Language, Cognition and Neuroscience*, 30(9), 1048–1060. <https://doi.org/10.1080/23273798.2014.946934>.
- Wiener, S., & Ito, K. (2016). Impoverished acoustic input triggers probability-based tone processing in mono-dialectal Mandarin listeners. *Journal of Phonetics*, 56, 38–51. <https://doi.org/10.1016/j.wocn.2016.02.001>.

- Wiener, S., & Turnbull, R. (2015). Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. *Language and Speech*, 0023830915578000.
- Wiener, S., Lee, C., & Tao, L. (2019). Statistical regularities affect the perception of second language speech: Evidence from adult classroom learners of Mandarin Chinese. *Language Learning*, 69(3), 527–558. <https://doi.org/10.1111/lang.12342>.
- Winke, P. M. (2007). Tuning into tones: The effect of L1 background on L2 Chinese learners' tonal production. *Journal of the Chinese Language Teachers Association*, 42(3), 21–55.
- Wu, C. (2011). *The evaluation of second language fluency and foreign accent*. University of Illinois at Urbana-Champaign.
- Wu, C., & Shih, C. (2012). *A corpus study of native and non-native vowel quality*. Presented at the Speech Prosody, Shanghai.
- Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America*, 143(4), 2013–2031. <https://doi.org/10.1121/1.5027410>.
- Yang, B. (2015). *Perception and production of Mandarin tones by native speakers and L2 learners*. New York: Springer.
- Yang, C. (2016). *The acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice*. Philadelphia, PA: John Benjamins Publishing Co.
- Yang, C., & Yu, A. C. L. (2019). The acquisition of Mandarin affricates by American second language learners. *Taiwan Journal of Linguistics*, 17(2), 91–122. [https://doi.org/10.6519/TJL.201907_17\(2\).0004](https://doi.org/10.6519/TJL.201907_17(2).0004).
- Yao, Y., & Sharma, B. (2017). What is in the neighborhood of a tonal syllable? Evidence from auditory lexical decision in Mandarin Chinese. *Proceedings of the Linguistic Society of America*, 2, 45. <https://doi.org/10.3765/plsa.v2i0.4090>.
- Zhang, H. (2010). Phonological universals and tonal acquisition. *Journal of the Chinese Language Teachers Association*, 45(1), 39–65.
- Zhang, H. (2014). The third tone: Allophones, sandhi rules and pedagogy. *Journal of the Chinese Language Teachers Association*, 49(1), 117–145.
- Zhang, H., & Xie, Y. (2020). Coarticulation effects of contour tones in second language Chinese. *Chinese as a Second Language Research*, 9(1), 1–30. <https://doi.org/10.1515/caslar-2020-0001>.

Exploring Fluency and Disfluency Features of Oral Performances in Chinese as a Second Language



Yuyun Lei

Abstract This study investigates various fluency and disfluency features of oral performances by second language (L2) learners of Chinese to explore how these features differ and develop at different levels of oral proficiency in L2 Chinese. Although fluency has been extensively researched, few studies have explored oral fluency in L2 Chinese, with the available ones addressing a small number of fluency features or a restricted range of learner proficiency. The present study extends this body of research by including L2 learners at various curricular levels and by examining a more comprehensive set of fluency features. Oral responses to a narrative task were collected from thirty-eight L2 learners of Chinese at a US university. Their responses were holistically scored on four different levels and were analyzed for eleven fluency and disfluency features, including features of the amount and rate of speech, pausing, and repairs. Results showed that features of the amount and rate of speech and silent pausing not only demonstrated strong relationships with score levels but could also distinguish among the various levels with more distinctive differences observed at higher score levels than at lower ones. These findings have important implications for the teaching and assessment of speaking in L2 Chinese.

Keywords Oral assessment · Fluency · Disfluency · Language proficiency · L2 Chinese

1 Introduction

Fluency has long been recognized as an important aspect of second language (L2) teaching and learning. Although fluency can be used to describe written and spoken language, it usually refers to oral fluency. At school, students consider developing oral fluency as one of their ultimate goals of learning an L2; teachers frequently assess fluency in their evaluations of students' oral performances. Outside the classroom, fluency is featured in rating scales of high-stake tests, such as the ACTFL Oral

Y. Lei (✉)

University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA

e-mail: ylei1academic@hotmail.com

© Springer Nature Singapore Pte Ltd. 2021

C. Yang (ed.), *The Acquisition of Chinese as a Second Language Pronunciation*, Prosody, Phonology and Phonetics, https://doi.org/10.1007/978-981-15-3809-4_13

281

Proficiency Interview (OPI), the Test of English as a Foreign Language (TOEFL), and the Hànyǔ Shuǐpíng Kǎoshì (HSK) Speaking Test, all of which have been used for purposes of university admission and employment selection. In the field of second language acquisition (SLA) research, fluency is a major component of spoken language ability, an essential criterion of L2 oral performance and development, as well as a reliable indicator of the cognitive processes of speech production (Housen and Kuiken 2009; Koponen and Riggensbach 2000; Segalowitz 2010).

Because of the significance of fluency in second language studies, a large volume of research has been devoted to understanding what constitutes fluency, how fluency develops, and whether fluency can predict oral performance and proficiency. While previous studies have indicated that a number of temporal measures can reliably represent fluency and predict oral performance and proficiency, there are mixed results regarding what features can best characterize fluency at different levels of oral proficiency, and whether these features can consistently distinguish across levels (Ginther et al. 2010; Iwashita et al. 2008; Tavakoli et al. 2020). In addition, the majority of fluency studies were conducted on learners of English as a second language (ESL). Only a handful of studies have examined fluency in oral performance by L2 learners of Chinese (e.g., Chen 2015; Jin and Mak 2013; Shih and Wu 2011; Zhai 2011). Those studies also tended to focus on small sample sizes, a restricted range of proficiency levels, or a limited number of fluency features. To extend this body of research, this study included L2 learners at different curricular levels in a university-level Chinese language program and investigated a comprehensive set of fluency and disfluency features. The aim of the study is to explore how fluency and disfluency features develop and distinguish across different levels of oral proficiency in L2 Chinese. The findings of the study are expected to further the understanding of fluency in L2 Chinese and have implications for the teaching and assessment of speaking in L2 Chinese.

2 Literature Review

2.1 Definitions of Fluency

Fluency has been conceptualized in first and second language research. Concerning first language (L1), Fillmore (1979) identified four types of fluency a native speaker could possess. The first type of fluency is the ability of a speaker to maintain speech flow with few pauses. The second type is the ability to talk in coherent, reasoned, and semantically and syntactically dense sentences. The third type relates to the ability to produce socially and contextually appropriate language. The last type has to do with the ability to use language creatively and imaginatively, and easily find novel ways to express ideas.

As regards L2, Lennon (1990) pointed out that there are broad and narrow senses of fluency. The broad sense of fluency is a cover term for a person's global language

ability, manifested in the perception of ease, eloquence, and smoothness of speech (Housen and Kuiken 2009). The broad sense encompasses the four types of fluency found in Fillmore's (1979) conceptualization. This concept of fluency is often seen in people's comments spoken in daily life, such as, "She speaks English fluently". On the other hand, the narrow sense of fluency is one of the components of spoken language ability, corresponding to Fillmore's (1979) first type of fluency. It deals with the speed and smoothness of oral delivery and can be measured by an array of temporal variables, such as speech rate, the number and length of silent pauses, and so on. The goal of L2 learning, as Lennon (1990) argued, is to "produce speech at the tempo of native speakers, unimpeded by silent pauses and hesitations, filled pauses, self-corrections, repetitions, false starts, and the like" (p. 390).

With a more recent understanding of the cognitive bases of fluency, Segalowitz (2010) proposed that under the narrow sense of fluency, three distinct aspects of fluency could also be identified. These aspects include cognitive fluency, utterance fluency, and perceived fluency. Cognitive fluency is defined as the speaker's ability "to efficiently mobilize and integrate the underlying cognitive processes responsible for producing utterances" (p. 48). Utterance fluency deals with the temporal features of utterances that "reflect the speaker's cognitive fluency" (p. 52). Perceived fluency refers to the "interferences listeners make about a speaker's cognitive fluency based on their perception of utterance fluency" (p. 48). Although fluency is not well understood and there are many definitions associated with it (Koponen and Riggensbach 2000), utterance fluency, represented by temporal variables, is thought to be the most readily measurable aspect of fluency (Segalowitz 2010; Tavakoli et al. 2020). This study thus focuses on utterance fluency and its manifested temporal measures.

2.2 *Temporal Measures of Fluency*

A variety of temporal variables has been developed and used to measure utterance fluency. Skehan (2003) and Tavakoli and Skehan (2005) have classified these variables into three categories: speed fluency, breakdown (dis)fluency, and repair (dis)fluency. Speed fluency is concerned with the rate and density of delivery, breakdown (dis)fluency focuses on the nature of pauses, and repair (dis)fluency deals with the repetitions and false starts that are used to repair speech during production. Following this classification, the section below reviews commonly used features under the three major categories of fluency measures.

2.2.1 **The Amount and Rate of Speech**

The most commonly used measure regarding the amount of speech is the phonation time ratio or speech time ratio. It is the ratio of the time spent speaking to the total response time, which includes speaking time and pausing time. A review of the literature finds that there are three most frequently reported rate measures: speech rate,

articulation rate, and mean length of run. These variables are calculated based on the number of syllables (or words, morphemes) produced per second (or per minute, per speech segment). Speech rate and articulation rate differ in their inclusion or exclusion of silent pauses. Mean length of run is the number of syllables produced between two silent pauses. When counting the total number of syllables, some researchers excluded self-corrected and repeated syllables and computed the rate measures based on pruned syllables (Lennon 1990; Iwashita et al. 2008). Compared to L1 speech, L2 speech is often characterized by a smaller phonation time ratio, slower rates, and shorter runs. Differences in these features are associated with varied proficiency levels (Cucchiariini et al. 2000, 2002; Ginther et al. 2010).

2.2.2 Pausing

There are two types of pauses that are commonly measured: silent or unfilled pauses and filled pauses. Researchers usually take measurements of the number and duration of silent and filled pauses. In the literature, there are disagreements about the cut-off point for silent pauses. The cut-off point can be as short as 0.1 s (Trofimovich and Baker 2006), to as long as 3 s (Fulcher 1996). Zellner (1994) suggested that pauses are more easily perceived if their duration is between 0.2 and 0.25 s. De Jong and Bosker (2013) found that a cut-off point for silent pauses of 0.25–0.3 s led to the highest correlation between the number of silent pauses and L2 proficiency levels. Therefore, in this study, any silence equal to or longer than 0.25 s is identified as a silent pause.

Filled pauses, or fillers, can be classified into non-lexical fillers and lexical fillers. Non-lexical fillers are sounds such as “uh” or “um” that “are not recognized as words and contain little or no semantic information” (Riggenbach 1991, p. 426). Lexical fillers are sounds that “are recognized as words, but in context contribute little or no semantic information” (p. 426), such as “you know” and “I mean”.

While pausing is a normal phenomenon that occurs in both L1 and L2 speech (Goldman-Eisler 1968), the frequency, length, and location varies (Chambers 1997). Despite cross-linguistic differences, L2 learners tend to pause more frequently and pause longer than native speakers (Riazantseva 2001). In addition, among L2 learners, lower-proficiency learners are more likely to have non-juncture pauses, i.e., pauses occurring at unpredictable places such as within constituents or between every word, whereas higher-proficiency learners tend to produce more juncture pauses, i.e., pauses occurring at syntactic boundaries (Hawkins 1971).

2.2.3 Repairs

Repairs are defined as the online modification of utterances (Tavakoli and Skehan 2005). Typical measures include the number of repetitions and the number of corrections, false starts, or reformulations. The number of repetitions is the number of partially or exactly repeated syllables (or words, phrases). Freed (2000) suggested

that a correction (or a grammatical repair) corrects a structural feature, while a false start “suggests a decision to rephrase because the speaker perceives the best form for the intended meaning was not originally selected” (p. 248). Although many classifications have been made on the types of repairs, some of these measures have often overlapped with one another. For example, a false start can be used to repair a grammatical structure. Therefore, the most of the researchers did not differentiate among the subcategories of repairs and only reported the total number of repairs or disfluencies (e.g., Cucchiari et al. 2000, 2002; Iwashita et al. 2008). In previous research, the frequency of repair features has not demonstrated a clear relationship with L2 proficiency levels. However, it has been suggested that L2 learners tend to use more corrections and repetitions than L1 speakers (Kahng 2014), and higher-proficiency learners are able to restart more quickly (i.e., “a smaller part of the original utterance is rejected before the restart”) than lower-proficiency learners (Rohde 1985, as cited in Riggensbach 1991, p. 427).

2.3 *Studies on Fluency*

2.3.1 **General Studies on Fluency**

Utilizing the above-mentioned fluency measures, two strands of fluency studies can be found in SLA research. The first strand investigates how fluency in oral production is affected by different conditions, such as task types (Skehan and Foster 1999), planning time (Mehnert 1998; Yuan and Ellis 2003), study abroad context (Freed 1995, 2000; Möhle 1984; Lennon 1990; Towell et al. 1996), and learning mode (Blake 2006). The second strand of fluency studies examines fluency features in relation to perceived fluency and proficiency ratings. Since this study focuses on the relationships between fluency features and different levels of oral proficiency (scores received on a test), the studies reviewed here are mainly concerned with the second strand.

Many studies have examined what temporal measures could contribute to perceived fluency ratings. Riggensbach (1991) asked twelve English instructors to rate six ESL students’ dialogues as either fluent or non-fluent. She found that through speech rate and the number of silent pauses she could distinguish fluent from non-fluent learners, but not with features of repairs and filled pauses. Cucchiari et al. (2000) examined to what extent fluency measures obtained by an automatic speech recognizer could predict expert ratings of fluency on read speech produced by 60 L2 learners of Dutch. The results showed that automatic fluency measures of rates and silent pauses were significantly correlated with expert fluency ratings. In 2002, Cucchiari and her colleagues expanded their 2000 study and included a set of spontaneous speech produced by 57 L2 learners. This time they found that articulation rate and the mean length of silent pauses barely correlated with expert fluency ratings, and the strength of the correlations with other rate and pausing features was also weaker with regard to spontaneous speech than with read speech (Cucchiari et al.

2002). In both studies, the number of filled pauses and repairs only showed weak correlations with fluency ratings. In addition to native experienced raters, Kormos and Dénes (2004) included nonnative experienced teachers in the fluency rating of speech samples collected from 16 Hungarian ESL learners. Similarly, speech rate, mean length of run, phonation time ratio, and the mean length of silent pauses were found to be important predictors of fluency ratings, and the number of filled pauses and the number of disfluencies had almost no relations to fluency ratings. However, the articulation rate and the number of silent pauses were not found to influence the judgments of fluency. Bosker et al. (2013) examined the relative contributions of speed, breakdown, and repair fluency measures to perceived fluency ratings. Besides the significant roles of speed and silent pausing measures in predicting perceived fluency ratings, they also found that repair fluency made a small but significant contribution to the perception of fluency.

A number of studies have also explored the relationships between temporal measures of fluency and performance scores assigned by raters on an actual test or an experimental task. Iwashita et al. (2008) analyzed 200 ESL test-takers' responses to five TOEFL iBT speaking tasks and investigated to what extent fluency features could distinguish different levels of oral performance. The results showed that speech rate, the number of silent pauses, and the total pausing time could distinguish the levels overall, but they failed to consistently separate adjacent levels. Based on 150 speech samples from a local oral English proficiency test, Ginther et al. (2010) found that all of the rate and silent pausing features could predict ESL test-takers' oral scores, but that none of the filled pauses and repair features were related to the scores. Baker-Smemoe et al. (2014) examined fluency features of excerpts taken from 86 ACTFL OPIs in five different languages, including Arabic, French, German, Japanese, and Russian. In a similar vein, speech rate, mean length of run, and the number and length of silent pauses were found to significantly correlate with the ACTFL proficiency levels. With 32 ESL learners' test performances on four different tasks of the British Council's Aptis Speaking Test, Tavakoli et al. (2020) observed that all the rate features and the mean length of silent pauses could distinguish two or more of the Common European Framework of Reference for Languages (CEFR) levels. Interestingly, they found that the frequency of silent pauses did not show significant differences across levels, but the frequency of filled pauses and repairs did. They also observed a lack of consistent differences between adjacent levels.

All these studies have suggested that the best predictor of fluency appears to be speech rate (e.g., Riggenbach 1991; Iwashita et al. 2008). Mean length of run and phonation time ratio were also found to be reliable predictors (e.g., Cucchiarini et al. 2000, 2002; Kormos and Dénes 2004). However, mixed results have been observed for other variables. A number of studies revealed that the number and length of silent pauses could predict fluency and proficiency ratings (e.g., Baker-Smemoe et al. 2014; Ginther et al. 2010). However, some studies found that only one of these features significantly correlated with fluency and proficiency ratings. For example, Kormos and Dénes (2004) and Tavakoli et al. (2020) reported that fluency ratings did not strongly correlate with pause frequency, but did correlate with pause length. In contrast, Cucchiarini et al. (2002) found that fluency ratings were significantly

correlated with pause frequency, but not with pause length. Aligned with Lennon's (1990) proposal that some of the fluency measures may be "peripheral", most of the studies found that filled pausing and repair features had less influence in predicting fluency and proficiency ratings (e.g., Ginther et al. 2010; Kormos and Dénes 2004). However, Bosker et al. (2013) and Tavakoli et al. (2020) observed that the number of repairs could predict proficiency ratings. No consensus has been reached on the best set of fluency features in predicting and distinguishing L2 oral proficiency levels. Hence, more research is needed to address this question.

2.3.2 Studies on L2 Chinese Fluency

Compared to the great volume of fluency research conducted about learners of ESL and other languages, the number of studies on fluency in L2 Chinese is rather small, and the scope of the research is also limited. The majority of the fluency studies on L2 Chinese deal with a small number of participants within a restricted range of proficiency levels or with a limited number of fluency features. For example, for the sake of sampling convenience, some studies only examined fluency in performances by L2 learners in a single Chinese language class. Zhai (2011) investigated fluency features of oral exams taken by ten L2 learners in an intermediate Chinese class. Although these students' oral performances were evaluated by a total of twelve Chinese instructors, the author only compared fluency features of responses by three students, whom she classified into the low, middle, and high groups according to their performance scores. Zhai and Feng (2014) compared the oral responses to the same picture narration task produced by twelve L2 learners in a beginning Chinese class over a two-month period. Although they found that students made significant improvements in speech rate and the mean length of silent pauses, these results are difficult to be generalized to higher-proficiency levels of L2 Chinese. Studies using a larger sample size were conducted by Shih and Wu (2011) and by Liu and Wu (2016), but they were only concerned with advanced learners. Shih and Wu (2011) asked 43 native speakers of Chinese to assign fluency ratings to snippets of natural conversations conducted by 48 speakers (11 natives and 37 L2 learners) in advanced Chinese classes. They found that speech rate and the number of vowels/syllables produced were strong predictors of perceived fluency among native listeners. Liu and Wu (2016) directed six L2 learners from the same advanced Chinese class and three native speakers to complete a monologue task. A total of 60 native listeners provided fluency ratings for the speech samples. Speech rate, mean length of run, and the mean length of silent pauses were found to be significantly correlated with fluency ratings, but the number of filled pauses and disfluencies showed weak correlations. Chen (2012) observed similar findings in her comparisons of the oral performances produced by three proficiency groups (16 intermediate learners, 16 advanced learners, and 16 native speakers). Additionally, she found nonsignificant differences in the number of silent pauses across proficiency groups. However, she did not include a beginning level in her study.

A number of studies have examined the relationships between fluency features and proficiency ratings in L2 Chinese. However, in those studies, only a few fluency features were explored. Adapting the tasks in Iwashita et al. (2008) to Chinese, Jin and Mak (2013) examined oral performances produced by 66 L2 learners of Chinese from advanced speaking courses. However, because they centered their study on features of complexity, accuracy, and fluency (CAF), only two fluency features were investigated. Their study revealed that speech rate and total pausing time had moderate correlations with holistically rated performance scores. Ye (2015) investigated the relationships among CAF features of speech samples produced by 45 L2 learners and their proficiency scores obtained from an HSK test. Based on the HSK test scores, participants were grouped into three proficiency groups. Speech rate and the length of silent pauses demonstrated significant and consistent differences across groups. Chen (2015) conducted a similar study of 29 L2 learners (12 intermediate and 17 advanced) and 15 native speakers of Chinese. Among CAF features, mean length of run and mean length of silent pauses were found to be significantly different across proficiency groups.

Based on these studies, it is difficult to make inferences about how fluency and disfluency features are manifested at different levels of oral performance in L2 Chinese, from low through advanced levels. Therefore, this study recruited L2 learners at different curricular levels in a typical university-level Chinese language program and explored a wide array of temporal measures of fluency, in an effort to find out to what extent these features differ across the levels of performance on an oral narrative task in Chinese, and whether these features can distinguish various levels. Specifically, the study aims to address the following research questions:

- (1) What are the relationships between fluency features of oral performances by L2 Chinese learners and their levels of oral proficiency (score levels)?
- (2) What fluency features can distinguish different levels of oral proficiency in L2 Chinese?

3 Methodology

3.1 Instrument

To elicit oral performance from L2 learners of Chinese, an oral narrative task was created. The oral narrative task was used, as it can elicit a paragraph-length discourse for analysis. The task asked participants to discuss an interesting or memorable trip they had taken. Each participant had 30 s to prepare and two minutes to respond. Paradigm (2007), a software program designed to run psycholinguistic experiments, was used to deliver the test.

3.2 Participants

A total of 38 L2 learners (15 females and 13 males) were recruited from the four curricular levels in the Chinese language program at a large US university. There were six from first-year Chinese, 12 from second-year Chinese, nine from third-year Chinese, and 11 from fourth-year Chinese. The majority of the participants are native speakers of English ($n = 33$); the remainder included native speakers of Korean ($n = 2$) and Cantonese ($n = 3$). Their average age was 21 (min = 18, max = 37, SD = 4.06).

3.3 Procedure

All the participants were tested individually in a phonetics laboratory at the university. Each participant wore a head-mounted microphone and sat in front of a computer. First, they were shown a topic on a computer screen. They were given 30 s to prepare for the topic. Then, they were prompted to start talking after hearing a beeping sound. They had two minutes to respond. A self-introduction task was used as a trial to assist them in familiarizing themselves with the procedure. The whole experiment was self-paced. Participants were encouraged to speak as much as possible, but they were also allowed to end the task before time expired. After finishing the task, they completed a language background questionnaire. Participants' responses were recorded by Paradigm (2007) and saved as .wav files (22 kHz, 16-bit sound).

3.4 Rating

Two experienced Chinese instructors rated all the oral responses by following a four-level holistic rating scale (see Appendix). The two instructors each had more than four years of teaching experience in college-level Chinese. The courses they taught involved weekly assessments of students' oral performances throughout the semester. The inter-rater reliability (Cohen's Kappa) was 0.85. Any disagreements between the two raters were resolved through discussion. Only one rating was used for later analysis.

3.5 Analyses

3.5.1 Fluency and Disfluency Features

All of the oral responses produced by the participants were manually transcribed and annotated using Praat (version 6.0.11, Boersma and Weenink 2016). Pause boundaries were marked and information was noted with regard to whether the sounds were intelligible syllables, fillers, or repairs. These were saved in time-aligned TextGrids. An additional coder coded 10% of the data and the inter-coder reliability reached 0.99.

A Python (version 3.6.3) script was developed to extract and compute fluency features from the TextGrids. Fluency measures were calculated from the entire speech sample elicited from each participant. The following list presents all the fluency and disfluency features examined in this study, as well as their various operations.

The amount and rate of speech

- (a) Phonation time ratio: The ratio of speech time to the total response time. Speech time is the total response time, excluding silent pausing time.
- (b) Speech rate: Total number of syllables divided by the total response time. It is presented as syllables per second.
- (c) Articulation rate: Total number of syllables divided by speech time. It is also presented as syllables per second.
- (d) Mean length of run: Total number of syllables divided by number of runs. A run is defined as a speech segment occurring between two silent pauses. A silent pause is a silence equal to or longer than 0.25 s.

Pausing

- (a) Number of silent pauses: Normalized number of silent pauses. Normalization was performed by dividing the total number of syllables by 100. It is presented as the number of silent pauses per 100 syllables.
- (b) Mean length of silent pauses: Silent pausing time divided by the total number of silent pauses.
- (c) Number of non-lexical fillers: Normalized number of non-lexical fillers. Non-lexical fillers are sounds such as “um” and “uh”.
- (d) Mean length of non-lexical fillers: Non-lexical filler time divided by the total number of non-lexical fillers.
- (e) Number of lexical fillers: Normalized number of lexical fillers. Examples of lexical fillers in Chinese are “zhège (这个)”, “nàge (那个)”, and “ránhòu (然后)”.

Repairs

- (a) Number of reformulations: Normalized number of syllables being abandoned due to false starts or self-corrections.

- (b) Number of self-repetitions: Normalized number of syllables being self-repeated due to disfluencies. Partial or exact repetitions intended for pronunciation corrections were grouped into reformulations.

3.5.2 Statistical Analyses

The measured fluency values were imported to R (version 3.6.1) to run all the statistical analyses. To address the first research question, a Spearman rank-order correlation test was computed among fluency features and score levels. The Spearman rank-order correlation test was selected as the rating scale is ordinal. To answer the second research question, a series of one-way analysis of variance (ANOVA) tests were performed to explore the extent to which the fluency features were different across score levels.¹

4 Results

4.1 Correlations Between Fluency Features and Proficiency Levels

A Spearman rank-order correlation test was conducted to examine the relationships between the measured fluency features and participants' received proficiency scores. The results are presented in Table 1.

All the amount and rate features of speech showed strong and positive correlations with proficiency scores. Among them, speech rate had the strongest correlation with score levels ($r = 0.81, p < 0.001$) followed by articulation rate ($r = 0.77, p < 0.001$), mean length of run ($r = 0.69, p < 0.001$), and phonation time ratio ($r = 0.66, p < 0.001$). Both the mean length of silent pauses ($r = -0.74, p < 0.001$) and the number of silent pauses ($r = -0.68, p < 0.001$) were strongly and negatively correlated with score levels. As for the filled pauses, the number and length of non-lexical fillers demonstrated very weak correlations with score levels ($r = -0.18, p = 0.272$ for the mean length of non-lexical fillers, and $r = -0.17, p = 0.306$ for number of non-lexical fillers). However, it is interesting to observe that the number of lexical fillers was positively and significantly correlated with proficiency scores ($r = 0.40, p = 0.013$). All the repair features had almost no relations to score levels ($r = -0.08, p = 0.647$ for the number of reformulations and $r = -0.10, p = 0.547$ for the number of self-repetitions).

¹Shapiro–Wilk normality tests were conducted to check whether the normality assumption was met for ANOVA tests. The number of lexical fillers did not reach the statistical threshold for normal distribution. Log and square root transformations were attempted, but none of them resulted in a better approximation of normal distribution. Therefore, the original data was retained for analysis, and the result was interpreted cautiously. Nonetheless, the ANOVA result aligned with the visual representation of the boxplot.

Table 1 Correlation matrix

	Score	PTR	SR	AR	MLR	MLSP	#SP	MLNF	#NF	#LF	#Ref	#Rep
Score	1											
PTR	0.66**	1										
SR	0.81**	0.90**	1									
AR	0.77**	0.54**	0.82**	1								
MLR	0.69**	0.88**	0.92**	0.71**	1							
MLSP	-0.74**	-0.85**	-0.87**	-0.66**	-0.69*	1						
#SP	-0.68**	-0.87**	-0.91**	-0.71**	-1.00*	0.67**	1					
MLNF	-0.18	-0.13	-0.15	-0.13	-0.06	0.10	0.09	1				
#NF	-0.17	-0.33*	-0.20	0.00	-0.31	-0.05	0.32*	0.37*	1			
#LF	0.40*	0.34*	0.35*	0.30	0.44**	-0.27	-0.44**	0.06	-0.09	1		
#Ref	-0.08	-0.08	-0.04	-0.03	-0.02	0.03	0.03	0.15	0.18	-0.05	1	
#Rep	-0.10	-0.46**	-0.41*	-0.24	-0.47**	0.24	0.47**	0.03	0.27	-0.09	-0.04	1

Note ** $p < 0.01$; * $p < 0.05$. PTR = phonation time ratio, SR = speech rate, AR = articulation rate, MLR = mean length of run, MLSP = mean length of silent pauses, #SP = number of silent pauses, MLNF = mean length of non-lexical fillers, #NF = number of non-lexical fillers, #LF = number of lexical fillers, #Ref = number of reformulations, and #Rep = number of self-repetitions

These results indicate that L2 learners at higher score levels spent more time speaking about the task than did learners at lower score levels. They also spoke faster and produced longer chunks of speech. While higher-level learners did not differentiate from lower-level learners in the non-lexical fillers they produced, higher-level learners tended to use more lexical fillers to buy time for thinking and in the meantime keep speech flowing. In contrast, lower-level learners paused in silence more frequently and for a longer period of time, producing hesitant and choppy speech. Given that higher-level learners may attempt to produce more complex ideas and maintain extended discourse, they did not reformulate or self-repeat less than lower-level learners. Overall, the results demonstrated the expected correlation patterns between fluency measures and score levels. The amount and rate features of speech had significant positive correlations with score levels, while silent pausing features demonstrated significant negative relationships. The filled pauses and repair features showed weak or no correlations with score levels.

4.2 *Distinguishing Fluency and Disfluency Features*

A series of one-way ANOVAs were conducted to explore the extent to which fluency measures were differed across score levels. Descriptive statistics of fluency measures are provided. Boxplots showing means with bootstrapped 95% confidence intervals (CI) are presented to corroborate significant results of level differences found in ANOVAs.

The amount and rate features of speech

Descriptive statistics of the amount and rate features of speech are presented in Table 2. A clear increasing trend was observed in all measures across score levels. The ANOVA results showed that the differences were statistically significant: phonation time ratio, $F(3, 34) = 18.02$, $p < 0.001$, $\eta^2 = 0.61$; speech rate, $F(3, 34) = 36.76$, $p < 0.001$, $\eta^2 = 0.76$; articulation rate, $F(3, 34) = 15.72$, $p < 0.001$, $\eta^2 = 0.58$; mean length of run, $F(3, 34) = 24.5$, $p < 0.001$, $\eta^2 = 0.68$. Post-hoc analyses revealed that Level 4 was significantly different from Level 3 with regard to all the amount and rate features except articulation rate, and Level 3 was distinct from Level 2 in all measures. These findings were confirmed by the non-overlapping confidence intervals (CIs), as shown in Figs. 1, 2, 3, and 4. However, Level 1 and 2 were not statistically different from each other.

Pausing

Table 3 presents descriptive statistics of pausing features across score levels. A decreasing trend was observed for silent pausing features, but the pattern was not clear for filled pauses. The ANOVAs showed significant results for the two silent pausing measures across score levels: the number of silent pauses, $F(3, 34) = 9.04$, $p < 0.001$, $\eta^2 = 0.44$, and the mean length of silent pauses, $F(3, 34) = 10.47$, $p < 0.001$, $\eta^2 = 0.48$. Post-hoc analyses revealed that participants at Level 4 paused

Table 2 Descriptive statistics of the amount and rate features across score levels

Fluency features	Score	<i>N</i>	Mean	SD	Min	Max
Phonation time ratio	1	8	0.41	0.10	0.23	0.53
	2	14	0.42	0.09	0.27	0.56
	3	10	0.54	0.08	0.37	0.63
	4	6	0.70	0.09	0.52	0.76
Speech rate	1	8	1.00	0.32	0.61	1.38
	2	14	1.17	0.31	0.59	1.72
	3	10	1.78	0.30	1.26	2.12
	4	6	2.65	0.43	2.14	3.17
Articulation rate	1	8	2.43	0.40	1.80	2.97
	2	14	2.80	0.43	2.01	3.62
	3	10	3.33	0.39	2.75	3.97
	4	6	3.80	0.44	2.99	4.17
Mean length of run	1	8	2.52	0.92	1.19	3.62
	2	14	2.57	0.64	1.53	3.89
	3	10	4.00	1.19	2.62	6.28
	4	6	6.52	1.52	3.96	8.40

Fig. 1 Boxplot for phonation time ratio

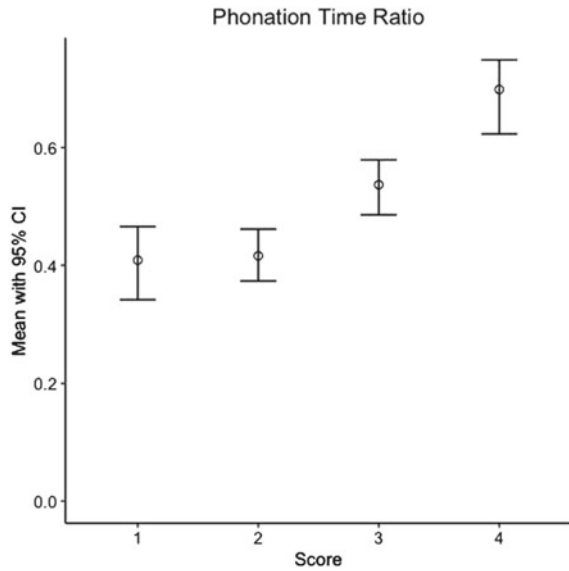


Fig. 2 Boxplot for speech rate

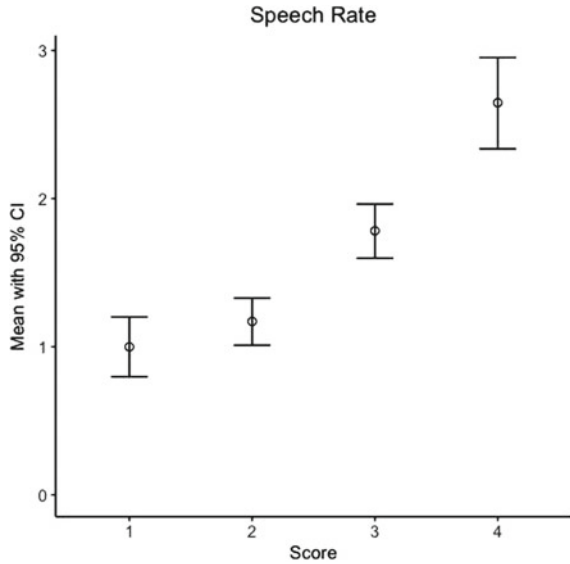
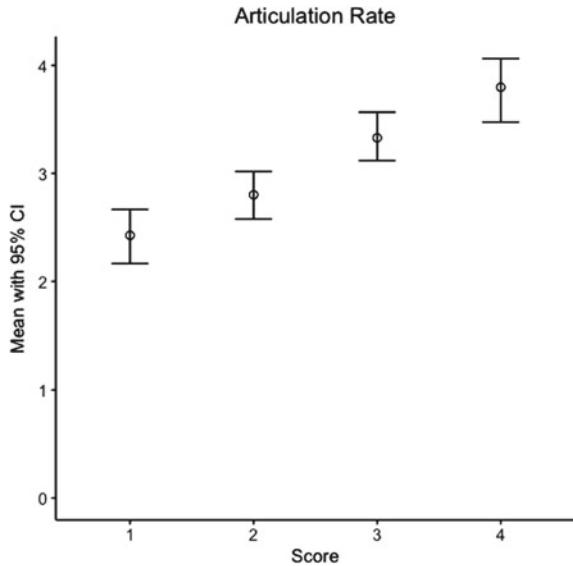


Fig. 3 Boxplot for articulation rate



significantly less and for shorter durations than did those at Level 3, which also was the case between Levels 3 and 2. Again, Level 2 and Level 1 were not statistically different from each other (see Figs. 5 and 6). As for the filled pauses, no significant results were observed for the non-lexical filler measures: the number of non-lexical fillers, $F(3, 34) = 0.94, p = 0.432, \eta^2 = 0.08$, and the mean length of non-lexical fillers, $F(3, 34) = 0.09, p = 0.967, \eta^2 = 0.007$. There was a significant main effect

Fig. 4 Boxplot for mean length of run

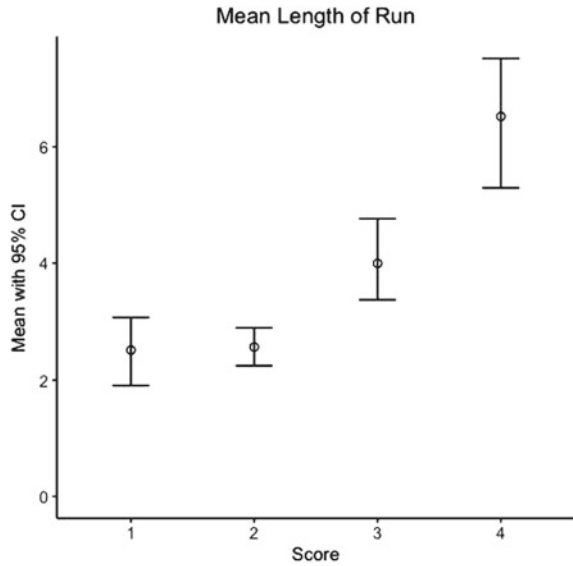


Table 3 Descriptive statistics of pausing features across score levels

Fluency features	Score	N	Mean	SD	Min	Max
Number of silent pauses	1	8	44.22	19.75	25.53	82.26
	2	14	40.32	11.03	25.00	63.04
	3	10	26.49	7.66	15.54	37.50
	4	6	15.79	4.47	11.61	24.18
Mean length of silent pauses	1	8	1.37	0.26	1.00	1.72
	2	14	1.19	0.47	0.59	2.16
	3	10	0.78	0.15	0.61	1.09
	4	6	0.51	0.17	0.39	0.83
Number of non-lexical fillers	1	8	14.42	14.03	0.00	45.16
	2	14	16.13	10.19	0.00	32.61
	3	10	13.76	5.24	7.05	20.48
	4	6	8.32	5.41	4.36	18.87
Mean length of non-lexical fillers	1	8	0.39	0.18	0.00	0.60
	2	14	0.41	0.17	0.00	0.69
	3	10	0.41	0.08	0.26	0.55
	4	6	0.38	0.05	0.32	0.45
Number of lexical fillers	1	8	0.36	0.76	0.00	2.13
	2	14	0.22	0.46	0.00	1.28
	3	10	0.65	1.24	0.00	4.03
	4	6	1.65	1.54	0.00	3.97

Fig. 5 Boxplot for number of silent pauses

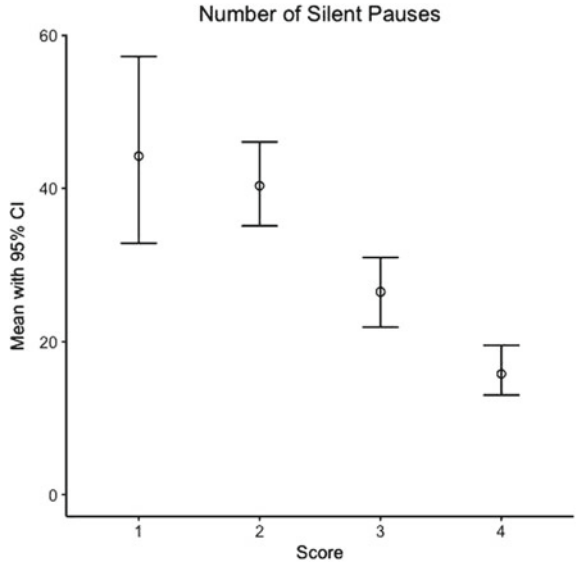
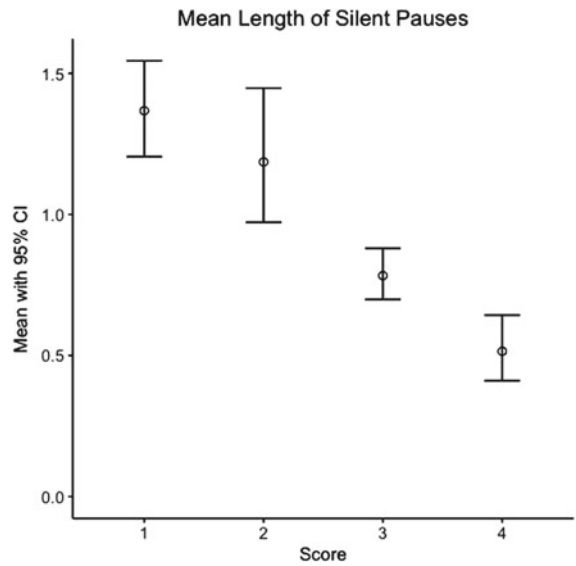
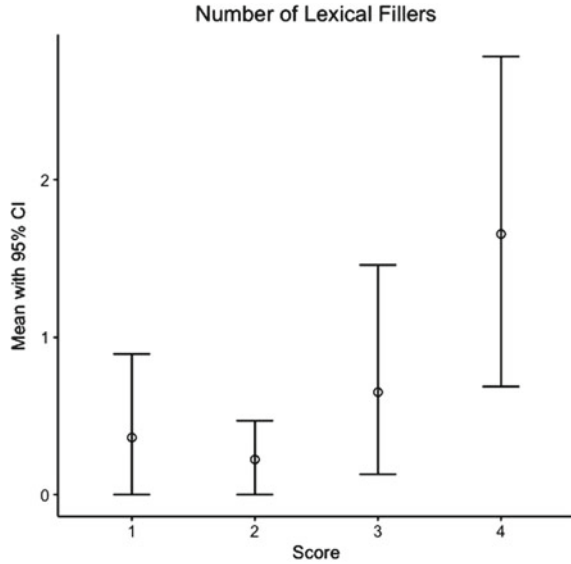


Fig. 6 Boxplot for mean length of silent pauses



of score level on the number of lexical fillers, $F(3, 34) = 3.18, p = 0.036, \eta^2 = 0.22$. However, a significant difference was observed between Level 2 and Level 4—not between adjacent levels (see Fig. 7).

Fig. 7 Boxplot for number of lexical fillers



Repair features

Descriptive statistics of repair features are listed in Table 4. Unlike the other two categories of fluency features, repair features did not show a clear increasing or decreasing trend across score levels. The ANOVA results confirmed that there were no significant level differences in the number of reformulations, $F(3, 34) = 1.63, p = 0.201, \eta^2 = 0.13$, or the number of self-repetitions, $F(3, 34) = 0.57, p = 0.64, \eta^2 = 0.05$.

Table 5 summarizes the findings of the statistical analyses applied to this section. There are distinctive differences among score levels in the features of the amount and rate of speech, silent pausing, and lexical fillers. For the separation of adjacent levels, Level 3 could be distinguished from Level 2 by all the amount and rate features of

Table 4 Descriptive statistics of repair features across score levels

Fluency features	Score	<i>N</i>	Mean	SD	Min	Max
Number of reformulations	1	8	5.10	3.4	0.00	9.52
	2	14	2.60	2.66	0.00	8.70
	3	10	2.87	1.46	0.88	5.71
	4	6	3.63	3.23	0.85	9.89
Number of self-repetitions	1	8	3.37	4.49	0.00	12.73
	2	14	3.54	2.37	0.00	8.94
	3	10	2.73	1.94	0.40	6.04
	4	6	1.87	2.14	0.00	5.84

Table 5 Summary of statistical analyses

Categories	Fluency measures	Score level
The amount and rate of speech	Phonation time ratio	4 > 3 > 2 = 1
	Speech rate	4 > 3 > 2 = 1
	Articulation rate	4 = 3 > 2 = 1
	Mean length of run	4 > 3 > 2 = 1
Pausing	Number of silent pauses	4 < 3 < 2 = 1
	Mean length of silent pauses	4 < 3 < 2 = 1
	Number of non-lexical fillers	No difference
	Mean length of non-lexical fillers	No difference
	Number of lexical fillers	4 > 2
Repairs	Number of reformulations	No difference
	Number of self-repetitions	No difference

Note “=” represents no significant difference, “>” represents significantly greater, “<” represents significantly less

speech, as well as by silent pausing features. Additionally, Level 4 could be separated from Level 3 by the same set of features excluding articulation rate. However, Level 2 and Level 1 were not statistically different from each other.

5 Discussion

The present study has attempted to find out how fluency and disfluency features differ and develop at different levels of oral performance in L2 Chinese. First, correlation results showed that all of the amount and rate features of speech had significant positive relationships with score levels, and the silent pausing features had significant negative relationships. These results are consistent with previous studies on relationships between fluency measures and proficiency levels (Baker-Smemoe et al. 2014; Ginther et al. 2010). The results also confirmed that there might be a distinction between primary and secondary fluency variables. The primary fluency features tend to correlate significantly with proficiency levels (Cucchiari et al. 2000; Lennon 1990). Those features include phonation time ratio, speech rate, articulation rate, mean length of run, and the number and length of silent pauses. On the other hand, the secondary fluency features, including filled pauses and repairs, are not always

present in speech production and tend not to demonstrate strong correlations with proficiency levels. However, because few studies have made a distinction between lexical fillers and non-lexical fillers when measuring filled pauses, it is interesting to observe in this study that learners with higher-proficiency scores produced more lexical fillers than did learners with lower-proficiency scores. Raupach (1984), in his study of formulae as fillers, found that due to the lack of a sufficient command of lexicalized fillers, lower-proficiency learners often used non-idiomatic sounds to fill pauses, such as “uh” and “um”. In contrast, higher-proficiency learners used idiomatic expressions to give themselves time to plan and continue their speech flow. These differences were observed in the use of lexical fillers in the present study. Although the number of lexical fillers could not consistently distinguish across score levels and the variability at higher levels was also comparatively large, the positive relationship with score levels suggests the importance of the use of lexical fillers in achieving a higher level of oral proficiency. With regard to repair features, only very weak correlations with score levels were observed. While learners at higher-proficiency levels may constantly monitor and repair their speech as they perceive the best form with which to express themselves when attempting to produce complex ideas, learners at lower levels have difficulty in retrieving simple lexical items or constructing basic structures, thus, repeating and restarting several times (Fulcher 1996; Kahng 2014). The sentences below are examples of repairs made by a Level 4 student and a Level 1 student. Words in brackets are repairs, and a silent pause is marked by a slash.

[我就] uh 我觉得[很有]/[很] uh, uh收获很大。(Level 4)

[I then] uh, I felt [was very]/[very] uh, uh (I) learned a lot.

[我吃很多] uh /[我吃很多好吃]/[我吃很多好吃饭]。(Level 1)

[I eat many] uh/[I eat many good eat]/I eat many good eat meal.

The student at Level 4 made three repairs as she attempted to express the idea that she learned a lot during the trip, which required advanced vocabulary and knowledge of structure. The student at Level 1 intended to say that she had many good dishes. She repaired twice, but none of her repairs were successful. The final sentence she produced contained more than one grammatical error (missing articles “le” and “de”, and misusing “饭”). This might explain the nonsignificant differences in the number of reformulations and self-repetitions that participants made across score levels. It is the quality of the repairs rather than their quantity that made a difference among score levels. The examples also seem to provide evidence for the previously noted observation that higher-proficiency learners are able to restart more quickly than are lower-proficiency learners (Rohde 1985). The student at Level 4 only rejected a small part of the sentence before she realized the correct expression, whereas the student at Level 1 repeated a larger part of the original sentence. A more detailed qualitative analysis of such repairs may reveal more consistent differences across proficiency levels.

In addition, the present study found that the amount and rate features of speech, as well as the silent pausing features, could distinguish two or more levels of oral proficiency in L2 Chinese. There were more significant differences in higher levels

than was the case for lower levels. This is shown in Table 5: Levels 2 and 3 could be distinguished by all of the amount, rate, and silent pausing features, and Level 4 could be separated from Level 3 by all of those features except articulation rate, but none of them could separate Level 1 from Level 2. This was in line with what Iwashita et al. (2008) and Baker-Smemoe et al. (2014) found in their studies, namely, that performances at lower levels were less distinctive. These findings suggest that lower-proficiency learners may resemble one another in degrees of disfluency, and there might be a threshold of fluency in order for L2 learners to achieve a higher level of oral proficiency. It is also possible that lower-level performances may be distinguished by other micro-level fluency features, such as non-juncture pauses or disfluency clusters (Riggenbach 1991; Kahng 2014). An investigation of micro-level fluency features across different proficiency levels is worthy of exploration. Additionally, the results aligned with previous studies of distinguishing features, in that no single fluency feature was able to consistently distinguish across proficiency levels (Kang and Yan 2018; Tavakoli et al. 2020). This lack of sensitivity to separate adjacent levels might indicate that temporal measures of fluency only partially characterize oral performance at each proficiency level. Oral performances at adjacent levels might be distinguished in other aspects of proficiency, such as accent, intonation, accuracy, lexical diversity, and grammatical complexity (Derwing et al. 2004; Kormos and Dénes 2004; Riggenbach 1991; Shih and Wu 2011; Wennerstrom 2000). A composite of performance features might provide a clearer picture of the differences observed at each proficiency level.

6 Conclusion

The present study sought to explore how fluency and disfluency features differ and develop at different levels of oral performance in L2 Chinese. The results showed that the amount and rate features of speech, as well as silent pausing features, were significantly correlated with score levels, and these features could also distinguish among score levels. The results suggest that the amount and rate of speech and silent pausing features could be reasonably selected as proxies of fluency.

The findings of the present study have important implications for the teaching and assessment of speaking in L2 Chinese. First, the findings can add to our understanding of the characteristics of speaking fluency in L2 Chinese across different proficiency levels. This can help guide fluency training in the language classroom. For example, the results suggest that fluency training is needed even in a beginning-level classroom, in which fluency often is not prioritized. To achieve a higher level of fluency, communication strategies such as the use of lexical fillers can also be taught in lower-level classes. Moreover, as fluency is an essential criterion of oral assessment, the findings of the study can also help improve scoring rubrics for oral assessment. As filled pauses and repair features did not show strong relationships with proficiency scores, instructors may not have to be overly concerned about such aspects of fluency when evaluating students' oral performances. Last, but not least,

since temporal measures of fluency can easily be detected and built into automatic scoring systems, the findings of the study have provided an empirical basis for the development of automated scoring systems in L2 Chinese.

Although the present study has extended the scope of research on L2 Chinese fluency by examining a more comprehensive set of fluency and disfluency features, as well as a wider range of proficiency levels, the study only analyzed oral performance on an oral narrative task. It has been suggested that task type can affect a learner's oral performance (Cucchiari et al. 2002; Derwing et al. 2004). Therefore, further investigation of a greater number of participants responding to a wider range of speaking tasks is needed.

Appendix

Scoring rubric	
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse
3	The response addresses the task appropriately but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas
2	The response addresses the task, but the development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence occur; meaning may be obscured in places
1	The response is very limited in content and/or coherence or is only minimally connected to the task, or speech is largely unintelligible

Note This scale is adapted from TOEFL iBT independent speaking rubric

References

- Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals*, 47(4), 707–728. <https://doi.org/10.1111/flan.12110>.
- Blake, C. G. (2006). *The potential of text-based Internet chats for improving ESL oral fluency* (Unpublished doctoral dissertation). Purdue University.
- Boersma, P., & Weenink, D. (2016). *Praat: Doing phonetics by computer* (Version 6.0.11) [Computer program]. <https://www.praat.org/>.
- Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, 30(2), 159–175. <https://doi.org/10.1177/0265532212455394>.
- Chambers, F. (1997). What do we mean by fluency? *System*, 25(4), 535–544. [https://doi.org/10.1016/s0346-251x\(97\)00046-8](https://doi.org/10.1016/s0346-251x(97)00046-8).

- Chen, M. (2012). 美国留学生汉语口语产出的流利性研究 [Chinese oral fluency of CSL learners of American English speakers]. *语言教学与研究 [Language Teaching and Linguistic Studies]*, 2, 17–24.
- Chen, M. (2015). 汉语作为第二语言自然口语产出的复杂度、准确度和流利度研究 [Complexity, accuracy and fluency in Chinese as second language acquisition]. *语言教学与研究 [Language Teaching and Linguistic Studies]*, 3, 1–10.
- Cucchiari, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, 107(2), 989–999. <https://doi.org/10.1121/1.428279>.
- Cucchiari, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, 111(6), 2862–2873. <https://doi.org/10.1121/1.1471894>.
- De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In *The 6th workshop on disfluency in spontaneous speech* (pp. 17–20).
- Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, 54(4), 655–679. <https://doi.org/10.1111/j.1467-9922.2004.00282.x>.
- Fillmore, C. J. (1979). On fluency. In D. Hymes, C. J. Fillmore, D. Kempler, & W. S. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85–101). Academic Press.
- Freed, B. F. (1995). What makes us think that students who study abroad become fluent? In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 123–148). Amsterdam: John Benjamin.
- Freed, B. F. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggensbach (Ed.), *Perspectives on fluency* (pp. 243–265). University of Michigan.
- Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13(2), 208–238. <https://doi.org/10.1177/026553229601300205>.
- Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27(3), 379–399. <https://doi.org/10.1177/0265532210364407>.
- Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. Academic Press.
- Hawkins, P. R. (1971). The syntactic location of hesitation pauses. *Language and Speech*, 277–288. <https://doi.org/10.1177/002383097101400308>.
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–473. <https://doi.org/10.1093/applin/amp048>.
- Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 24–49. <https://doi.org/10.1093/applin/amm017>.
- Jin, T., & Mak, B. (2013). Distinguishing features in scoring L2 Chinese speaking performance: How do they work? *Language Testing*, 30(1), 23–47. <https://doi.org/10.1177/0265532212442637>.
- Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, 64(4), 809–854. <https://doi.org/10.1111/lang.12084>.
- Kang, O., & Yan, X. (2018). Linguistic features distinguishing examinees' speaking performances at different proficiency levels. *Journal of Language Testing & Assessment*, 1(1), 24–39. <https://doi.org/10.23977/langta.2018.11003>.
- Koponen, M., & Riggensbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggensbach (Ed.), *Perspectives on fluency* (pp. 5–24). University of Michigan Press.
- Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145–164. <https://doi.org/10.1016/j.system.2004.01.001>.
- Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3), 387–417. <https://doi.org/10.1111/j.1467-1770.1990.tb00669.x>.
- Liu, Y., & Wu, X. (2016). 汉语二语学习者口语产出的流利度研究 [Exploring speaking fluency in the speech of L2 Chinese learners]. *华文教学与研究 [TCSOL Studies]*, 64(4), 32–41.

- Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, 20(1), 83–108. <https://doi.org/10.1017/s0272263198001041>.
- Möhle, D. (1984). A comparison of the second language speech production of different native speakers. In H. W. Dechert, D. Möhle, & M. Raupach (Eds.), *Second language productions* (pp. 26–49). Gunter Narr Verlag.
- Perception Research Systems. (2007). *Paradigm stimulus presentation* [Computer program]. <https://www.paradigmexperiments.com>.
- Raupach, M. (1984). Formulae in second language speech production. In H. W. Dechert, D. Möhle, & M. Raupach (Eds.), *Second language productions* (pp. 114–137). Gunter Narr Verlag.
- Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition*, 497–526. <https://doi.org/10.1017/s027226310100403x>.
- Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, 14(4), 423–441. <https://doi.org/10.1080/01638539109544795>.
- Rohde, L. (1985). Compensatory fluency: A study of spoken English produced by four Danish learners. *Learner Discourse*, 43–69.
- Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.
- Shih, C., & Wu, C. (2011). Evaluating second language fluency. *Proceedings of VLSP, 2011*, 38–41.
- Skehan, P. (2003). Task-based instruction. *Language Teaching*, 36(1), 1–14.
- Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49(1), 93–120. <https://doi.org/10.1111/1467-9922.00071>.
- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). John Benjamins.
- Tavakoli, P., Nakatsuhara, F., & Hunter, A. M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal*, 104(1), 169–191. <https://doi.org/10.1111/modl.12620>.
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84–119. <https://doi.org/10.21832/9781853597688-011>.
- Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, 28(1), 1–30. <https://doi.org/10.1017/s0272263106060013>.
- Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 102–127). University of Michigan.
- Ye, W. (2015). 英语母语者汉语口语水平发展研究 [The oral Chinese language development of native English speakers]. *南京师范大学文学院学报 [Journal of School of Chinese Language and Culture of Nanjing Normal University]*, 4, 170–174.
- Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, 24(1), 1–27. <https://doi.org/10.1093/applin/24.1.1>.
- Zellner, B. (1994). Pauses and the temporal structure of speech. In E. Keller (Ed.), *Fundamentals of speech synthesis and speech recognition* (pp. 41–62). Wiley.
- Zhai, Y. (2011). 口语流利性主观标准的客观化研究 [An objective study on the subjective criteria of speaking fluency]. *语言教学与研究 [Language Teaching and Linguistic Studies]*, 5, 79–86.
- Zhai, Y., & Feng, H. (2014). 基于“看图说话”任务的汉语学习者口语流利性发展研究 [A study of Chinese learners' speaking fluency development with picture description activity]. *华文教学与教育 [TCSOL Studies]*, 56(4), 1–7.

The Role of Vocabulary Knowledge in Second Language Speaking Fluency: A Mixed-Methods Study



Yu Liu

Abstract The present study examined the relationship between L2 learners' task-related lexical access and their utterance fluency. Two groups of American learners of Chinese participated in the study. The first group of learners ($N = 15$) took part in a vocabulary test requiring them to translate 198 words related to four speaking tasks from L1 to L2. Learners' accuracy and reaction time were recorded. Learners then completed four speaking tasks. Six features of their utterance fluency were measured. The second group of learners ($N = 13$) participated in a stimulated recall interview, which was conducted to obtain additional details regarding how lexical access affects utterance fluency in L2 speech. According to the results, significant correlations were found among vocabulary size and all three facets of utterance fluency: speed fluency (speech rate, mean length of runs), breakdown fluency (mean length of silent pauses, number of silent pauses), and repair fluency (number of disfluencies), excluding the number of filled pauses. However, among all fluency measures, only speech rate was significantly correlated to lexical retrieval speed. Moreover, stimulated recall responses revealed that around two-third of the disfluencies were reported to be caused by vocabulary-related issues. The findings confirmed that efficient task-related lexical access was crucial for producing fluent speech in second language.

Keywords Lexical access · Vocabulary size · Lexical retrieval speed · Utterance fluency · Cognitive fluency · Second language

1 Introduction

Speech production is a complex cognitive process, which involves five main stages including message generation, lexico-grammatical encoding, morpho-phonological encoding, phonetic encoding, and monitoring (Levelt 1989, 1999). Compared to first language speech production, second language (L2) speech production appears

Y. Liu (✉)
Brigham Young University, Provo, UT, USA
e-mail: rachelyuliu@byu.edu

to be less automatic and more attention consuming. Noticeable disfluent features can always be found in low proficient L2 learners' speech such as slower speech rate, frequent and longer pauses, frequent repetitions, incomplete sentences, lexical and syntactic errors, and simpler expressions. These features can be categorized into three dimensions: fluency, accuracy, and complexity (Housen and Kuiken 2009).

In terms of second language speaking fluency, previous studies from different disciplines have explored the fluidity nature in L2 speech, trying to answer the following four main questions:

- How is L2 speech different from L1 speech?
- Which temporal features can best predict listeners' judgments on L2 speech?
- What makes L2 speech less fluent?
- How can L2 speaking fluency be improved in a second language classroom?

Some studies investigated the characteristics of L2 learners' speech by measuring its temporal variables and studying their patterns. Studies have found that L1 and L2 speech are different in terms of speed, length of run, and silent pauses. Additionally, silent pauses within a clause were found to be salient in L2 speech (Bosker et al. 2013; Derwing et al. 2009; Kahng 2014). Some studies related objective measures of L2 learners' utterance fluency to listeners' subjective ratings (Bosker et al. 2014; Derwing et al. 2004; Kormos and Dénes 2004; Révész et al. 2016; Riggensbach 1991; Rossiter 2009). In these studies, speech rate and silent pauses were found to be significant predictors of subjective ratings. However, after investigating the correlation between listeners' ratings of fluency and accuracy in L2 speech, Kormos and Dénes (2004) argued that listeners' subjective judgments on fluency may be distracted by their impression on speakers' accuracy. Some other studies attempted to explain the disfluent behaviors in L2 speech from cognitive perspectives (De Jong et al. 2013; Kahng 2014; Segalowitz and Freed 2004; Segalowitz 2010; Towell et al. 1996). It is suggested that L2 learners' speaking fluency is strongly related to their personal speaking style, and linguistic skills (e.g., lexical access speed and efficiency, speed of morpho-syntactic processing) were found to be significant predictors of L2 speaking fluency. Concerning pedagogic suggestions, formulaic instruction was proposed to be an effective way of improving learners' speaking fluency in second language classroom (Boers et al. 2006; Segalowitz 2010; Wood 2002, 2009). Results from previous studies also revealed that repetitive tasks (such as 4/3/2 tasks, namely repeated talks to different partners in four minutes, three minutes and two minutes) would facilitate speaking instruction, therefore improving learners' speaking fluency (De Jong and Perfetti 2011; Gatbonton and Segalowitz 2005). In addition to the above-mentioned tasks, Segalowitz (2016) suggested to include sociolinguistic nature while investigating L2 fluency. He argued that L2 speaking fluency was the outcome of the operation of a dynamical system where cognitive, motivational, social, sociolinguistic, pragmatic, and psycholinguistic considerations interact in complex ways.

Although the existing research has built a solid foundation on describing the nature of L2 speaking fluency, there are only a few explanatory studies focusing on the internal factors that affect learners' fluidity (De Jong et al. 2013; Kahng 2014;

Segalowitz and Freed 2004). Therefore, the present study aims to explore the relationship between vocabulary knowledge and L2 speaking fluency with mixed methods, combining experiment and stimulated recall interview approaches. Specifically, we focus on vocabulary size and lexical accessing speed as vocabulary knowledge. In this chapter, we will first review current literature of the relation between L2 learners' vocabulary and their speaking fluency. We will then present an experiment investigating how task-related L2 lexical access interacted with six measures of L2 utterance fluency. Next, we will explore how various vocabulary-related issues caused disfluencies in L2 speech based on learners' responses in a stimulated recall interview. A discussion of the relation between L2 lexical access and L2 utterance fluency will then follow.

2 Second Language Speaking Fluency

Fluency refers to the fluidity of one's speech. It is a multidimensional construct (Segalowitz 2010: 7). Fluency can be seen as a reflection of listener's subjective judgment on speaker's speaking behavior ('perceived fluency'); it is observable and measurable by features such as speech rate, pausing, hesitation, and repetitions ('utterance fluency'). It can also be explained by identifying mechanisms and processes underlying fluency and disfluency phenomenon ('cognitive fluency'). From the perspective of cognition, fluency is the result of 'the rapid speed, automaticity, and efficiency of the underlying mechanisms' (Rehbein 1987: 104), and thus, fluency reveals 'how efficiently a speaker is able to mobilize and temporally integrate, in a nearly simultaneous way, the underlying processes of planning and assembling an utterance' (Segalowitz 2010: 165).

What makes second language speech less fluent? Kormos (2006: 168) argued that the bilingual production is not significantly different from the one constructed for monolingual speakers (Levelt 1989, 1999) except for two features: (1) the incorporation of L2 concepts, lemmas (word forms), lexemes (syntactic and morphological features), and syllable programs, and (2) a new knowledge store for the declarative knowledge of L2 rules. On the one hand, before L2 knowledge is turned into procedural knowledge from declarative knowledge, processes of L2 production cannot run as parallel as in L1. Lack of automaticity in language processing leads to disfluent phenomenon. On the other hand, fluency issues arise because of learners' limited attentional capacity. When learners focus on processing the language at a specific level due to their incapability in L2, their attention is thus reduced in other areas. L2 speech is less fluent as a result. In addition, in the 'model of bilingual speech production' proposed by Kormos (2006: 168), which is adapted from Levelt's 'blueprint of the monolingual speaker' (1999), Kormos modified the grammatical encoding stage as lexico-grammatical encoding. It revealed the close relationship between morphological encoding and grammatical encoding in language processing. Figure 1 presents the model of L2 speech production based on Levelt's (1999) monolingual

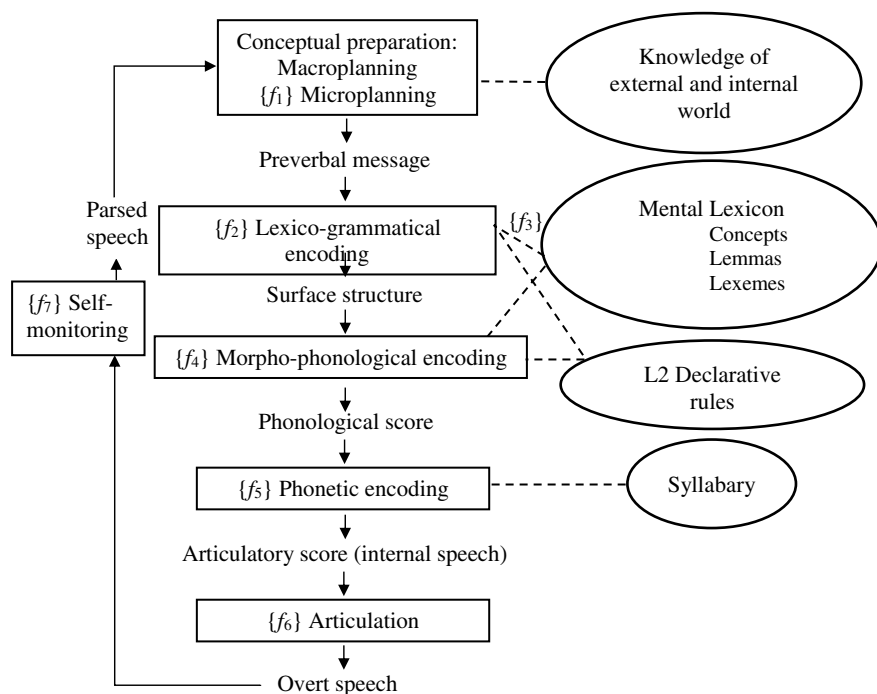


Fig. 1 Model of L2 speech production, adapted from Kormos (2006: 168), Segalowitz (2010: 9), Levelt (1999, Fig. 4.1)

model incorporating Kormos' (2006: 168) modifications regarding the features of bilingual speech processing.

Segalowitz (2010: 8–17) found seven vulnerability points in L2 speech production where underlying processing difficulties could be associated with L2 speech dysfluencies. In Fig. 1, the $\{f\}$ symbols indicate the vulnerability points occur at different stages: microplanning, lexico-grammatical encoding, morpho-phonological encoding, phonetic encoding, articulation, and self-perception. According to Segalowitz (2010), when formulating the preverbal message, learners may not know the L2 lexical items for microplanning (speech preparation), which may have a negative impact on fluency ($\{f_1\}$). At the lexico-grammatical encoding and morpho-phonological level, learners may slow down their speech when they have difficulty in retrieving and utilizing the appropriate lemmas, lexemes ($\{f_2\}$) as well as L2 grammatical rules ($\{f_3\}$) during the formulation of the surface structure. L2 fluency can be compromised when learners do not have automatic access to syllable program ($\{f_4\}$). Fluency issues can arise if learners exert too much effort to attempt to select the appropriate gestural scores ($\{f_5\}$) or execute the scores ($\{f_6\}$). Learners' self-monitoring and self-correction may also slow down the speech or cause pauses or repetitions ($\{f_7\}$).

3 Vocabulary and L2 Cognitive Fluency

In the review of L2 speech production, Kormos (2006) claimed that L2 speech processing is lexically driven. Without knowing the L2 lexical items, learners are unable to prepare the speech to represent the exact message they plan to express. If they fail to retrieve the correct L2 lexicon to match with the preverbal message, then lexico-grammatical encoding, morpho-phonological encoding, and phonetic encoding will be hard to operate. Therefore, an efficient lexical access is regarded as the key to successful speech production.

Lexical access is multifaceted; in the present literature, three categories have been distinguished in measuring lexical access (Anderson and Freebody 1981; Daller et al. 2007; Meara 1996; Milton and Fitzpatrick 2013):

1. *Vocabulary breadth*: the number of words a learner knows regardless of the form they are known in or how well they are known. Vocabulary breadth is also referred to as vocabulary size. Word selection in most studies is based on word frequency bank lists (De Jong et al. 2013; Koizumi and In'nami 2013; Uchihara and Clenton 2020).
2. *Vocabulary depth*: how well or how completely words are known. Vocabulary depth is a rich concept that consists of various aspects. According to Nation's (2001: 27) description of 'what is involved in knowing a word,' vocabulary knowledge includes form (spoken, written, word parts), meaning (form and meaning, concepts and referents, associations), and use (grammatical functions, collocations, constraints on use). Read (2004) proposed that vocabulary involves word form and meaning, as well as associational knowledge, collocation knowledge, inflectional knowledge, and derivational knowledge. Meara and Wolter (2004) extended the vocabulary depth by including knowing the network words. Measuring vocabulary depth is less manageable because it is difficult to find a concept that holds together the variety of elements (Milton 2010).
3. *Vocabulary fluency*: the automaticity with which the words a person knows can be recognized and processed. It is also referred to as processing speed or lexical retrieval speed. Reaction time (RT) is recorded to measure vocabulary fluency in a vocabulary test (De Jong et al. 2013; Koizumi and In'nami 2013).

Since utterance fluency reflects the automaticity of language processing, previous studies tried to relate utterance fluency with cognitive fluency, especially with vocabulary knowledge or lexical access. Segalowitz (2010: 103–106) proposed three important components of L2 cognitive fluency: L2-specific processing speed, processing stability, and processing flexibility. In the previous studies, lexical access has been found to play a key role in L2 speech production (De Jong et al. 2013; Hilton 2008; Koizumi and In'nami 2013; Uchihara and Clenton 2020; Uchihara and Saito 2019). It has also been found that efficient lexical access leads to fluent speech (De Jong et al. 2013; Kahng 2014; Segalowitz and Freed 2004). Segalowitz and Freed (2004) examined the relation between L2 oral fluency—as measured by speech rate, mean run length containing no silent pauses or hesitations greater than 400 ms, mean

length of run without filled pauses ('uhms' and 'uhs'), and longest fluent run—and the speed and efficiency of lexical access. They invited the participants to complete a semantic classification task, requiring them to judge if a word on the computer screen referred to a living or non-living object. The speed of lexical access was indexed by reaction time, while the efficiency of lexical access was indexed by the coefficient of variation of the reaction time. Results showed that lexical access speed and lexical access efficiency were significantly correlated to mean length of run without filled pauses. De Jong et al. (2013) explored the relationship among L2 speaking fluency and vocabulary knowledge and lexical retrieval speed. Utterance fluency was measured by calculating articulation rate, number of silent pauses, mean duration of silent pauses, number of filled pauses, number of corrections, and number of repetitions. For assessing vocabulary knowledge, participants were asked to fill out the omitted words in meaningful sentences. 116 target words were selected from each frequency band of 1000 words between words ranked 1–10,000 according to the Corpus of Spoken Language. For lexical retrieval speed, participants reaction time was recorded after they named the pictures as fast and accurately as possible. They found significant correlation among vocabulary knowledge, speed of lexical retrieval, and L2 learners' speaking fluency. Kahng (2014) investigated different qualitative patterns in the stimulated recall responses by the lower and higher proficiency learners. Learners' comments revealed that lower proficiency learners looked for words or selected words more frequently in language processing compared to the higher proficiency learners. Lower proficiency learners also reported that the reduced fluency was associated with their limited L2 competence.

4 The Present Study

Previous research in L2 Chinese acquisition has investigated the features of L2 Chinese utterance fluency (Chen 2012, 2013; Dai 2007; Liu 2014) and its relationship with perceived fluency (Liu and Wu 2016; Zhai 2011). Hu and Wang (2017) found strong predictive power of reaction time for sentence construction and attention shift cost to L2 Chinese speed fluency (speech rate, mean length of utterance). However, to our knowledge, there is no published studies so far relating lexical access/vocabulary knowledge to L2 Chinese speaking fluency.

As mentioned before, lexical access has been reported to significantly affect L2 speaking fluency (De Jong et al. 2013; Segalowitz and Freed 2004). While measurements of utterance fluency have been agreed upon widely by researchers, measurements of lexical access vary across studies. Segalowitz and Freed (2004) assessed learners' ability of lexical classification (living or non-living), whereas De Jong et al. (2013) tested learners' productive vocabulary knowledge of selected words of different frequency bands via a C-test. In both studies, learners' vocabulary knowledge was represented by randomly selected target words, independent from the speaking tasks learners completed. In order to get a closer look into how task-related lexical access interact with L2 learners' utterance fluency, in the present study, we

chose to investigate learners' vocabulary size and lexical retrieval speed within four tasks of different text types. We narrowed down the scope of investigation into task-related lexical access because that at the microplanning stage, after a message is constructed, the speech learners seeking to prepare is closely associated with the communicative task they are dealing with. Word selection is limited within the task. Besides, instead of following Segalowitz and Freed (2004) or De Jong et al.'s (2013) protocol, we chose a translation approach to test learners' vocabulary knowledge. Specifically, in the present study, learners' vocabulary size was assessed based on the accuracy of their performance in translating a task-related word list from L1 into L2. The motivation for the choice of this approach is that L1 exerts a considerable influence on the use of L2 vocabulary. L1 is active during L2 lexical processing in both beginning and more-advanced learners, that form-meaning link is often established via L1 translations (Schmitt 2008). By using a translation approach, we were able to explore learners' form-meaning matching behaviors based on the efficiency of the translations. Stimulated recall responses in Kahng's (2014) study revealed details of disfluency being affected by difficulty in lexical retrieval. Stimulated recall interview is an effective way to get an idea of what learners are thinking when processing L2 speech. Hence, in this study, we adopted a mixed-methods approach combining quantitative analysis in an experiment with qualitative inquiry through conducting stimulated recall interviews.

The aim of the current study was to explore how L2 learner's task-related lexical assess (measured by vocabulary size and lexical retrieval speed) relate to their utterance fluency, including speed fluency (measured by speech rate, mean length of runs), breakdown fluency (measured by the number of silent pauses, mean length of silent pauses, the number of filled pauses), and repair fluency (measured by the number of repairs, repetitions and restarts). The following research questions were addressed in this study:

1. Is L2 learners' task-related vocabulary size correlated with the utterance fluency measures?
2. Is L2 learners' task-related lexical retrieval speed correlated with the utterance fluency measures?
3. Based on stimulated recall responses, what is the role of lexical access in producing fluent speech among all reasons that cause disfluency in L2 speech?

5 Method

5.1 Participants

Two groups of American learners of Chinese participated in this study on a voluntary basis. The mean age was 22 years old ($SD = 4.1$) and their first language was English. They were enrolled in a third-year Chinese course at a US university. None of them was familiar with any other languages. There were 15 learners in the first

group—an experiment group (10 males, 5 females) and 13 learners in the second group—a stimulated recall interview group (10 males, 3 females). According to the instructor of the course, these participants' Chinese proficiency levels ranked at ACTFL intermediate-high to advanced-low levels (ACTFL 2012).

5.2 *Experiment*

5.2.1 **Materials**

A task-related test was created to investigate how lexical access interacts with L2 utterance fluency within four tasks. Before the experiment, we invited six Chinese native speakers from the same university (ages 19–22, four females and two males) to complete four speaking tasks, which were also used to test fifteen L2 learners. By doing this, we were able to set up the reference on the basis of native speakers' productive vocabulary.

Four speaking tasks represented four text types, carrying four different communicative functions: instructive, descriptive, explanatory, and argumentative. In the first task, both Chinese native speakers and L2 learners were asked to introduce the city where the university was located. In the second task, they described their first day at the university. In the third task, they were presented a data chart and were invited to explain the income gap between males and females of different age groups. In the last task, they talked about their opinions on a given topic, specifically 'what kind of professors are good college professors?' These tasks were not culturally specific; the native speakers and the participants shared similar experiences at the same university. We assumed that most of the vocabulary output by these two groups should be within a limited range when they completed the same tasks.

Based on the vocabulary Chinese native speakers used in the four tasks, we compiled a list of 198 vocabulary items that were most commonly used (being used at least three times by different speakers). All words were listed in a random order, controlling the effect derived from word frequency and task order. The list was then translated into English by the researcher for the vocabulary test.

5.2.2 **Procedure**

There were two parts in the experiment. The first part was a vocabulary test. In this part, participants were instructed to translate the words on the vocabulary list orally from English into Chinese as fast as possible. They were instructed to respond 'I don't know' if they did not know the answer. They were not given pre-task planning time. The whole process was timed in order to record lexical retrieval speed. If participants were able to say the target words or their synonyms, the answers were rated as correct; otherwise, the answers were rated as incorrect. Two L2 Chinese teachers rated the vocabulary test independently; there was no disagreement between two ratings.

Participants took a five-minute break after completing the first part and then continued to finish the second part. The second part was a speaking test consisting of four monologue tasks. Participants completed four speaking tasks, which were the same as the ones completed by the six Chinese native speakers. For each task, participants had one minute to prepare and ten minutes to speak. Participants' speech was recorded through a recording software 'Audacity' with the stereo setting set at 44,100 Hz. The experiment was conducted in the researcher's office individually, administered by the researcher. The total time commitment for each participant in this experiment was about 1.5–2 h.

5.2.3 Measures and Statistical Procedures

L2 learners' lexical access is represented by both vocabulary size and lexical retrieval speed. Vocabulary size was measured based on the accuracy rate in the vocabulary test. Lexical retrieval speed was measured by calculating the average response time for each word in the vocabulary test. As for speaking performance, all participants' speech samples were first transcribed by a Chinese native speaker. Afterward, they were encoded manually by the researcher as described in detail below. Then, six variables of the following three categories were measured for utterance fluency: speed fluency (speech rate, mean length of runs); breakdown fluency (mean length of silent pauses, number of silent pauses, number of filled pauses), and repair fluency (number of disfluencies). A script programmed in PRAAT (Boersma and Weenink 2010) was used to detect silent pauses with a minimum silence duration set to 250 ms. We were therefore able to measure speech rate, mean length of runs, mean length of silent pauses, and the number of silent pauses. Filled pauses such as *en* (嗯 'um'), *ránhòu* (然后 'and then'), *jiùshì* (就是 'that is'), *nàgè* (那个 'that'), as well as disfluencies such as repetitions, restarts, or repairs, were extracted manually from the transcripts of the speech samples. The number of filled pauses and the number of disfluencies were then calculated.

Table 1 lists the calculation methods used to measure L2 speaking performance in this study.

5.3 Stimulated Recall

5.3.1 Tasks and Materials

The stimulated recall interview was conducted in a quiet office. Learners' speech was also recorded through 'Audacity' with the stereo setting set at 44,100 Hz. The experiment was conducted in the researcher's office individually, administered by the researcher. The total time commitment for this session was about 1.5–2 h. It includes two parts. In the first part, learners were asked to complete two communicative tasks in a dialogue manner with the researcher. The learners had no preparation time before

Table 1 Measures of L2 utterance fluency and their calculation methods

	Variables	Calculation methods
Speed fluency	Speech rate	The total number of syllables divided by total time
	Mean length of runs	The average number of syllables produced in utterances between pauses of 0.25 s and above
Breakdown fluency	Mean length of silent pauses	The total length of pauses above 0.25 s divided by the total number of pauses above 0.25 s
	Number of silent pauses	The total number of pauses over 0.25 s divided by the total amount of time spent expressed in seconds and is multiplied by 60
	Number of filled pauses	The total number of filled pauses such as divided by the total amount of time spent expressed in seconds and is multiplied by 60
Repair fluency	Number of disfluencies	The total number of disfluencies such as repetitions, restarts, and repairs divided by the total amount of time expressed in seconds and multiplied by 60

completing the task. In order to understand the fluency issues that occur in a natural conversation, the learners were able to ask the researcher for help when they did not know the vocabulary. The researcher avoided to interrupt the respondent speech as much as possible during the conversation. When the learners had difficulty generating content, the researcher would extend the question to encourage longer speech.

In the first task, the learners were presented a cartoon map of USA with icons of special features of each state (see Appendix). They were asked to introduce the USA based on the maps. As the learners are all Americans, very familiar with their country, this topic was moderate in difficulty. They know this country very well, especially for certain states and cities. In the second task, the learners were invited to answer following questions, described the details and explained: (1) What is your major? Why did you choose this major? (2) Tell me what you have learned in this major course. (3) In your opinion, is your major related to your future work? Why? (4) Which majors do you think are more useful and which are less useful? Because this topic was more abstract and professional in nature, this topic was more difficult for learners.

In the second part, the researcher replayed the recordings of the two conversations in the two tasks and paused at locations where disfluency issues appeared. The learners were invited to describe what they were thinking to respond to interview questions related to the disfluencies, such as: ‘What were you thinking at that time when you paused, repeated, or repaired?’ The subjects could answer in Chinese or English. Learners’ responses were also recorded.

5.3.2 Encoding

All the recordings including learners' stimulated recall responses were transcribed into texts and then encoded. Pauses longer than 0.5 s were marked with '//', speech extension was marked with '~', and the stimulated recall responses were marked with '()'. For example,

1. Chinese: 然后你也会看到在~//佛罗里达州//佛罗里达//佛罗里达州有鳄鱼
English: Then you will also see that there is a crocodile in ~ // Florida State// Florida / Florida State.
2. Chinese: 我//我只去过//那个佛//佛州 //只去过一次
English: I // I have only been to // that Flo // Flo State // only go there once.

6 Results

6.1 Experiment

Regarding the vocabulary test, it took participants an average of 516.5 s to complete the vocabulary test. The average reaction time for each word was 2.6 s. All participants correctly translated at least half of the words in the vocabulary list, with the accuracy rate ranging from 58 to 96%. The average accuracy rate was 81%, which indicates that these learners are familiar with most of the words on the list. Regardless, none of them could translate all the words correctly. Table 2 shows each participant's accuracy and average reaction time for each word in the vocabulary test. This accuracy represents the learners' task-specific receptive vocabulary size. Reaction time shows how fast lexical retrieval was. The stronger the link between the conceptual messages and the L2 lexical items (e.g., the concept of causal relation '*because*' and the Chinese words '因为 *yinwei*'), the more words were translated correctly, and the faster the reaction time was.

Table 2 Accuracy and average reaction time in the vocabulary test

Students ($N = 15$)	Accuracy	Reaction time per word (s)
1	0.68	3.3
2	0.58	4.3
3	0.88	2.84
4	0.85	1.68
5	0.74	2.72
6	0.84	2.96
7	0.89	1.95
8	0.91	2.13

(continued)

Table 2 (continued)

Students ($N = 15$)	Accuracy	Reaction time per word (s)
9	0.73	2.34
10	0.78	2.43
11	0.67	3.36
12	0.95	2.03
13	0.75	2.24
14	0.95	2.08
15	0.96	2.78
Mean (SD)	0.81 (0.12)	2.61 (0.68)

Table 3 shows L2 learners' utterance fluency in four speaking tasks. To determine the degree of the relationship between vocabulary size and all measures of utterance fluency, Pearson's correlations were calculated. Table 4 presents the correlations among vocabulary size, lexical retrieval speed, and all the measures of utterance fluency. A significant correlation was found between task-specific vocabulary size and all fluency measures except for the number of filled pauses ($r = 0.012, p = 0.926$). In particular, L2 learners' vocabulary size was strongly correlated with speed fluency (speech rate, $r = 0.375, p = 0.003$; mean length of runs, $r = 0.354, p = 0.005$), with breakdown fluency (mean length of silent pauses, $r = -0.256, p = 0.048$; number of silent pauses, $r = -0.35, p = 0.006$), and with repair fluency (number of disfluencies, $r = -0.285, p = 0.027$).

Pearson's correlations also demonstrated a high correlation between task-specific lexical retrieval speed, which was measured by the average reaction time for each word in the vocabulary test, and speech rate ($r = -0.379, p = 0.003$). No significant correlation was found between lexical retrieval speed and other fluency measures, including mean length of runs ($r = 0.076, p = 0.565$), breakdown fluency (mean length of silent pauses, $r = 0.023, p = 0.862$; number of silent pauses, $r = 0.147, p = 0.262$; number of filled pauses, $r = -0.127, p = 0.332$), or with repair fluency (number of disfluencies, $r = 0.189, p = 0.148$). See Table 4.

6.2 Stimulated Recall

According to the stimulated recall responses, we found that 70 disfluent AS-units¹ were related to planning what to say. 280 disfluent units were related to non-content reasons, which accounts for 25% of all the fluent phenomenon. Disfluent phenomena related to content planning were not included in the investigation. Table 5 presents a summary of causes of disfluency in the stimulated recall interviews.

¹An AS-unit is 'a single speaker's utterance consisting of an independent clause or subclausal unit, together with any subordinate clause(s) associated with it' (Foster et al. 2000: 365).

Table 3 Participants' speaking performance in four speaking tasks (N = 15)

	Task 1		Task 2		Task 3		Task 4		Average of all tasks	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1. Speech rate	2.49	0.47	2.69	0.64	2.14	0.49	2.62	0.49	2.48	0.52
2. Mean length of runs	0.79	0.62	0.82	0.72	0.67	0.20	0.64	0.19	0.73	0.43
3. Mean length of silent pauses	106.44	54.35	98.44	37.45	98.90	39.16	102.70	40.11	101.62	42.77
4. Number of silent pauses	0.54	0.21	0.47	0.16	0.61	0.18	0.56	0.17	0.55	0.18
5. Number of filled pauses	9.97	5.52	8.74	4.45	10.29	4.20	8.85	5.47	9.46	4.91
6. Number of disfluencies	5.75	3.07	5.25	2.97	4.11	2.58	5.51	2.18	5.15	2.70

Table 4 Correlation between lexical access and utterance fluency

	Accuracy of vocabulary test	Speed of lexical access	Speech rate	Mean length of runs	Mean length of silent pauses	Number of silent pauses	Number of filled pauses	Number of disfluencies
Accuracy of vocabulary test								
Speed of lexical access	-0.703 ^a							
Speech rate	0.375 ^a	-0.379 ^a						
Mean length of runs	0.354 ^a	-0.076	0.255 ^b					
Mean length of silent pauses	-0.256 ^b	0.023	-0.236	-0.732 ^a				
Number of silent pauses	-0.350 ^a	0.147	-0.568 ^a	-0.526 ^a	0.461 ^a			
Number of filled pauses	0.012	-0.127	-0.068	-0.404 ^a	0.589 ^a	0.12		
Number of disfluencies	-0.285 ^b	0.189	0.098	-0.097	0.131	0.008	-0.212	

^aCorrelation is significant at the 0.01 level

^bCorrelation is significant at the 0.05 level

The disfluent phenomenon in spoken language unveiled learners' cognitive activities in dealing with problems in the L2 speech production. Twenty-one types of responses were classified into three main categories: related to L2 incompetence, not related to L2 incompetence, and corrections, as summarized in Table 5. Most responses mentioned that learners had difficulty in planning or assembling speech before language output, which accounts for 86.43% of the disfluencies. 13.57% of the responses pointed to corrections after language output. 81.07% of the disfluencies were due to learners' incompetence in L2 at pronunciation, vocabulary, and grammar levels. 5.36% were caused by reasons not related to learners' L2 incompetence, such as feeling time pressure, difficulties in planning the speech in L1, or just wanting to slow down so that the listener could hear more clearly.

Table 5 Summary of causes of disfluency

					Percentage	
Before language output	Related to L2 incompetence	Pronunciation	1. Attempted to select the correct L2 pronunciation among similar ones	0.36	86.43	
			2. Simulated the pronunciation of L1 word with two L2 phonemes to assemble the transliterated word	2.86		
			3. Replaced unfamiliar L2 pronunciation with approximate L2 phonemes	0.36		
		Vocabulary	4. Attempted to retrieve a L2 word to translate L1	21.79		
			5. Failed to retrieve the L2 target word to translate L1 first. Then used another way to express, or explained with an example, or avoided using the target word, or used L1 word directly, or used a superordinate word to replace the target word	23.92		

(continued)

Table 5 (continued)

			Percentage
		6. Failed to retrieve the target L2 word at first. Replaced it with another word with similar meaning, and then recalled the target word and used it	0.71
		7. Attempted to select the correct L2 word among synonyms	8.93
		8. Attempted to select the correct L2 word among words with similar radicals	0.36
		9. Attempted to find a better L2 word but failed. Then gave up and used the word that they could recall at the time	3.21
		10. Attempted to select the correct L2 morpheme among different morphemes	0.71
		11. Not sure if the L2 morphemes were used correctly	0.71

(continued)

Table 5 (continued)

				Percentage	
			12. Not sure if the combination of the L2 word was correct	4.29	
		Grammar	13. Not sure if the L2 grammar was correct	7.14	
			14. Unable to express the complex meaning with corresponding L2 sentence. So simplified the sentence structure	2.14	
	Not related to L2 incompetence	15. Time pressure		1.07	
		16. Difficulties in planning the sentence in L1		3.93	
		17. Slowed down to make sure the listener heard clearly		0.36	
After language output	Correction	18. Attempted to correct the wrong pronunciation		4.29	13.57
		19. Attempted to correct the wrong word		5.71	
		20. Attempted to correct the wrong grammar		3.57	

It should be noted that 64.63% of the disfluencies appeared when the learners struggled to process vocabulary effectively, which is the main reason of disfluencies among all the responses. 3.58% disfluencies were related to processing pronunciation, whereas 9.28% were related to processing grammar.

Among all the causes at the vocabulary level, 21.79% reported that the learners encountered difficulty in translating L1 into L2. Example (1) is a comment regarding to silent pauses when the learner attempted to retrieve L2 words to translate L1.

(1) Chinese: 我不喜欢血//blood//血

English translation: but // I don't like blood (L2) // blood (L1) // blood (L2)

Stimulated recall response: It's just I've heard it said and like I've seen it. I know it's 'xue (blood),' but I've never said it, so I wasn't sure.

24.63% of the responses commented on failing to translate L1 words into L2. Some of them compensated the issue by explaining with examples, avoiding the use of the target words, using L1 words directly, or using superordinate words to replace the target words. In some other cases, the learners recalled the L2 words later after hesitations. Example (2) is a comment regarding silent pauses and a filled pause because that the learner failed to retrieve the L2 word to translate the L1 target word. He used another way to express it. In example (3), the learner gave up after failing to translate L1 and used code-switching strategy to finish the sentence.

- (2) Chinese: 那个时候也有那个//呃//找金属的
 English translation: at that time also had that // uh // gold searching
 Stimulated recall response: I don't know how to say 'Gold Rush' in Chinese.
- (3) Chinese: 亚洲学学到的就是//呃//我觉得最重要就是//呃//中//呃//中国//中国在亚洲//呃//该怎么说呢//呃//做的//就是他们的//他们的role
 English translation: Asian studies learns that is // uh // I feel the most important is just // uh // Chi- // uh // China // China in Asia // uh // How should I say this? // uh // do // I mean their // their role
 Stimulated recall response: I know I want to say 'role,' but I don't know how to say it in Chinese.

Reasons related to selecting words account for 12.5% of the disfluencies. Example (4) shows that the learner had a filled pause and sound extension in her speech when she tried to select the correct L2 word among synonyms.

- (4) Chinese: 呃//所以我应该会选择在这边~生活
 English translation: Uh, therefore I probably will choose this place to ~ live
 Stimulated recall response: I was thinking of 'to survive,' but it's not quite fitting, 'live' was better.

5.71% of the comments reported disfluencies coming from issues in word assembling. In example (5), the learner used fillers twice and repeated part of the target word because he was struggling in assembling the word.

- (5) Chinese: 呃//还有西//呃//西南部有很多的//cactus//就代表这//这一个部分比较少会下雨
 English translation: uh // also the west // uh // south western part has many // cactus // just represents this // this part rains comparatively less
 Stimulated recall response: I was thinking if it was 'south west' or 'west south.'

There are 5.71% of the comments mentioned that after speech was produced, learners' perceived errors in terms of vocabulary use. They therefore corrected their speech by repeating the words with the correct forms or restating part of the speech with the correct words. In example (6), when the learner heard himself used an incomplete word, he paused and restated the sentence with the correct form.

- (6) Chinese: 因为那边的水非常温~暖//呢//跟加州的水比起来非常温暖
English translation: because that side's water is super wa~rm // uh // compared to California's water is super warm
Simulated recall response: I originally just wanted to say 'warm (wen),' then I added the character 'nuan,' because I felt that 'wen' wasn't complete.

7 Discussion

The present study examined the relationship between L2 learners' task-related lexical access and their utterance fluency. Prior work has shown that vocabulary knowledge affects L2 learners' speaking performance, but little is known about how task-related lexical access interact with utterance fluency. Moreover, most research so far only includes quantitative analysis. There has been so far no study examining L2 Chinese cognitive fluency from the perspective of learners' vocabulary. To overcome these limitations, this study examined how Chinese L2 learners' vocabulary size and lexical retrieval speed relate to their utterance fluency, with both quantitative analysis and quantitative interviews. Results from the experiment and the stimulated recall responses both confirmed the key role lexical access plays in L2 speech production and its close relation with utterance fluency. We discuss our findings with respect to (a) vocabulary size and L2 utterance fluency, (b) lexical retrieval speed and L2 utterance fluency, and (c) disfluency causes and L2 utterance fluency.

Through a vocabulary test requiring learners to translate 198 words from L1 to L2 related to the four tasks they completed later in the experiment, we were able to examine learners' task-related vocabulary size and lexical retrieval speed. In response to the first research question, 'Is L2 learners' task-related vocabulary size correlated with the utterance fluency measures?', this study found significant correlations among vocabulary size and all three facets of utterance fluency: speed fluency (speech rate, mean length of runs), breakdown fluency (mean length of silent pauses, number of silent pauses), and repair fluency (number of disfluencies), except for the number of filled pauses. The better learners did in the vocabulary test, the more L2 vocabulary they could retrieve from their mental lexicon, the easier they generated speech. Though the current study adopted a different method to examine L2 learners' vocabulary size, this finding is in line with previous study (De Jong et al. 2013) and confirmed that with a larger L2 vocabulary inventory, word-meaning links were easier to establish. Therefore, less effort was put into lexico-grammatical encoding, morpho-phonological encoding, phonetic encoding as well as self-monitoring, so that learners' speech output was more fluent with faster speech rate, less filled pauses, hesitations, repairs, or repetitions.

Learners' speaking performance reflects the automaticity in learners' cognitive processing system, which is a bundle of three features (Segalowitz 2010): fast (processing speed), parallel (cognitive activities are executed simultaneously), and effortless (few attention resources required). In addition to learners' vocabulary size, it is also important to investigate the efficiency of lexical access. Concerning

the second research question, 'Is L2 learners' task-related lexical retrieval speed correlated with the utterance fluency measures,' only speech rate among all fluency measures was significantly correlated to lexical retrieval speed. The more efficient lexical access was, the faster learners retrieved L2 words from mental lexicon, the faster they talked. A close link regarding speed attribute was found between cognitive fluency and utterance fluency. This finding is partially in line with previous studies (De Jong et al. 2013; Segalowitz and Freed 2004). In De Jong et al.'s study (2013), lexical retrieval speed was significantly correlated with silent pauses, filled pauses, and repetitions. Lexical access efficiency was found to be correlated with the mean length of filler-free run in Segalowitz and Freed's (2004) study. Our results show that lexical retrieval speed was correlated with speech rate only. Mixed results were found in different research. It may be explained by that fact that different methods were used to analyze L2 learners' lexical access speed in these three studies. Another possible reason is that the speaking tasks that were used to elicit L2 speech varied among three studies. Nevertheless, different kinds of evidence from different research revealed that a fluent speech required high efficiency of lexical access.

Stimulated recall interviews were conducted to obtain more details of how lexical access affect utterance fluency in L2 speech. Answering the third research question, 'what is the role of lexical access in producing fluent speech among all reasons that cause disfluency in L2 speech?', learners' responses confirmed that lexical access plays a key role in producing fluent speech, as around two-thirds of the disfluencies were reported to be caused by vocabulary-related issues. Moreover, learners' comments also revealed that they struggled with accessing L2 words at different stages along the process of speech production. At the microplanning stage, after message was generated ($\{f_1\}$), learners tried to retrieve L2 words from their long-term memory ($\{f_3\}$), their fluency was reduced when they could not efficiently translate L1 words into L2 words, select the correct words among synonyms, or find better words. When they failed in linking L1 words and L2 words, they had to respond quickly by using compensational strategies to complete the expression. This also caused their speech to be less fluent. At the lexico-grammatical stage ($\{f_2\}$), learners' fluency was affected when they encountered problems in regard to choosing the correct morphemes to assemble words or when they lacked confidence in the words they assembled. After speech was articulated, they found lexical errors in their self-perception ($\{f_7\}$). They corrected errors by replacing them with the correct forms, which led to less fluent speech. L2 speech production is lexical driven in that word-meaning links are central to using a language (Kormos 2006; Segalowitz 2010). Two routes of L2 lexical access were found in the stimulated recall responses in the current study: a direct route that learners used to retrieve L2 words to match

the abstract message generated in the macroplanning stage; and an indirect route that learners translated L1 words into L2 words, as word-meaning links are stronger in L1 than L2. This finding supports the 'hierarchical model' (Potter et al. 1984) that proposed two ways in which concepts are related to words in L2. Combining results from quantitative analysis and qualitative inquiry, the crucial role of lexical access in L2 speech production was supported in this study.

8 Conclusion

The present study showed that efficient task-related lexical access was crucial for producing fluent speech in second language. Vocabulary size was found to affect all facets of speaking fluency (speed fluency, breakdown fluency, and repair fluency), whereas lexical retrieval speed only related significantly to speech rate. The findings suggested that to improve L2 learners' speaking fluency, it would be helpful to design tasks that can promote L2 lexical access, such as tasks to expand L2 vocabulary size, to reinforce L1-L2 links, to strengthen memory of L2 word forms, and to introduce formulaic sequences to reduce attention effort. Rather than testing L2 learners' general vocabulary size, this study focused on learners' vocabulary size within specific tasks. We believe that by using this more focused approach, it was possible to investigate more closely the dynamics between L2 learners' lexical access and their speaking fluency. However, the number of vocabularies in the experiment was limited. It would be better for future studies to include more tasks with different topics and text types to increase the vocabulary number. It is also necessary to consider individual differences of productive vocabulary in completing the same tasks, especially the 'vocabulary gap' between native speakers and non-native speakers in relation to native-referenced vocabulary test design. In addition, to improve the reliability of the experiment for future work, a larger number of participants at different proficiency levels may be included. With a larger scale of investigation and replicated studies, it would be able to further discuss how lexical access affects L2 speech. Another limitation is that the current study compared data from two different groups of participants completing different speaking tasks in the experiment and the stimulated recall session, respectively. The validity of the findings may be affected by the group differences as well as task differences. Hence, for future research, it would be better to conduct stimulated recalls with the same group of participants after they completed the tasks. In addition, it would also be useful to explore more about the link between L2 learners' task-specific receptive vocabulary knowledge and productive vocabulary knowledge by comparing the words in the vocabulary test as well as in learners' speech.

Appendix

Material used in the first task of stimulated recall interviews.



References

- ACTFL, J. (2012). *ACTFL proficiency guidelines*. Alexandria, VA: American Council on the Teaching of Foreign Languages. <https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/english/speaking>.
- Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). Newark, DE: International Reading Association.
- Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, 10(3), 245–261.
- Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, 30(2), 159–175.
- Bosker, H. R., Quené, H., Sanders, T., & De Jong, N. H. (2014). The perception of fluency in native and nonnative speech. *Language Learning*, 64(3), 579–614.
- Boersma, P., & Weenink, D. (2007). Praat (Version 6.1.38) [Software]. Latest version available from www.praat.org (accessed 2 January 2021): <https://www.fon.hum.uva.nl/praat/>
- Chen, M. (2012). Meiguo liuxuesheng hanyu kouyu chanchu de liulixing yanjiu 美国留学生汉语口语产出的流利性研究 [Chinese oral fluency of CSL learners of American English speakers]. *语言教学与研究 [Language Teaching and Linguistic Studies]*, 2, 17–24.
- Chen, M. (2013). Meiguo liuxuesheng hanyu kouyu chanchu de yunlu bianjie tezheng yanjiu 美国留学生汉语口语产出的韵律边界特征研究 [The characteristics of oral Chinese prosodic boundaries of American English speaking CSL learners]. *世界汉语教学 [Chinese Teaching in the World]*, 1, 95–104.

- Dai, Y. (2007). Di er yuyan xuexizhe hanyu huihua xiuzheng xianxiang yanjiu 第二语言学习者汉语会话修正现象研究 [A study on conversational repair of the Chinese second language learners]. *汉语学习 [Hanyu Xuexi]*, 6, 69–75.
- Daller, H., Milton, J., & Treffers-Daller, J. (Eds.). (2007). *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press.
- De Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, 61(2), 533–568.
- De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, 34(5), 893–916.
- Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, 54(4), 655–679.
- Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, 533–557.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). A unit for all reasons: The analysis of spoken interaction. *Applied Linguistics*, 21(3), 354–375.
- Gatbonton, E., & Segalowitz, N. (2005). Rethinking communicative language teaching: A focus on access to fluency. *Canadian Modern Language Review*, 61(3), 325–353.
- Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal*, 36(2), 153–166.
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–473.
- Hu, W., & Wang, J. (2017). Di er yuyan kouyu renzhi liulixing dui kouyu nengli de yuce zuoyong 第二语言口语认知流利性对口语能力的预测作用 [Predictive power of L2 oral cognitive fluency to L2 oral competence]. *世界汉语教学 [Chinese Teaching in the World]*, 31(1), 105–115.
- Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, 64(4), 809–854.
- Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching & Research*, 4(5).
- Kormos, J. (2006). *Speech production and second language acquisition*. Routledge.
- Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145–164.
- Levelt, W. J. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levelt, W. J. (1999). Producing spoken language. In *The neurocognition of language* (pp. 83–122).
- Liu, F. (2014). Liuxuesheng hanyu kouyu ziwo xiuzheng yanjiu 留学生汉语口语自我修正研究 [A study on conversational repair of the Chinese second language learners]. *华文教学与研究 [TCSOL Studies]*, 1, 42–48.
- Liu, Y., & Wu, X. (2016). Hanyu eryu xuexizhe chanchu de liulixing yanjiu 汉语二语学习者产出的流利性研究 [Exploring speaking fluency in the speech of L2 Chinese learners]. *华文教学与研究 [TCSOL Studies]*, 64(4), 32–60.
- Meara, P. (1996). The vocabulary knowledge framework. In *Vocabulary acquisition research group virtual library* (pp. 1–11).
- Meara, P., & Wolter, B. (2004). V_Links: Beyond vocabulary depth. *Angles on the English Speaking World*, 4, 85–96.
- Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 211–232).
- Milton, J., & Fitzpatrick, T. (Eds.). (2013). *Dimensions of vocabulary knowledge*. London: Macmillan International Higher Education.
- Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

- Potter, M. C., So, K. F., Von Eckardt, B., & Feldman, L. B. (1984). Lexical and conceptual representation in beginning and proficient bilinguals. *Journal of Verbal Learning and Verbal Behavior*, 23(1), 23–38.
- Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined. In *Vocabulary in a second language* (pp. 209–227).
- Rehbein, J. (1987). On fluency in second language speech. In *Psycholinguistic models of production* (pp. 97–105).
- Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, 37(6), 828–848.
- Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, 14(4), 423–441.
- Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review*, 65(3), 395–412.
- Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.
- Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching*, 54(2), 79–95.
- Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, 26(2), 173–199.
- Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, 12(3), 329–363.
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84–119.
- Uchihara, T., & Clenton, J. (2020). Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research*, 24(4), 540–556.
- Uchihara, T., & Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The Language Learning Journal*, 47(1), 64–75.
- Wood, D. (2002). Formulaic language acquisition and production: Implications for teaching. *TESL Canada Journal*, 1–15.
- Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *Canadian Journal of Applied Linguistics*, 12(1), 39–57.
- Zhai, Y. (2011). Kouyu liulixing biao zhun de keguan hua yanjiu 口语流利性标准的客观化研究 [An objective study on the subjective criteria of speaking fluency]. *语言教学与研究 [Language Teaching and Linguistic Studies]*, 5, 79–86.