Wireless Networks (2021) 27:3471–3481
https://doi.org/10.1007/s11276-019-02220-2

Analysis of phonemes and tones confusion rules obtained by ASR

Gulnur Arkin[1] · Askar Hamdulla[1] · Mijit Ablimit[1]

Published online: 1 January 2020

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 is also used to recognize phonemes of the pronunciation of Chinese mandarin learners. 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 pronunciation can be obtained, which may give some help to the learners to learn the Chinese auxiliary language system and the corresponding pronunciation evaluation model.

Keywords Non-mother tongue · Chinese mandarin · 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. 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, it is very important 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 [1–3]. In recent years, with the development of science and technology, online education has become more and more popular [4–6]. 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 [7–11]. 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.

The aim of this work is to develop automatic instruments for language learning. We attempt to develop a mispronunciation detection system to effectively highlight pronunciation errors made by Uyghur (L1) learners of Mandarin Chinese (L2). Our long-term goal is to design effective pedagogical and remedial instructions for pronunciation improvement. The target learners are Students who are native Uyghur speakers seeking to improve their pronunciation in mandarin Chinese. In [12] was conducted to discriminate a confusing pair of phonemes, e.g. the correct English pronunciation and its Japanese pronunciation marked by non-native speaker's accent. Another related work was presented in [13, 14], where the monophone substitutions for English pronunciation variability from Korean speakers were obtained by analyzing phonetics and speech recognition results. The phoneme confusions were used to modify the state-trying to adapt the acoustic models for accented speech recognition. In our research, a method is developed to identify the pronunciation errors at the phoneme level, so as to provide corrective feedback to support self-learning to improve English pronunciation using a CALL system.

✉ Askar Hamdulla
askar@xju.edu.cn

[1] Institute of Information Science and Engineering, Xinjiang University, Ürümqi 830046, China

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 [15]. 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 [16], and the Altaic language belongs to the agglutimother tongue language [17]. 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 [4]. When learners learn another language, the old phonetic perception and production systems play an auxiliary or interference role [18–21].

Since some phonemes in L1 pronunciation do not exist in L2 pronunciation, learners habitually replace the pronunciation by their mother tongue phonemes which are similar on pronouncing perception and generation [6]. Pronouncing mode and part as L1 pronunciation but still have certain differences, and the difference generated by this replacement will cause possible pronouncing confusion and errors. For instance, based on standard pronunciation phonemes, learners produce a continuous oral pronunciation when pronouncing a phoneme/kuo/sound. In this continuous oral pronunciation, phonemes/u/or/o/may be used instead of phoneme/uo/, mainly because of their similarities with/uo/in perception and production. This work focuses on automatically detecting such pronunciation errors caused by language transfer effects, using the speech recognition system with the predicted word mispronunciations on continuous speech. 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 [22]. For non-mother tongue speakers of languages, they attach particular importance to computer-aided L2 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 L1 whose mother tongue language belongs to the Altai language family, and they have some L2 learning experience. Uyghur pronunciation is different from Mandarin pronunciation, some of its pronunciation cannot be found in L2 pronunciation, based on this, learners habitually use L1 pronunciation as a benchmark to learn L2, focusing on pronunciation perception and aspects of producing, pronouncing ways and parts, finding the mother tongue phonemes that are similar but slightly different with L2 to replace the pronunciation, thus the difference caused by the substitution may cause confusion of pronunciation or certain pronunciation errors [23]. To sum up, through the collection of learners' pronunciations, we can get a comparative analysis of L1 cross-linguistic phonological contrast in linguistic and phonemic when they say L2 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.

This paper is organized as follows. In Sect. 2, we introduce briefly the Experimental data and preparation. In Sect. 3, we show the results and analyze for Phoneme confusion rules obtained by speech recognition system and learner's mobile phone APP. We end our paper with a brief conclusion in the last section.

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 to 26 years (Means 23), and their mother tongue 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 mother tongue-tongue-using Minority Nationality Students), and their learning time on Chinese mandarin is more than 10 years, their Chinese MHK oral test scores are above 45.

2.2 Experimental method

The learners may have non-knowledge errors in their actual pronunciation, to find this confusion, in this paper, we use a data-driven method [6, 24, 25] that does not depend on prior knowledge to generate additional confusion rules, and it is confused with the rules of prior knowledge as a supplementary phoneme.

Figure 1 shows the basic flow of the data-driven approach. The main method is, firstly to identify the phonemes of L2 learners whose mother tongue is L1 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 $x(i)|$ $y(j)$:

$$P\,x(i)|y(j) = (y(j)/x(i)) * 100\% \quad (1)$$

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

2.3 Experimental process

Each speaker was sitting in a sound booth during the experiment, and the microphone was 5 cm 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 16HZ. Participants' read materials are Chinese sentences, each participant needs to record 300 Mandarin sentences (50 * 300 = 15,000) and each sentence contains 5 to 11 words. The recognition results are obtained through Chinese speech recognition system of Tsinghua University. Our goal is to detect the categories of phoneme errors that may exist in the continuous oral pronunciation of L2 learners. Based on the theory of language transfer, between L1 and L2, in terms of linguistics, phonology and phonology, systematic contrastive analysis of translingual phonological differences in phoneme pronunciation location and mode, perception and production shall be done, predicting phoneme error rules that learners may cause mispronunciation. The following Table 1 is a detailed description of the project.

The objective of this work is to derive salient mispronunciations made by L1 learners of L2. Therefore, according to the method of this paper, in order to make it more convenient for L1 learners' to learn L2, and can carry it with them, we are developing an automatic evaluation application for real-time mispronunciation (the mobile phone APP), which is shown in the following series of figures (Fig. 2, 3, 4, 5). From the Figs. 2, 3, 4 and 5 one can see that, according to the given sentence and phoneme sequence pronunciation of the L2 speaker, the L1 learner can systematically identify the learner's L2 and phoneme pronunciation sequence and compare them, diagnose the types of pronunciation errors and provide scores. The range of score is excellent, ordinary and poor. The range of score more than 80 is excellent, the range of score more than 60 is ordinary and the range of score below 60 is poor.

Step one: Start the mobile phone APP, and you can see Fig. 2.

Step two: Click the start button to display the interface in Fig. 3. Compare and start grading learners 'pronunciation.

Step three: Click on the result button to appear the following interface (Figs. 4, 5, 6): L2 learners' speech recognized by the speech recognition system, and compared with the standard Mandarin speech to score. Learners can also correct their mistakes by repeatedly listening to the pronunciation of standard L2, which can achieve a

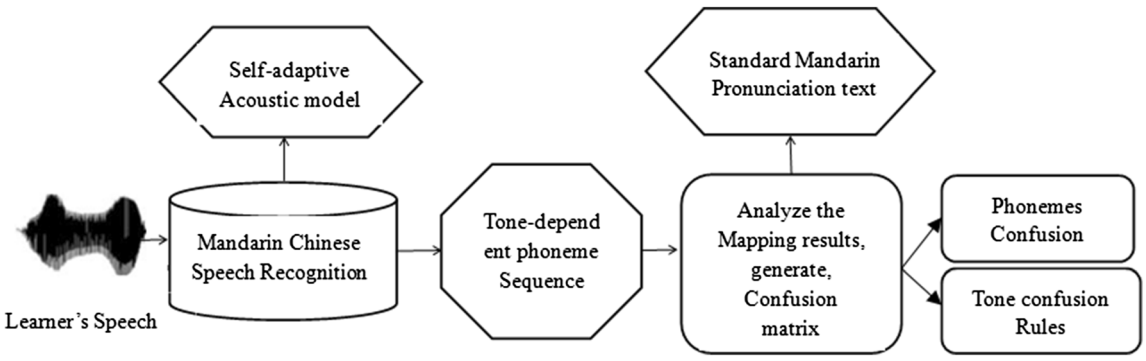

Fig. 1 The flow chart of the confusion rules of phonemes and tones generated by the results of ASR recognition is analyzed

Table 1 The specifications of the project

Project name	By using React and Redux, build a simple detection mobile phone APP
Technical support	JavaScript, React, Redux, Antd
Development environment	Windows 10
Development tools	WebStorm
Project description	React is a JAVASCRIPT library for building user interfaces. React has high performance, whose code logic is very simple, and more and more people have begun to pay attention to and use it. Redux is a JavaScript state container that provides predictable state management, which allows you to build consistent applications, running in different environments (client, server, mother tongue applications), and easy to test. Antd is a front-end CSS framework provided by Ant Financial Services Group's technical department. APP uses create-react-app scaffolding tools to create projects [26]. Complete the construction and rendering of the page in the form of React components (React advocates the concept of all components), the interface uses the Antd CSS framework for style rendering, and the data is stored in JSON format for local JSON files. Through the import of text resources and voice resources for React interface rendering [27], finally integrate the project to complete APP

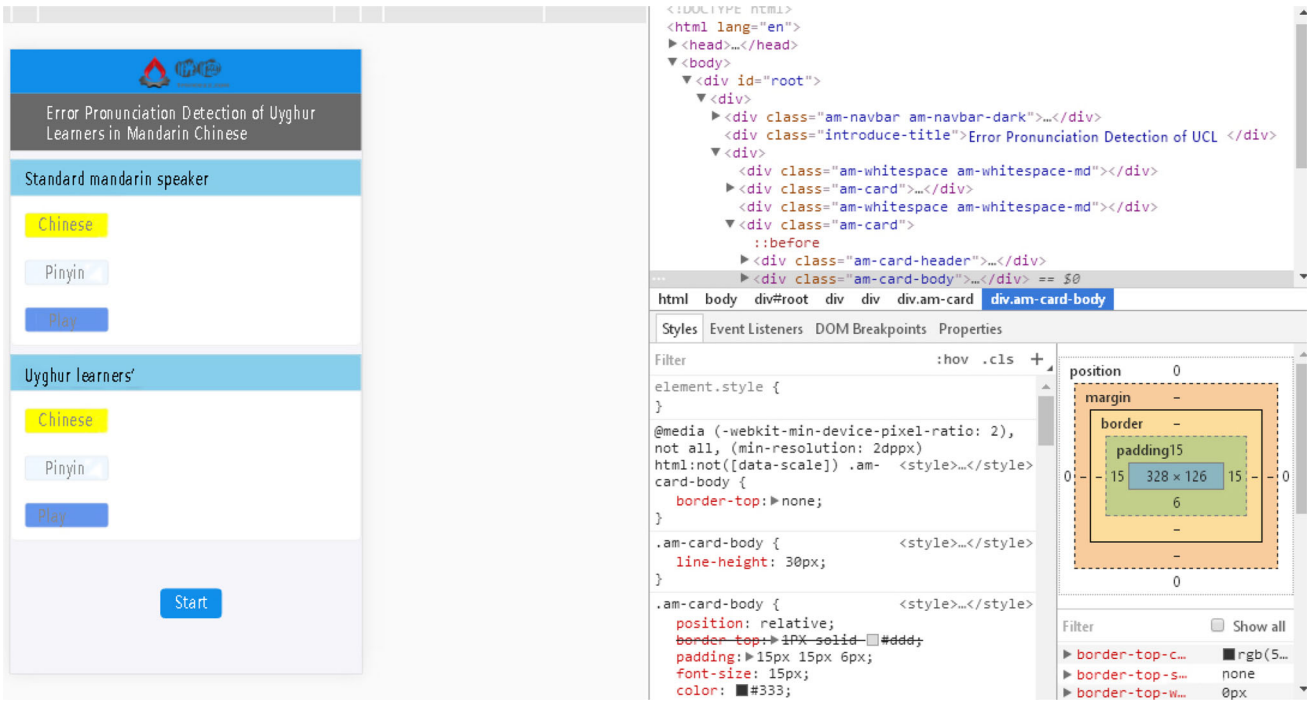

Fig. 2 Interface for the beginning of detection and scoring of L2 learners (first)

good learning opportunity. At the same time, this study is of great help to L2 learners and even foreigners who learn Mandarin.

Step four: Similarly, click on the next sentence, and then click on the results to see, there may also be ordinary and poor levels of learners:

Although the above research results have been achieved in the analysis of the pronunciation error stage of L1 learners', automatic pronunciation evaluation is after all a new immature research direction, and there are still many aspects worthy of further solution in the future.

3 Results and analysis

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

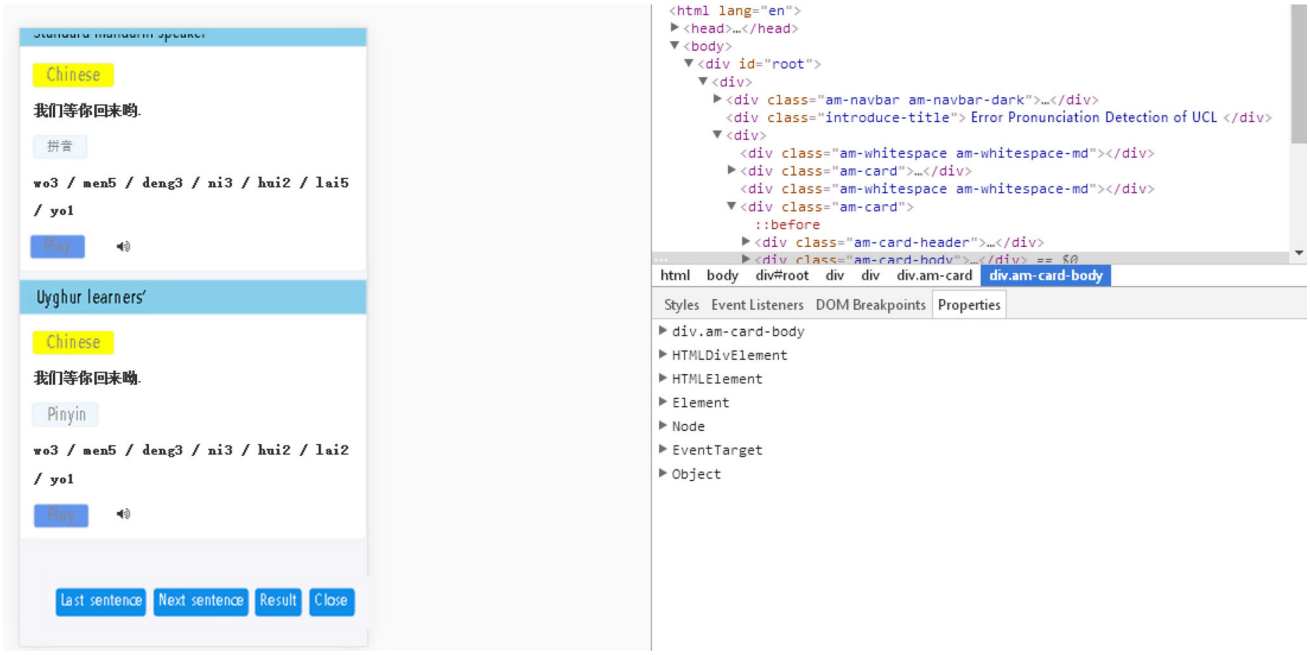

Fig. 3 Interface for the beginning of detection and scoring of L2 learners (second)

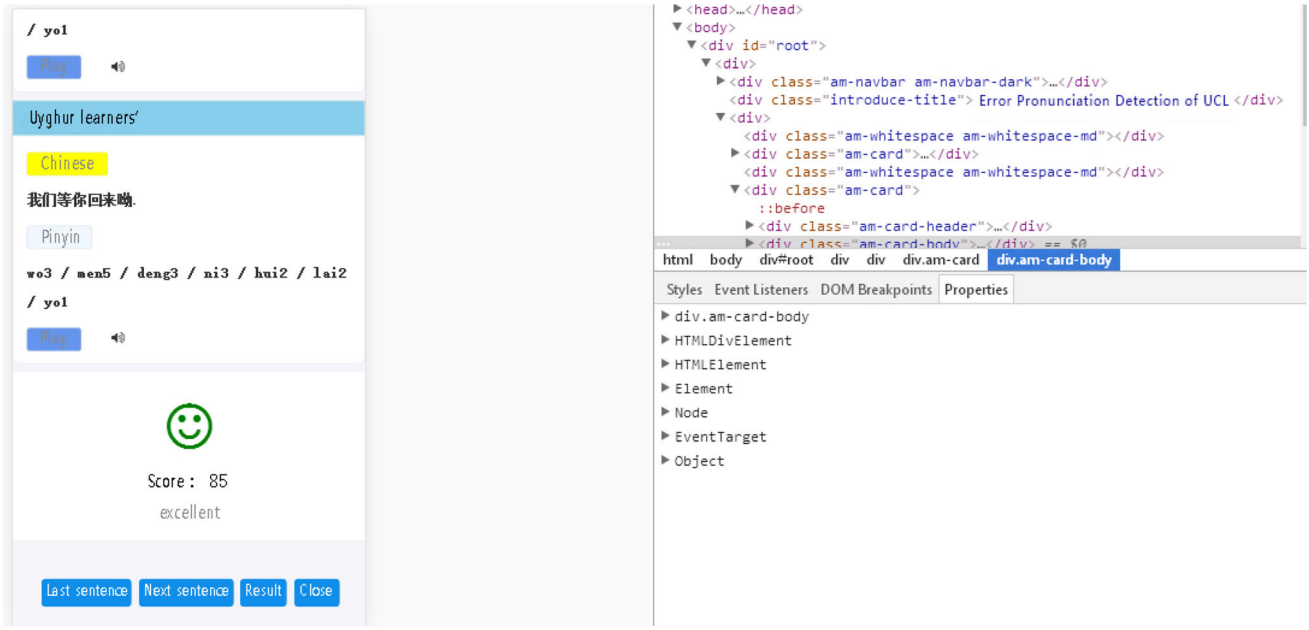

Fig. 4 Testing result and scoring of L2 learners (excellent)

due to the integration of these prior knowledge [22, 28–31], 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 [32]. The misunderstanding of non-knowledge among L1 learners 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 [33, 34], 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 [11, 15] it automatically generate the relevant rules, and this rule is mainly for additional phoneme confusion. Next, we use the above mobile phone APP results (that is, automatic speech recognition) to find out the list of

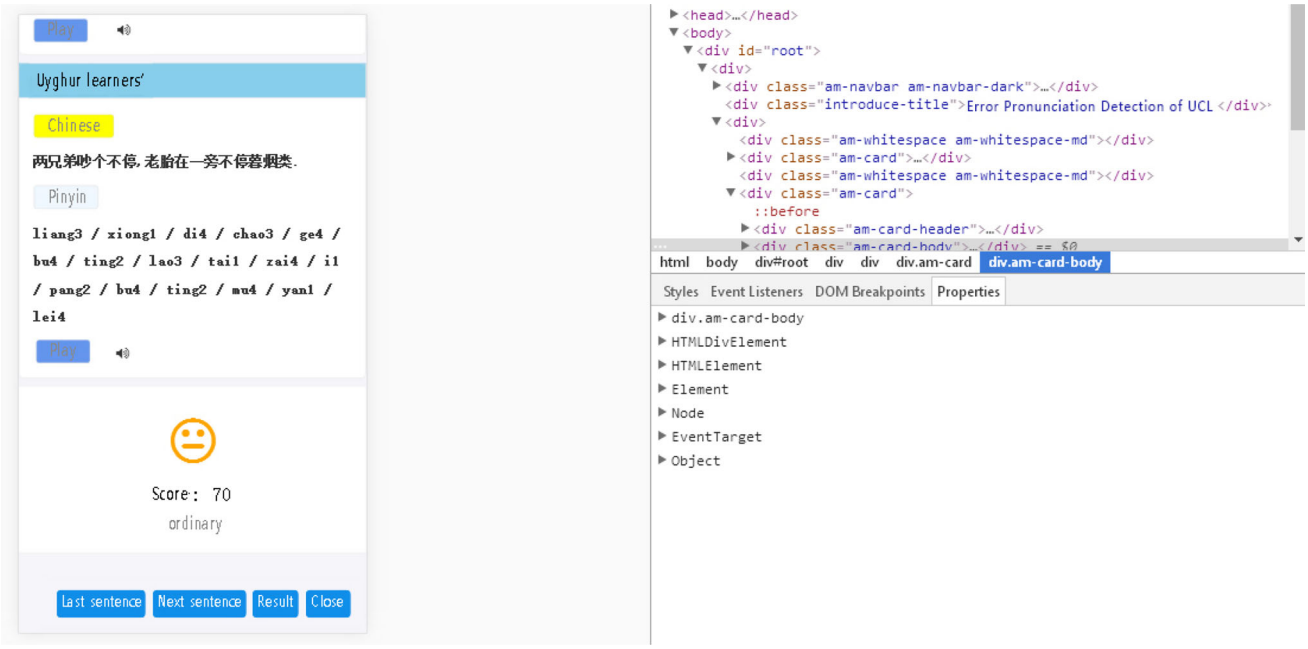

Fig. 5 Testing result and scoring of L2 learners (ordinary)

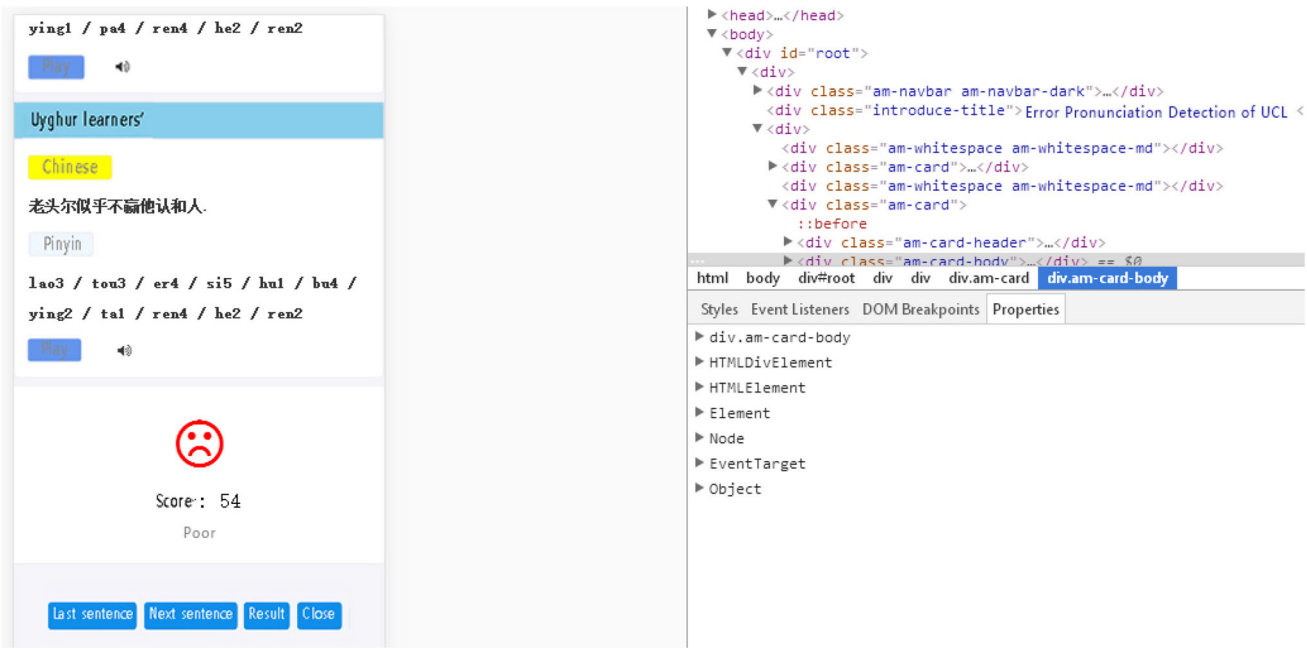

Fig. 6 Testing result and scoring of L2 learners (poor)

phonemes that may be mispronounced to generate a confusion matrix, and filter each monophthong, diphthong and consonant table mapping according to the threshold to generate their own confusion rules and perform error analysis. In a summary, we tabulate the possible phonetic confusions between L1 and L2 in a form of phone-to-phone mapping, which are then used to produce a lexicon with erroneous pronunciation variations.

3.1 Vowel confusion matrices and rules

The data-driven approach is mainly used here to get the rules of phoneme confusion, and the flow chart is as follows. These predicted confusions are incorporated into a pronunciation lexicon to generate additional, erroneous pronunciation variants of each word. Meanwhile we found out phoneme list with possible pronunciation mistakes by

Table 2 The 1st–12th columns of vowel confusion matrix obtained from data driven experiment

	a	o	e	i	u	v	er	i.e.	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
i.e.	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

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.

From table above, the confusable phone may be substituted, deleted or inserted in the continuous speech. 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 2 and 3, 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 (see Tables 2, 3).

3.2 Consonant confusion matrices and rules

The purpose of finding confusion matrices is to detect L1 learners' continuous spoken pronunciations and some phoneme errors. After phoneme confusion rules are constructed, the next step is to combine it with speech recognition around confusion rules. To prepare for the

Table 3 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, i.e., o	51.63
i	e,i.e.,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, i.e.	46.95
i.e.(ĕ)	i			i	60.72
ai	a,i	i		a, i, e, ei	52.58
ei	e			e, i	44.48
ao	a	a	u	o, ou, u, i.e.	45.96
ou	u	o		u, ao, ai, a	52.37
ia	i.e.	i,a		i.e., ou	49.72
i.e.					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

establishment of a set of automatic detection methods specifically for phoneme wrong pronunciation of Uyghur people learning L2 pronunciation and to prepare for preparation materials. Similarly, Tables 4, 5 and 6 below are confusion matrices and confusion rules of consonants.

Focusing on phoneme to phoneme, and the confusion rules they produce, we have replaces all standard pronunciation phonemes that can be confused in all words with a combination approach to generate extended pronunciation words that cover both standard pronunciation and possibly incorrect pronunciation dictionary. Speech recognition exploits the generated extended pronunciation dictionary to do linguistic constraints on the recognition of the phoneme level, so as to detect wrong phonemes of learner pronunciation. At the same time, it improves the accuracy, and the phoneme confusion and speech recognition also rely on it.

Table 4 The 1st–11st columns of consonant confusion matrix obtained from data driven experiment

	b	p	m	f	z	c	s	d	t	n	l
b	408	4	1	1	4	0	0	0	2	8	6
p	7	89	4	0	0	0	0	0	2	0	0
m	2	5	329	0	0	0	0	4	0	1	4
f	0	1	0	195	0	1	1	0	0	4	0
z	4	0	0	0	371	6	8	5	1	0	0
c	0	0	0	0	3	111	23	2	0	8	0
s	0	0	0	0	1	4	73	0	0	0	0
d	2	0	4	0	4	0	0	671	2	3	0
t	0	4	0	1	0	1	0	3	334	0	4
n	8	0	0	4	0	7	0	5	0	293	11
l	4	0	4	0	1	0	2	1	4	6	490
j	0	0	4	0	0	4	0	5	0	0	4
q	0	0	0	2	0	0	0	0	0	0	0
x	4	8	0	0	1	0	4	5	4	1	4
g	0	0	0	0	1	0	0	0	4	0	0
k	1	0	0	0	0	0	0	3	0	0	0
h	0	0	0	1	0	0	0	4	2	3	4
zh	8	0	4	0	3	0	0	4	0	0	3
ch	8	0	0	0	4	2	4	2	1	0	4
sh	0	6	0	0	4	1	8	0	3	2	6
r	4	0	0	0	0	0	0	0	0	0	0

Table 5 The 12st–21st columns of consonant confusion matrix obtained from data driven experiment

	j	q	x	g	k	h	zh	ch	sh	r
b	0	1	4	0	0	0	8	7	0	4
p	0	0	8	0	0	0	0	0	4	0
m	3	0	0	0	0	0	4	0	0	0
f	0	0	0	0	0	2	0	0	1	0
z	3	1	1	0	0	0	14	5	8	0
c	6	0	0	0	0	0	0	3	6	0
s	0	0	4	0	0	0	0	4	1	0
d	5	2	3	0	4	6	1	4	2	0
t	0	10	4	4	2	0	0	0	4	0
n	0	0	1	0	0	0	1	0	0	0
l	3	0	4	0	0	4	4	3	3	0
j	511	13	15	0	0	4	3	0	0	1
q	7	270	4	0	0	4	0	5	0	0
x	10	9	424	7	0	14	7	0	1	0
g	0	0	3	346	1	0	7	0	0	0
k	0	0	0	2	117	5	0	0	0	0
h	3	7	6	2	4	505	1	0	12	0
zh	0	0	8	8	0	1	419	7	9	0
ch	0	7	1	0	0	0	15	233	1	0
sh	0	0	4	2	0	12	3	1	602	0
r	0	0	0	0	0	0	1	0	1	185

4 Analysis of tone confusion rules

Chinese is a tonal language, and Mandarin consists of four standard tones, as well as soft tones, and the role of tone lies in distinguishing meaning and must exist in Chinese syllable structure. The Uyghur language belongs to phonograms, when they read exactly, they only need to look at it, while Chinese belongs to ideogram, they cannot figure out the pronunciation only from Chinese literal characters, not to mention to know the tone. Therefore, Uyghur learners simply focus on practicing the pronunciation of Chinese and do not place the tone in an important position.

For Uyghur students whose mother tongue language is non-tonal language, the tone is not easy to master for most learners. Chinese tone itself has a complex structure, which is not easy to master, at the same time it affects the pronunciation, and it should be learned, so the pronunciation teaching becomes very difficult. In view of this, we begin with the tone pronunciation recognition of continuous speech and recognize the tone of Uyghur learner's pronunciation. Here we will look at the tone confusion matrix of the Uyghur learner's speech recognition result and frequency. (T1—the 1st tone, T2—the 2nd tone, T3—the falling-rising tone, T4—the falling tone).

Table 7 shows the confusable tone and their correct frequency list. From the Table 7 we find that the average correct frequency rate is not sounds good for nonnative speakers every tone, respectively. Meanwhile when nonnative speakers learn Chinese tones, they less sensitive to change Chinese tones and they all have errors. Maybe among the important reasons are the following two: on the one hand; their mother language is non-tone language so this is results in lack of tone experience. On the other hand, their speaking rate, duration and intensity also affect the tones of the learners' speech [34–36].

The purpose is to know the students how pronounce in specific Chinese Mandarin, how to use the tone, focusing on the high degree of naturalness and high precision of speech synthesis and recognition, so as to achieve the goal of practical application of Mandarin in minority areas, to improve the overall education level in ethnic minority areas, it can also detect the pronouncing accuracy of phoneme in Uyghur spoken pronunciations, in order to provide a little reference for the above issues.

Table 6 The consonant confusion rules list obtained from data driven experiment

Target phoneme	Replace	Delete	insert	Misreading	ASR correct recognition probability (%)
b	p, t, l			p, f, t, l, n	48.68
p	b, t			b, t, m	47.30
m	n			b, n, p, l	45.00
f	n			c, s, p, n	44.52
z	s, c			c, s, b, d, t	52.92
c	z, s			z, s	53.88
s	z, c			z, c	36.13
d	b, t			b, z, t	44.00
t	p, f, d			l, p, c, d	49.77
n	f, m, ng			b, f, c, l, m, ng	4.92
l	b, z			t, n, g	48.03
j	q, zh			q, ch, zh, x, h	46.49
q	j, ch			x, j.h, ch	40.78
x	sh			j, qh, sh, g	48.29
g	k			zh, x, k	11.35
k	g			g, h	39.00
h	g, k			q, x, k, g, j	52.33
zh	ch, j			x, g, ch, j	47.29
ch	zh, q			zh, q, x, z, c	46.78
sh	x			g, zh, ch, x	44.46
r	l			b, l	50.40

Table 7 The confusable tone and frequency list

Tone	T1	T2	T3	T4	T5	Correct frequency
T1	1553	175	64	158	10	40.18%
T2	156	1668	95	166	19	37.35%
T3	153	127	1500	232	16	25.46%
T4	368	282	145	2700	13	45.83%
T5	54	258	99	250	488	48.90%

5 Conclusions

At present, the research on language learning and oral pronunciation testing (evaluation) is becoming more and more popular, and most scholars are keen on interactive language learning, especially for the automatic examination (evaluation) of oral pronunciation quality. In this paper, we mainly analyze the result s of 50 Uyghur learners through Chinese speech recognition. And in order to make it more convenient for Uyghur learners to learn Mandarin, by using the speech recognition system of the speech Technology Research Center of the Institute of Information Technology of Tsinghua University, a simple mobile phone APP is constructed by using React and Redux.

Secondly, it introduces in detail the two cases in which learners may mispronounce phonemes. First, according to the causes of learners' mispronunciation, the phoneme confusion is predicted by the method of cross-language phonological contrast, which is aimed at Uyghur learners. Their personal factors can also lead to confusion. Second, with the help of data-driven, the additional phoneme confusion is summarized and sorted out, and the relevant situation of the two methods is explained clearly. 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 next step, the work plan will expand the standard pronunciation dictionary by the resulting phoneme confusion rules and phoneme error categories, and lay a certain foundation for the study of learners' Putonghua evaluation system and Call (Computer Assisted Language Learning) system by combining the resulting speech rules with speech recognition.

Acknowledgments This work was supported by the National Natural Science Foundation of China (NSFC; Grants 61662078, and 61633013), National Key Research and Development Plan of China (2017YFC0820602).

References

1. Ito, A., Lim, Y.-L., Suzuki, M., & MaKino, S. (2007). Pronunciation error detection for computer-assisted language learning

system based on error rule clustering using a decision tree. *Acoustical Science and Technology, 28*(2), 131–133.
2. Stanley, T., & Hacioglu, K. (2012). Improving L1-specific phonological error diagnosis in computer assisted pronunciation training. In *INTERSPEECH 2012* (pp. 827–830). ISCA.
3. Wang, Y. B., & Lee, L. S. (2012). Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training. In *ICASSP 2012* (pp. 5049–5052). IEEE.
4. Jiang, D., Wang, W., Shi, L., & Song, H. (2018). A compressive sensing-based approach to end-to-end network traffic reconstruction. *IEEE Transactions on Network Science and Engineering, 5*(3), 1–12.
5. Jiang, M., Jiang, L., Jiang, D., et al. (2017). Dynamic measurement errors prediction for sensors based on firefly algorithm optimize support vector machine. *Sustainable Cities and Society, 2017*(35), 250–256.
6. Wang, F., Jiang, D., & Qi, S. (2019). An adaptive routing algorithm for integrated information networks. *China Communications, 7*(1), 196–207.
7. Jiang, D., Zhang, P., & Lv, Z. (2016). Energy-efficient multi-constraint routing algorithm with load balancing for smart city applications. *IEEE Internet of Things Journal, 3*(6), 1437–1447.
8. Jiang, D., Li, W., & Lv, H. (2017). An energy-efficient cooperative multicast routing in multi-hop wireless networks for smart medical applications. *Neurocomputing, 220,* 160–169.
9. Witt, S. M. (1999). *Use of speech recognition in computer-assisted language learning*. Cambridge: Cambridge University.
10. Ye, H., & Young, S. J. (2005). Improving the speech recognition performance of beginners in spoken conversational interaction for language learning. In *INTERSPEECH 2005* (pp. 289–292). ISCA.
11. Jiang, D., Huo, L., & Song, H. (2018). Rethinking behaviors and activities of base stations in mobile cellular networks based on big data analysis. *IEEE Transactions on Network Science and Engineering, 1*(1), 1–12.
12. Qian, X. J., Meng, H., & Soong, F. K. (2011). On mispronunciation lexicon generation using joint sequence multigrams in computer-aided pronunciation training (CAPT). In *INTERSPEECH 2011* (pp. 865–868). ISCA.
13. Tsubota, Y., Kawahara, T., & Dantsuji, M. (2002) Recognition and verification of English by Japanese students for computer-assisted language learning system. In *Proceedings of ICSLP* (pp. 1205–1208).
14. Oh, Y. R., Yoon, J. S., & Kim, H. K. (2007). Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. *Speech Communication, 49,* 59–70.
15. Peter Ladefoged, F., & Keith Johnson, S. (2015). *A course in phonetics* (7th ed.). Haidian: Peking University Press.
16. Thurgood, G., & La Polla, R. J. (2003). *The Sino-Tibetan languages*. London: Routledge.
17. Jiang, D., Huo, L., & Li, Y. (2018). Fine-granularity inference and estimations to network traffic for SDN. *PLoS ONE, 13*(5), 1–23.
18. Huo, L., Jiang, D., & Lv, Z. (2018). Soft frequency reuse-based optimization algorithm for energy efficiency of multi-cell networks. *Computers & Electrical Engineering, 66*(2), 316–331.
19. Xiangru, Z., & Zhining, Z. (1985). *Uyghur language*. China: National press.
20. Shifeng, F. (2009). *Experimental phonology exploration*. China: Peking University Press.
21. Lo, W. K., Zhang, S., & Meng, H. M. (2010). Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system. In *INTERSPEECH 2010* (pp. 765–768).
22. Zhu, J., Song, Y., Jiang, D., et al. (2018). A new deep-Q-learning-based transmission scheduling mechanism for the cognitive Internet of things. *IEEE Internet of Things Journal, 5*(4), 2375–2385.
23. Troung, K. F. (2004). *Automatic pronunciation error detection in Dutch as a second language: An acoustic-phonetic approach*. Utrecht: Utrecht University.
24. Jiang, D., Wang, Y., Lv, Z., et al. (2019). Big data analysis-based network behavior insight of cellular networks for industry 4.0 applications. *IEEE Transactions on Industrial Informatics*. https://doi.org/10.1109/tii.2019.2930226.
25. Jiang, D., Huo, L., Lv, Z., et al. (2018). A joint multi-criteria utility-based network selection approach for vehicle-to-infrastructure networking. *IEEE Transactions on Intelligent Transportation Systems, 19*(10), 3305–3319.
26. Huo, L., & Jiang, D. (2019). Stackelberg game-based energy-efficient resource allocation for 5G cellular networks. *Telecommunication System, 23*(4), 1–11.
27. Sun, M., Jiang, D., Song, H., et al. (2017). Statistical resolution limit analysis of two closely-spaced signal sources using Rao test. *IEEE Access, 2017*(5), 22013–22022.
28. Dong, B., & Zhao, Q. W. (2006). Automatic scoring of flat tongue and raised tongue in computer-assisted mandarin learning. In *ISCSLP 2006* (pp. 2–7). IEEE.
29. Chen, L., Jiang, D., Song, H., et al. (2018). A lightweight end-side user experience data collection system for quality evaluation of multimedia communications. *IEEE Access, 6*(2018), 15408–15419.
30. Sun, M., Jiang, D., Song, H., et al. (2017). Statistical resolution limit analysis of two closely-spaced signal sources using Rao test. *IEEE Access, 5,* 22013–22022.
31. Wang, S. J., & Li, H. Y. (2011). Research on the evaluation of spoken language scale intelligence for second language learning. *Chinese Journal of information science, 25*(6), 142–148.
32. Gass, S., & Selinker, L. (1992). *Language transfer in language learning* (pp. 22–113). Amsterdam: John Benjamins Publishing Company.
33. Wang, L., Feng, X., & Meng, H. M. (2008). Automatic generation and pruning of phonetic mispronunciations to support computer-aided pronunciation training. In *INTERSPEECH 2008* (pp. 22– 26).
34. Wang, L., Feng, X., & Meng, H. M. (2008). Mispronunciation detection based on cross-language phonological comparisons. In *ICALIP 2008* (pp. 307–311).
35. Arkin, G., & Hamdulla, A. (2018). Tone investigation of non-native Chinese speakers based on acoustic features. *Technical Acoustics, 37*(6), 572–578.
36. Arkin, G., & Hamdulla, A. (2018). Tone analysis of non-native Chinese speakers based on rules and statistics. *Journal of Applied Acoustics, 37*(3), 366–372.