The Contribution of Phonological and Fluency Factors to Chinese L2 Comprehensibility Ratings: A Case Study of Urdu-speaking Learners

Rian Bao[1], *Linkai Peng*[2], *Yingming Gao*[3], *Jinsong Zhang*[1]

[1]School of Information Science, Beijing Language and Culture University, Beijing, China
[2]Youdao, NetEase, Beijing, China
[3]School of Artificial Intelligence, University of Posts and Telecommunications, Beijing, China
borooooo@163.com, penglinkai96@gmail.com, yingming.gao@bupt.edu.cn, jinsong.zhang@blcu.edu.cn

Abstract

The present study set out to investigate the crucial phonological and fluency aspects that influence listeners' judgement of Chinese L2 comprehensibility. 180 speech samples elicited from 20 Urdu-speaking learners of Chinese were subjectively rated by native speakers of Chinese for comprehensibility scores, and then objectively analyzed in terms of phonological (segment error ratio, high & low FL error ratio, FL of segment substitutions, tone error ratio, and average maximum posterior possibility) and fluency features (articulation rate, pause ratio, FL of pause errors, and sentence length). The results showed that comprehensibility was significantly related to most of the features, and tone error ratio has the strongest correlation with comprehensibility judgements ($r = -.588$). Furthermore, multiple linear regression analyses revealed that these features have a combined contribution to the comprehensibility ($r^2 = .6286$). This study offers: (1) an empirical evidence for comprehensibility-related features of Chinese L2 speech, (2) the adaptation of Brown's FL principle on Chinese, (3) a strong proof that lexical tone errors greatly impair understanding when the segmental information is not sufficient.

Index Terms: second language comprehensibility, functional load

1. Introduction

Comprehensibility is an important measure for second language (L2) learners. It is defined as how easily and smoothly listeners understand L2 speech [1]. Since comprehensibility is based on listeners' perception and likely to simulate real-world interactions, it is an appropriate goal for successful communication and has been consistently emphasized in prior research [2, 3]. From a methodological point of view, L2 comprehensibility is most often measured based on listeners' intuitive judgements on a 9-point scale (1 = difficult to understand, 9 = easy to understand) [1, 4, 5].

To date, much work has shown that comprehensibility judgement of L2 English is mostly linked to phonological and fluency aspects. For example, segmental errors [1], word stress [6], speech rate [7, 8], and pause errors [9, 10] were proved to be associated with comprehensibility judgements. Segmental errors can be further divided as high functional load (an information-theoretic measure that computes contribution of phonological contrast to successful word identification, hereafter referred to as FL) errors and low FL errors, and high FL segmental errors were proven to impair comprehensibility more than low FL errors [11, 12]. Apart from phonological and fluency features, comprehensibility is also linked with lexical and grammatical features in extemporaneous tasks such as picture description [1, 13], IELTS or TOEFL speaking task [14] and argumentative speech [15]. Lexical and grammatical features are proven to be contingent on task complexity, and the contribution grows as complexity increases while phonological and fluency features of L2 speech have a consistent contribution to all tasks [14].

Apart from focusing on individual linguistic features, a range of studies has sought to determine the combined contribution of the measures mentioned above to L2 comprehensibility. Isaacs and Trofimovich have shown that type frequency, token frequency and word stress error contribute most to comprehensibility out of 19 linguistic measures [13]. Recently, automatic speech recognition technology deep neural network was also used to model English L2 comprehensibility [16], and averaged maximum posterior probabilities and averaged posterior gaps to natives were calculated, which demonstrated significant contribution to the model, and the model explained 67.7% of the variance in the listeners' comprehensibility.

As for Chinese Mandarin (Throughout the paper, the term Chinese will be used to specifically refer to Mandarin Chinese), some pilot studies investigated the relationship between linguistic features and comprehensibility. Segment features were proven to contribute more to comprehensibility compared to intonation and rhythm [17, 18], while lexical tone was shown to have no significant inhibitory effect on comprehension in quiet environments [19, 20]. One of the shortcomings of these studies is that the research materials were mostly synthesized speech and the content was repetitive, which might affect comprehensibility judgement. Furthermore, other comprehensibility-related features of Chinese have not been evaluated so far. For this reason, more research is needed to unpack what features are related to L2 Chinese comprehensibility and how these features account for overall comprehensibility.

Clearly, when investigating linguistic features that underlie Chinese L2 comprehensibility, the characteristics of Chinese cannot be ignored. There are two striking characteristics of Chinese that are different from English. First, Chinese is a tonal language, in which different tones are used to distinguish lexical meanings. Research suggests tone errors are very frequent among L2 Chinese speakers [21]. Second, written Chinese is character-based with each character representing a morpheme-syllable, and there are no visible word boundary demarcations in Chinese text [22]. Therefore, the process of word segmentation is needed before word identification, and this is also a difficulty for second language learners in the read speech [23].

To understand the construct of Chinese L2 comprehensibility, the current study is going to examine the related features based on knowledge of the comprehensibility-related aspects of English and Chinese-specific aspects. The present study aimed at: First, designing an experiment to find out which features contribute to Chinese L2 comprehensibility; Second, discovering to what degree these features simulate native listeners' comprehensibility judgement; Pearson's correlates coefficient and Multiple linear regression were used to determine the relative weights of the features.

2. Method

2.1. Materials

The speech materials used in present study were selected from BLCU-SAIT corpus [24], which is an interlanguage speech corpus of L2 learners of Chinese, contains simple sentence recordings from each speaker. 180 sentences from 20 Urdu-speaking learners (12 male and 8 female) from Pakistan were selected in the present study, whose mispronunciation are relatively diverse compared with other learners (for example, Thai- and Korean-speaking learners). All the speakers were students from Beijing Language and Culture University, and 9 sentences were selected from each speaker. The selected speech recordings contain different categories of pronunciation errors, such as segment, tone, and pause errors. The average sentence length is 13.2 syllable (SD = 4.41). 90 sentences were used to analyze related features for comprehensibility and were used for linear regression model, and the other 90 were used to validate the predictive power of the model.

2.2. Raters and comprehensibility judgements

We recruited six graduate students to participate in the comprehensibility rating experiment. All the raters are native speakers of Chinese, were all born in China and raised by monolingual parents, and none of them reported hearing disorder. All of them have no experience of teaching Chinese to speakers of other languages. They were all aged between 22 and 26 (M=23.5, 3 females and 3 males).

The comprehensibility rating tasks were conducted individually in a quiet room using the Praat's ExperimentMFC [25], and all the rating results can be automatically recorded in the software. Each rater listened to the audio through a set of headphones on the researcher's laptop. Before the data collection, the investigator trained all the raters. First, the raters familiarized themselves with the listening materials (9 sentences with standard comprehensibility ratings). Then, each rater practiced the rating procedure through three trails of experiments. In the formal experiment, they were asked to pay attention on the effort it takes to understand the sentences. If they can understand the sentence very easily, then this sentence is highly comprehensible, and vice versa.

In the formal rating experiment, 180 sentences were divided into two groups following the principle of non-repetition of the sentence content. Each group was rated by three raters. 102 sentences were played for each rater in a randomized order. 90 of them are from our research materials, and the other 12 sentences from other L2 speakers outside the scope of current study were included in both groups and used as computing inter-rater agreement. Each sentence can be played only one time. After hearing a sample, they made an intuitive judgement using a 9-point scale (1 = hard to understand, 9 = easy to understand). The whole session took around 30 minutes.

2.3. Comprehensibility measures

The phonological features extracted for comprehensibility-based analysis are segment error ratio, the number of high & low FL error ratio, FL of segment substitutions, tone error ratio, and averaged maximum posterior possibility. Fluency features extracted are articulation rate, pause ratio, and sentence length, and FL of pause errors is also computed. As for lexical and grammatical features, since the materials used here are read speech of simple sentences with no grammatical error, we don't include any lexical and grammatical features in this study. The measures used to calculate all the features are listed below:

- Segment error ratio: All the initial and final errors and their substitutions were annotated by a professional linguistic student, then the number of substitutions were divided by the sum of total initials and finals.
- High or low FL segmental error ratio: We applied Brown's FL principle [26] to Chinese, and the *cumulative frequency*, *probability of occurrence*, *occurrence and stigmatization in native accents*, *acoustic similarity*, *structural distribution of phonemes*, *lexical sets*, *number of minimal pairs*, *number of minimal pairs belonging to the same parts of speech*, *frequency of members of minimal pairs*, *number of common contexts in which members of minimal pairs occur* and *phonetic similarity* were adopted from the original model and calculated from a pinyin-transcribed corpus from Chinese TV show contains 300,000 sentences. All the measures were weighted equally by calculating the standardized score (T-score). After the calculation, we ranked the FL scores on the 10-point scale. Since there are almost two thousand initial and final pairs, the phonemic pairs of initials and finals shown in Table 1 are the ones which are conflated by the L2 speakers in the current study. Using this FL rank-ordering, we further divided the contrasts into high and low FL contrasts: 1 to 5 as low FL, and 6-10 as high FL. High or low FL segmental error ratio was defined as the number of high or low FL phonemic substitutions divided by the sum of total number of initials and finals.
- FL of segment substitutions: This was defined as the sum of the FL rank scores (in Table 1) of all the segmental substitutions divided by the total number of syllables articulated.
- Tone error ratio: This was defined as the number of tone substitutions divided by the total number of syllables articulated. Tone substitutions were annotated by professional linguistic students.
- Articulation rate: Articulation rate is equal to the number of syllables per second, which was calculated by dividing the number of syllables by the duration without pauses.
- Pause ratio: Pause ratio was calculated by dividing the number of unfilled pauses (longer than 400 ms) by sample duration without the silence before and after the sentence.
- FL of pause errors: Chen et al. [27] proposed a FL method to measure the contribution of Chinese prosodic boundary. A wrong-boundary or pausing error that causes ambiguity would be calculated as high FL in this model. Therefore, this model can be used to quantify the severeness of wrong pauses. FL of pause errors was defined as the increase of mutual information (MI) between text transcription and pinyin caused by wrong prosodic boundaries adding. When a wrong boundary is added between a word, the word hypothesis graph (WHG) will decrease and the MI between text transcription and pinyin will increase. MI is defined

Table 1: *Rank ordering of Chinese initial and final pairs conflated by L2 learners in the current study.*

FL rank	*Initials*	*Finals*
10	sh-zh, q-j, q-x	i-u, i-ing, u-ai, u-e, ian-ing, u-a, i-in, i-ie, u-uei, i-ai, eng-ang, i-a, u-ou, i-e, ian-in, ao-e, i-uei
9	h-g, t-d, ch-zh, ch-sh	i-ia, ao-ou, an-ai, ang-ong, ao-uo, ian-an, an-a, an-en, ian-iang, uan-an, uei-uo, u-uen
8	z-d, q-b	e-en, ang-a, uo-ong, ing-iang, a-e, eng-e, u-iou, ai-iao, ao-iao, ü-iou, ang-ing, ü-u
7	k-g, h-k	ua-e, ou-uo, ai-ei, uan-uen, ie-e, eng-ing, uan-ian, ian-uen, iou-üe, ong-uang, iao-uo
6	p-b, c-s, s-z, l-r	uang-ang, ei-ie, ü-ou, uen-uo, iang-eng, u-o, en-in, a-o, iong-iang
5	c-sh, c-z, z-zh	o-ao, uo-ua, uai-iao, an-ang, ou-o, u-uo
4	g-q, c-ch	in-ing
3	zh-j	en-eng
2	x-sh	-
1	ch-q, s-x	ang-iang

in Equation 2, where $\mathrm{MI(W, F)}$ and $\mathrm{MI(W, F_\alpha)}$ is the MI before and after all phonemes of α are merged, and $W_1^{'}, W_2^{'}, \ldots, W_m^{'}$ are all text sequences sharing the same transcription F. $P(W_i^{'})$ is the probability of the text sequence. FL of a wrong prosodic boundary is defined in (2), where α represents prosodic boundary. The FL is computed based on tri-gram language model trained with the same Chinese TV corpus used in the measure of FL of initials and finals.

$$\mathrm{MI}(W, F) = \lim_{n \to \infty} -\frac{1}{n} \log \sum_{i=1}^{m} \mathrm{P(W_i^{'})}, \quad (1)$$

$$\mathrm{FL_{MI}}(\alpha) = \frac{\mathrm{MI}(W, F_\alpha) - \mathrm{MI}(W, F)}{\mathrm{MI}(W, F)} \quad (2)$$

- Average maximum posterior probability: We trained a TDNN-GMM [28] based system with the AISHELL-1 dataset. Open source Kaldi toolkit [29] and the corresponding scripts[1] are used and we achieved word error rate of 8.9% on the test set. Once the model is trained, the maximum posterior probability score of given utterances on initial-final level at each time was selected, and then averaged over time, i.e., average maximum posterior probabilities. The maximum posterior probability score is a phonological measure that indicates a speaker's intended pronunciation. The higher the average is, the more distinct the pronunciation of the utterance is.
- Sentence length: This was defined as the total number of syllables articulated in a sentence.

[1] https://github.com/kaldi-asr/kaldi/blob/master/egs/aishell/s5/run.sh

3. Results

3.1. Correlations between comprehensibility and features

To achieve the first goal of this study, which is to find the features contribute to Chinese L2 comprehensibility, we examined the correlation between ten phonological and fluency features and listener-based comprehensibility judgements, and all the features were log-transformed.

Table 2: *Pearson correlation coefficients between L2 speech features and listener-based judgement of L2 comprehensibility* ($*** \leq 0.001, ** \leq 0.01, * \leq 0.05$).

Features	*Corr.*	*Sig.*
Tone error ratio	-.588	***
Pause ratio	-.532	***
Segment error ratio	-.509	***
Articulation rate	.468	***
High FL segmental error ratio	-.444	***
FL of segment substitutions	-.444	***
Average maximum posterior possibility	.428	***
FL of Pauses	-.291	***
Low FL segmental error ratio	-.201	**
Sentence length	.031	

In terms of inter-rater reliability, Pearson's r was computed among three raters' scores from the same group. The strength of correlations is relatively high, which varied from r = .658 to r = .828, and the inter-group reliability is r = .805. Then Pearson correlations were computed to examine the strength of the relationship between mean comprehensibility and ten features (see in Table 1). The result shows that most of the features have significant correlations with comprehensibility. Tone error ratio (r=-.588), pause ratio (r=-.532) and segment error ratio (r=-.509) have the strongest correlation with comprehensibility, while sentence length doesn't show significant correlations.

3.2. Comprehensibility modeling

To answer the second question of this study, stepwise multiple linear regression model was used to examine the extent to which L1 raters' comprehensibility judgements can be tied to all the features. The final model summary of stepwise multiple regression of all the variables can be seen in Table 1.

The model tested here is: Average comprehensibility scores given by three raters = Intercept + tone error ratio + pause ratio + segment error ratio + articulation rate + high FL segmental error ratio + FL of segment substitutions + average maximum posterior probability + FL of pauses + low FL segmental error ratio + sentence length.

The model significantly explained 62.86% of the variance in the raters' comprehensibility judgement, F (5, 84) = 31.13, $p < .001$. The best predictive feature selected in the first step is tone error ratio (B = -2.7066) followed by pause ratio (B = -1.7212), segment error ratio (B = -1.1046), sentence length (B = 1.8247), FL of pauses (B = -0.1918) and average maximum posterior probability (B = 5.69467). High & low FL segmental error ratio, FL of segment substitutions, articulation rate, and average maximum posterior probability were excluded from the model in the stepwise regression analysis.

We further calculated the predicted comprehensibility scores based on the result of the regression model in Table 3 (raw predictor values were multiplied by unstandardized B), and then compared the predicted scores with raters' scores (see in

Table 3: *Results of stepwise multiple regression analysis for predicting comprehensibility scores ($r^2 = 0.6286, p < .001$).*

Predicted variables	***Predictor variables***	***Unstandardized B***	***T***	***p***
Comprehensibility scores	Constant	-2.1270	-2.126	< 0.05*
	Tone error ratio	-2.7066	-7.368	< 0.001***
	Pause ratio	-1.7212	-3.799	< 0.001***
	Segment error ratio	-1.1046	-4.258	< 0.001***
	Sentence length	1.8247	2.233	< 0.05*
	Log-FL of pauses	-0.1918	-1.757	0.083

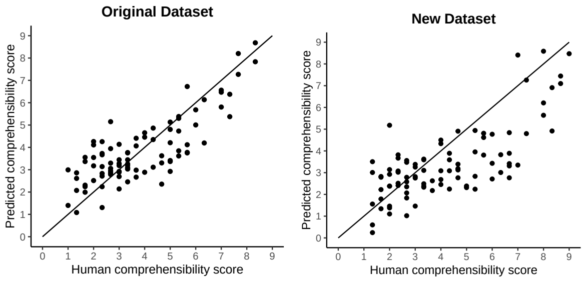

Figure 1: *The relationship between listener-based comprehensibility scores and predicted comprehensibility scores (left: original dataset, r=0.806; right: new dataset, r=0.738).*

left side of Figure 1). A significant correlation was found between predicted scores and human comprehensibility scores (r = 0.806, $p < .001$), which was comparable with the correlations between human raters.

3.3. Model Validation

The regression model demonstrated that these phonemic and fluency features can successfully simulate listener-based comprehensibility judgements. To further validate the predictability of these features for comprehensibility, we applied the regression formula to a new L2 speech data from the other ten Pakistani students as mentioned in 2.1. The correlation between predicted scores and listener-based scores was relatively strong (r = .738, visually summarized in right side of Figure 1.), which is also close to inter-rater agreements. The results indicate that these features successfully simulated human listeners' judgements.

4. Discussion

In this study, the comprehensibility of Chinese Mandarin L2 speech was evaluated by human raters, and then several features of the speech were extracted and calculated to study how they are related to the comprehensibility scores.

Tone accuracy has relatively strong correlation with comprehensibility, indicating that lexical tone has a strong impact on Chinese L2 comprehensibility rating. Previous research has shown that missing of tone information does not impede understanding with complete context in quiet [19, 20], but tone features came out as the strongest correlation in current study. The current result does not conflict with prior conclusions, because our result shows that tone has a strong inhibitory effect on comprehensibility when context information is not complete.

High FL segmental errors showed stronger correlation with comprehensibility than low FL errors, which is consistent with previous research [12]. When utilizing the FL rank as a continuous variable, it also showed significant correlation with comprehensibility judgements, indicating that this is an effective quantitative method for relative importance of different levels of phonological components. Furthermore, the current study took a first step to seek the impact of wrong prosodic boundaries on comprehensibility of Chinese speech. Wrong prosodic boundary may cause wrong sentence parsing, and the likeliness of the ambiguity caused by wrong pauses is what we call FL in current study, which showed a mild correlation with comprehensibility, and have a contribution to the linear model. Future studies should further pursue to what extent the prosodic boundaries affect comprehension.

Articulation rate and pause ratio have relatively strong correlation with comprehensibility in line with prior research [13, 15, 16]. Average maximum posterior probability didn't show as strong correlation with L2 comprehensibility as that of English [16]. A possible reason is that the average maximum posterior probabilities of sentences from same speaker are similar, so it can only distinguish the overall distinctiveness of pronunciation between different people. Since the dataset in current study contains only 20 speakers, the accurate comprehensibility scores are difficult to be obtained by this means.

5. Conclusion

In this paper, Chinese L2 comprehensibility-related factors were evaluated in the context of simple-sentence read speech of Pakistani students. There has been a lot of research on English L2 comprehensibility and its related features. Due to the scarcity of resources, there are few systematic studies on Chinese L2 comprehensibility.

Our results highlight that when interacting with other factors, tone contributes strongly to comprehensibility, and its correlation with comprehensibility is stronger than segmental features and fluency features. We took a first step to examine comprehensibility-related factors based on diverse speech materials, and the combined contribution of the different measures to Chinese L2 comprehensibility was determined. Our result showed that these phonemic and fluency features explained 62.86% of comprehensibility score. The regression model is able to provide predicted comprehensibility scores since the correlation between the predicted scores and listener-based scores (r = .738-.803) are comparable to inter-listener agreement (r = .658-.828), which provided insights for automatic assessment of Chinese L2 comprehensibility.

6. Acknowledgements

This study was supported by advanced Innovation Center for Language Resource and Intelligence (KYR17005), and Wutong Innovation Platform of Beijing Language and Culture University (19PT04), and the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University (21YCX177). Jinsong Zhang is the corresponding author.

7. References

[1] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language learning*, vol. 45, no. 1, pp. 73–97, 1995.

[2] J. M. Levis, "Changing contexts and shifting paradigms in pronunciation teaching," *TESOL quarterly*, vol. 39, no. 3, pp. 369–377, 2005.

[3] K. Saito, S. Webb, P. Trofimovich, and T. Isaacs, "Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness, and sense relations," *Studies in Second Language Acquisition*, vol. 38, no. 4, pp. 677–701, 2016.

[4] T. M. Derwing and M. J. Munro, "Accent, intelligibility, and comprehensibility: Evidence from four l1s," *Studies in second language acquisition*, vol. 19, no. 1, pp. 1–16, 1997.

[5] P. Trofimovich and T. Isaacs, "Disentangling accent from comprehensibility," *Bilingualism: Language and Cognition*, vol. 15, no. 4, pp. 905–916, 2012.

[6] J. Field, "Intelligibility and the listener: The role of lexical stress," *TESOL quarterly*, vol. 39, no. 3, pp. 399–423, 2005.

[7] J. Anderson-Hsieh and K. Koehler, "The effect of foreign accent and speaking rate on native speaker comprehension," *Language learning*, vol. 38, no. 4, pp. 561–613, 1988.

[8] M. J. Munro and T. M. Derwing, "Modeling perceptions of the accentedness and comprehensibility of l2 speech the role of speaking rate," *Studies in second language acquisition*, vol. 23, no. 4, pp. 451–468, 2001.

[9] O. Kang, D. Rubin, and L. Pickering, "Suprasegmental measures of accentedness and judgments of language learner proficiency in oral english," *The Modern Language Journal*, vol. 94, no. 4, pp. 554–566, 2010.

[10] D. Crowther, P. Trofimovich, and T. Isaacs, "Linguistic dimensions of second language accent and comprehensibility: Nonnative listeners' perspectives," *Journal of Second Language Pronunciation*, vol. 2, no. 2, pp. 160–182, 2016.

[11] M. J. Munro and T. M. Derwing, "The functional load principle in esl pronunciation instruction: An exploratory study," *System*, vol. 34, no. 4, pp. 520–531, 2006.

[12] Y. Suzukida and K. Saito, "Which segmental features matter for successful l2 comprehensibility? revisiting and generalizing the pedagogical value of the functional load principle," *Language Teaching Research*, vol. 25, no. 3, pp. 431–450, 2021.

[13] T. Isaacs and P. Trofimovich, "Deconstructing comprehensibility: Identifying the linguistic influences on listeners' l2 comprehensibility ratings," *Studies in Second Language Acquisition*, vol. 34, no. 3, pp. 475–505, 2012.

[14] D. Crowther, P. Trofimovich, T. Isaacs, and K. Saito, "Does a speaking task affect second language comprehensibility?" *The Modern Language Journal*, vol. 99, no. 1, pp. 80–95, 2015.

[15] S. Suzuki and J. Kormos, "Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech," *Studies in Second Language Acquisition*, vol. 42, no. 1, pp. 143–167, 2020.

[16] K. Saito, K. Macmillan, M. Kachlicka, T. Kunihara, and N. Minematsu, "Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies," *Studies in Second Language Acquisition*, pp. 1–30, 2022.

[17] L. Zhang, "Effect of segmental and prosodic information for accentedness and comprehensibility of chinese l2 speech," *Chinese Teaching in The World*, vol. 29, no. 2, p. 8, 2015.

[18] C. Yang, J. Chu, S. Chen, and Y. Xu, "Effects of segments, intonation and rhythm on the perception of l2 accentedness and comprehensibility," in *The Acquisition of Chinese as a Second Language Pronunciation*. Springer, 2021, pp. 233–255.

[19] A. D. Patel, Y. Xu, and B. Wang, "The role of f0 variation in the intelligibility of mandarin sentences," in *Speech Prosody 2010-Fifth International Conference*, 2010.

[20] J. Wang, H. Shu, L. Zhang, Z. Liu, and Y. Zhang, "The roles of fundamental frequency contours and sentence context in mandarin chinese speech intelligibility," *the Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. EL91–EL97, 2013.

[21] N. F. Chen, D. Wee, R. Tong, B. Ma, and H. Li, "Large-scale characterization of non-native mandarin chinese spoken by speakers of european origin: Analysis on icall," *Speech Communication*, vol. 84, pp. 46–56, 2016.

[22] R. Hoosain, "Psychological reality of the word in chinese," in *Advances in psychology*. Elsevier, 1992, vol. 90, pp. 111–130.

[23] D. Shen, S. P. Liversedge, J. Tian, C. Zang, L. Cui, X. Bai, G. Yan, and K. Rayner, "Eye movements of second language learners when reading spaced and unspaced chinese text." *Journal of Experimental Psychology: Applied*, vol. 18, no. 2, p. 192, 2012.

[24] B. Wu, Y. Xie, L. Lu, C. oCao, and J. Zhang, "The construction of a chinese interlanguage corpus," in *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2016, pp. 183–187.

[25] P. Boersma and D. Weenink. (1992) Praat: doing phonetics by computer [computer program].

[26] A. Brown, "Functional load and the teaching of pronunciation," *TESOL quarterly*, vol. 22, no. 4, pp. 593–606, 1988.

[27] Y. Chen, Y. Xie, B. Wu, and J. Zhang, "A study on functional load of chinese prosodic boundaries under reduction of syllable information," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.

[28] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.