Analysis of Phonemes and Tones Confusion Rules Obtained by ASR

Gulnur Arkin and Askar Hamdulla(✉)

Institute of Information Science and Engineering, Xinjiang University, Urumqi 830046, China
askar@xju.edu.cn

Abstract. This paper is based on the exploration of the effective method of erroneous phoneme pronunciation of Chinese mandarin learners whose mother tongue is Uyghur and the solution of major problems of language education, concerning the learner's pronunciation, it uses a different method, namely data-driven approach, and the Automatic Speech Recognition (ASR) is also used to recognize phonemes of the pronunciation of Chinese mandarin learners whose native language is Uyghur. The phoneme sequence is identified and then the standard pronunciation phonemes corresponding to the recognized phonemes are used as the target phonemes to obtain the mapping relation of each target phoneme and recognition phoneme, thus the possible phoneme error categories and possible erroneous rules in Uyghur learners' pronunciation can be obtained, which may give some help to the Uyghur learners to learn the Chinese auxiliary language system and the corresponding pronunciation evaluation model.

Keywords: Non-native speakers of Chinese · Speech recognition · Confusion rules · Pronunciation evaluation

1 Introduction

With the continuous development of the global economy, exchanges and cooperation in political, economic, cultural and educational fields among various countries have become more and more frequent (Jiang and Wang et al. 2018a; Jiang et al. 2016). Travelling and learning abroad is also increasingly common. Therefore, in addition to the mother tongue, many people choose another language as the second language. In ethnic minority areas of China, the Mandarin Chinese, as a national language, has been very crucial from primary school, junior high school, high school to college. Efficient spoken language learning requires one-on-one, face-to-face interaction between teachers and students. However, this approach is constrained by space, time and economic conditions (Ito et al. 2007; Stanley and Hacioglu 2012; Wang and Lee 2012). In recent years, with the development of science and technology, online education has become more and more popular. The Cloud-centric powerful computing resources, highly popularized mobile smart devices and rapidly developed voice processing technologies have enabled computer-assisted language learning System (CALL) to become more and more popular (Witt 1999; Ye and Young 2005; Qian et al. 2011).

H. Song et al. (Eds.): SIMUtools 2019, LNICST 295, pp. 469–476, 2019.
https://doi.org/10.1007/978-3-030-32216-8_45

However, the detection and diagnosis of pronunciation errors at the phonemic level, as a core module of the CALL system, still need to be further improved in accuracy.

Each language has its own vowel phonemic system. There are thirty-two phonemes in the phonological system of the modern Uyghur standard language. Eight of which are vowel phonemes but no diphthong. In contrast, Chinese has not only single vowels but also diphthongs and four triphthongs. In fact, it does not have any practical significance in terms of the phoneme itself, however, its function is very huge, the main difference lies in distinguishing the meaning and specifically distinguishing the different languages speed and different words (Ladefoged and Johnson 2015) Mandarin Chinese and Uyghur belong to the Sino-Tibetan language and Altaic languages respectively, and there are great differences in phonetics between these two languages, Chinese belongs to the isolated language (Thurgood and LaPolla 2003), and the Altaic language belongs to the agglutinative language (Zhao and Zhu 1985). There exists hierarchical relationship among phonemes, syllables, words, sentences, specifically, how the phonemes to form syllables, how syllables to form a specific single word, and how the word to form a sentence to express a certain meaning, these are the horizontal combination of phonetics, belonging to the scope of horizontal combination (Shifeng 2009). When learners learn another language, the old phonetic perception and production systems play an auxiliary or interference role (Lo et al. 2010). Certain types of phonetic errors are the product of interference effect, they are predictable, interpretable and understandable, and many research institutes have begun to pay attention to this issue, and they started to explore the system. For non-native speakers of languages, they attach particular importance to computer-aided Chinese language learning systems, which are especially suitable for ethnic minority areas. Now, relevant literature on the comparison of two languages almost can't be found, while this literature is to record the confusion rules, in order to improve the quality of pronunciation, and strive for accuracy.

The methodologies involved in this study are mainly aimed at Uyghur learners (L1) whose native language belongs to the Altai language family, and they have some Mandarin learning experience. Uyghur pronunciation is different from Mandarin pronunciation, some of its pronunciation cannot be found in Mandarin Chinese pronunciation, based on this, learners habitually use Uyghur pronunciation as a benchmark to learn Chinese Mandarin, focusing on pronunciation perception and aspects of producing, pronouncing ways and parts, finding the native phonemes that are similar but slightly different with Mandarin Chinese to replace the pronunciation, thus the difference caused by the substitution may cause confusion of pronunciation or certain pronunciation errors (Troung 2004). To sum up, through the collection of learners' pronunciations, we can get a comparative analysis of Uyghur cross-linguistic phonological contrast in linguistic and phonemic when they say Mandarin Chinese and wrong pronunciation characteristics with the data-driven method, so as to draw a reasonable phoneme confusion rule. At the same time, we will devote ourselves to exploring how to summarize the phonemic confusion between L1 and L2, establishing an experimental database based on the phoneme analysis, and how to combine the speech recognition technology and phonemic confusion to evaluate the accuracy of phoneme pronunciations, so as to establish an automatic detection method for phonetic erroneous pronunciation specifically designed for Uyghur who studies Chinese. Therefore, this study has actual theoretical research value.

2 Experimental Data and Preparation

2.1 Experimental Subjects

50 Uyghur speakers' sound recordings have been collected, all of them are students of Xinjiang University, aged from 20–26 years (Means 23), and their native language is Uyghur, Mandarin Chinese is their second language, they do not have language listening problems, and their parents are Uyghur, who use Uyghur as a communicative language in daily communication. 50 speakers were born in Xinjiang, fluent in Uyghur (the Native-tongue-using Minority Nationality Students), and their learning time on Chinese mandarin is more than ten years, their Chinese MHK oral test scores are above 45.

2.2 Experimental Process

Each speaker was sitting in a sound booth during the experiment, and the microphone was five centimeters from the speaker. The voice used in the experiment was collected in a dedicated recording studio, using equipment like a laptop, external sound card, microphone and some interconnecting data lines. The use of external sound card can adjust the volume of sound, reduce the noise, and monitor the situation of the plosive sound, etc. Recordings were under computer control by a program in Matlab, each data sampling point is digitized into bits, and the sampling rate is 16 Hz. Participants' read materials are Chinese sentences, each participant needs to record 300 Mandarin sentences (50 * 300 = 15000) and each sentence contains 5 to 11 words. The recognition results are obtained through Chinese speech recognition system of Tsinghua University.

3 Data-Driven Approaches

3.1 Processes and Methods

It is well-known that people habitually form a system of relative perception production when acquiring language (e.g., mother tongue). When people learn another language (for example, ethnic minority areas mainly learn Chinese, bilingual learning), the original system might produce auxiliary or interference effect, it promotes learners to learn another language when playing an auxiliary role, and the pronunciation errors are often specific when interfering. Fortunately, they are predictable, interpretable and understandable, the burden of recognition system will be reduced due to the integration of these prior knowledge (Dong and Zhao 2006; Wang et al. 2011), giving full play to the role of recognition.

The comparative analysis of cross-linguistic phonology, as a linguistic transfer theory, mainly focuses on the comparison of L1 and L2 (Gass and Selinker 1992). The misunderstanding of non-knowledge among learners (Uyghur) can cause confusion. We focus on learners' phonetic aspect, using a different method, namely data-driven, to carry on a relevant test to it, namely the automatic phoneme recognition, to analyze the recognition result emphatically, specially to mainly dissect some wrong pronunciation

that produced among them, emphatically to discuss around wrong pronunciation and standard pronunciation, studying the related mapping relation between them, with this particular mapping relation (Jiang et al. 2018b) it automatically generate the relevant rules, and this rule is mainly for additional phoneme confusion. The data-driven approach is mainly used here to get the rules of phoneme confusion, and the flow chart is as follows (Fig. 1).

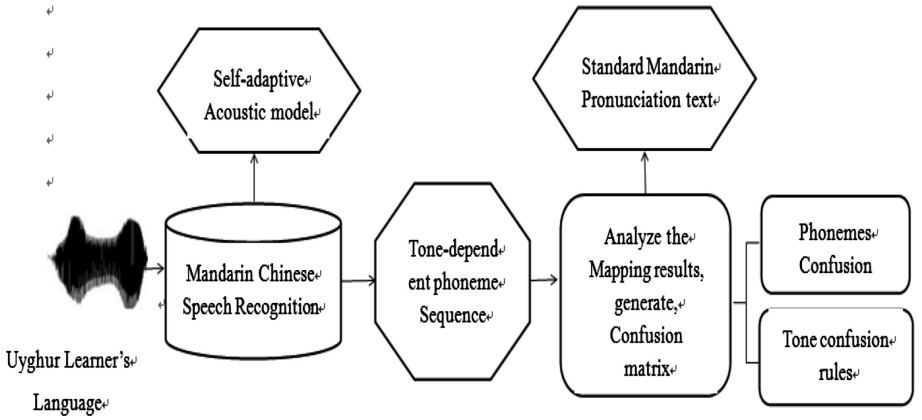

Fig. 1. The flow chart of the confusion rules of phonemes and tones generated by the results of ASR recognition is analyzed

The main method is, firstly to identify the phonemes of Chinese mandarin learners whose mother tongue is Uyghur based on the phoneme, and each Uyghur speaker is transformed automatically in the recognition to obtain the recognition sequence. Secondly, the standard phonetic phoneme is used as the target phoneme to obtain the mapping relation of the target phoneme x (i) and the recognition phoneme y (j). The relationship can be replaced, deleted, inserted and misread. Meanwhile, to count the mapping and calculate frequency) P x(i)| y(j):

$$P\,x(i)|y(j) = (Times\ of\ recognition\ phoneme\ y(j)/Times\ of\ target\ phoneme\ x\,(i)) * 100\% \quad (1)$$

Finally, the recognition accuracy is obtained by confusing the rules and calculating the frequency.

3.2 Phoneme Confusion Matrices and Rules

Vowel Confusion Matrices and Rules

We found out phoneme list with possible pronunciation mistakes by the way of automatic speech recognition to generate confusion matrix, and filtered each vowel, diphthong and consonant table map by threshold value to generate their own confusion rules (Table 1).

Table 1. The 1st–12th columns of vowel confusion matrix obtained from data driven experiment

	a	o	e	i	u	v	er	ie	ai	ei	ao	ou
a	341	0	4	4	0	0	0	0	6	0	6	0
o	0	80	0	0	4	0	0	0	0	0	0	0
e	1	2	695	9	5	0	0	2	0	0	3	8
i	5	0	26	1148	12	0	0	3	21	5	0	0
u	0	3	7	11	508	0	0	0	0	0	5	1
v	0	0	0	0	1	150	0	0	0	0	0	0
er	1	0	0	0	0	0	185	1	0	0	0	0
ie	0	0	0	4	0	0	0	167	0	0	0	0
ai	4	0	1	15	2	0	0	0	478	2	2	4
ei	0	0	1	5	0	0	0	0	0	133	0	0
ao	1	2	14	0	4	0	0	1	0	0	256	7
ou	0	3	4	1	4	0	0	0	3	0	1	232
ia	1	5	0	0	1	0	0	1	0	0	0	2
ua	0	0	0	0	0	0	0	0	0	0	0	0
uo	0	0	17	0	6	0	1	0	2	0	0	2
ve	0	0	0	0	0	1	0	0	0	0	0	0
iao	0	0	15	1	1	2	0	0	0	0	1	1
iou	0	0	0	0	0	0	0	0	0	0	0	0
uai	0	0	0	0	0	0	0	0	0	0	0	3
uei	0	0	0	0	0	0	0	0	0	0	0	0
an	2	0	0	5	0	0	0	0	0	1	4	0
en	4	0	2	16	0	0	0	0	0	0	0	0
in	0	0	8	7	0	0	0	1	0	0	0	0
vn	0	0	0	0	0	0	0	0	0	0	0	0
ian	0	0	0	5	0	4	0	2	1	0	0	0
uan	0	0	0	1	0	0	0	0	2	0	0	0
van	0	0	0	0	0	0	0	0	0	0	0	0
uen	0	0	0	0	0	0	0	0	0	0	0	0
ang	0	0	3	0	0	0	0	0	0	0	3	0
eng	0	0	0	0	4	4	0	0	0	0	1	0
ing	0	0	0	2	0	0	0	0	0	0	0	0
ong	4	0	1	4	1	0	0	0	8	4	0	1
iang	0	0	0	0	2	0	0	0	0	0	0	0
uang	0	0	0	0	0	0	0	0	0	0	0	0
ueng	0	0	0	0	0	0	0	0	0	0	0	0
iong	0	0	0	0	0	0	0	0	0	0	0	0

Table 2. The confusion rules list of monophthong, diphthong that obtained from data driven experiment

Target phoneme	Replace	Delete	Insert	Misreading	ASR correct recognition probability
a	e		i	e, i, ia	52.14%
o	u		u	u, ou	58.39%
e	a, i			a, i, ie, o	51.63%
i	e, ie, ai		e, a	e, a, ai, ei	49.54%
u	o, ou		o	ou, o, ao, e	45.51%
v	u				45.45%
er				a, ie	46.95%
ie(ě)	i			i	60.72%
ai	a, i	i		a, ie, ei	52.58%
ei	e			e, i	44.48%
ao	a	a	u	o, ou, u, ie	45.96%
ou	u	o		u, ao, ai, a	52.37%
ia	ie	i, a		ie, ou	49.72%
ie					0.00%
ua					49.18%
uo	u, ou	u, o		u, o, ou,	46.45%
ve				v	60.00%
iao	ao	i		ao, e, ou	55.33%
iou					0.00%
uai				ou	54.62%
uei					0.00%
an	en, in			en, in, ang	50.69%
en	in			an, in, eng	45.64%
in	en			an, en, vn	42.29%
vn					42.85%
ian	iao		ng	uo, iao, iang, ing	43.16%
uan	van	n	i, ng	ing, uai, ua	49.43%
van	uan			ian, uan,	42.59%
uen					0.00%
ang	an		n	an, en, eng, ong	50.72%
eng	ang		o, a	ang, ong, iang	41.94%
ing	ang	i	a, u	ian, uan, ang, iang	42.34%
ong				uan, uang	51.60%
iang		i		ing, ang	47.88%
uang		u		ang	71.66%
ueng	uang			uang, uo	33.33%
iong				eng	40.00%

From Table above, the confusion matrices and confusion rules of vowels and diphthongs can be seen. If the target phoneme is a consonant, but it is recognized as a vowel in the specific recognition, we neglect this recognition error, and the reverse is equally true. Vowel phoneme mapping, statistical frequency and recognition accuracy, all of which have presented in the tables, the first column is the target phoneme, the recognition phoneme that the behavior matches the target phoneme, without listing the mapping of phoneme that not to be considered (Table 2).

4 Conclusions

In this paper, we mainly analyze the results of 50 Uyghur learners through Chinese speech recognition. The samples are collected and the sample information is obtained through recognition on them, and based on this, the error analysis is carried out, and then the relevant rules and laws are obtained. It has introduced two cases of learners' possible pronunciation erroneous phonemes in detail; first, according to the reasons leading to erroneous pronunciation of learners, the cross-linguistic phonological comparison method is used to predict phonemic confusion, for Uyghur learners, their personal factors will also lead to confusion. With the help of data-driven, the extra phoneme confusion is summarized and the related situation of the two methods is clarified precisely, hoping that the phoneme confusion rules can provide linguistic priori knowledge for speech recognition. In the meantime, the next step is to establish a pronunciation evaluation system specifically targeting Uyghur Chinese learners by using the pronunciation rules obtained from this study in combination with speech recognition.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (NSFC; grant 61462085, 61662078, and 61633013).

References

Ito, A., Lim, Y.-L., Suzuki, M.: Pronunciation error detection for computer-assisted language learning system based on error rule clustering using a decision tree. Acoust. Sci. Technol. **28** (2), 131–133 (2007)

Stanley, T., Hacioglu, K.: Improving L1-specific phonological error diagnosis in computer assisted pronunciation training. In: INTERSPEECH 2012 [S.l.]: ISCA, pp. 827– 830 (2012)

Jiang, D., Wang, W., Shi, L., Song, H.: A compressive sensing-based approach to end-to-end network traffic reconstruction. IEEE Trans. Netw. Sci. Eng. (2018a). https://doi.org/10.1109/tnse.20182877-597

Wang, Y.B., Lee, L. S.: Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training. In: ICASSP 2012 [S.l.], pp. 5049 –5052. IEEE (2012)

Jiang, D., Zhang, P., Lv, Z.: Energy-efficient multi-constraint routing algorithm with load balancing for smart city applications. IEEE Internet Things J. **3**(6), 1437–1447 (2016)

Witt, S.M.: Use of Speech Recognition in Computer-Assisted Language Learning. [S.l.], Cambridge University (1999)

Ye, H., Young, S.J.: Improving the speech recognition performance of beginners in spoken conversational interaction for language learning. In: INTERSPEECH 2005. [S.l.]: ISCA, pp. 289 –292 (2005)

Jiang, D., Huo, L., Song, H.: Rethinking behaviors and activities of base stations in mobile cellular networks based on big data analysis. IEEE Trans. Netw. Sci. Eng. **1**(1), 1–12 (2018b)

Qian, X.J., Meng, H., Soong, F.K.: On mispronunciation Lexicon generation using joint sequence multigrams in computer-aided pronunciation training (CAPT). In: INTERSPEECH 2011. [S.l.]: ISCA, PP. 865 –868 (2011)

Ladefoged, P., Johnson, K.: A Course in phonetics, 7th edn. Peking University Press, Peking (2015)

Thurgood, G., LaPolla, R.J.: The Sino-Tibetan Languages. London Routledge, China (2003)

Zhao, X., Zhu, Z.: Uyghur Language. National press, China (1985)

Shifeng, F.: Experimental Phonology Exploration. Peking University Press, China (2009)

Lo, W.K., Zhang, S., Meng, H.M.: Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system. In: INTERSPEECH 2010, pp. 765– 768 (2010)

Troung, K.: Automatic Pronunciation Error Detection in Dutch as a Second Language: an Acoustic-Phonetic Approach, Utrecht University (2004)

Dong, B., Zhao, Q.W.: Automatic scoring of flat tongue and raised tongue in computer-assisted Mandarin learning. In: ISCSLP2006. [S.l.]: IEEE, pp. 2–7 (2006)

Wang, S., Li, H.: Research on the evaluation of spoken language scale intelligence for second language learning. Chin. J. Inf. Sci. **25**(6), 142–148 (2011)

Gass, S., Selinker, L.: Language Transfer in Language Learning, pp. 22–113. John Benjamins Publishing Company, Amsterdam (1992)

Zhang, R.: Research on automatic evaluation method of Mandarin Chinese pronunciation. Harbin Institute of Technology, pp. 72–101 (2013)