The Importance of Lexical Tone for Sentence Understanding: Utilizing Functional Load Principle to Simulate Comprehension Process

Rian Bao
School of information Science
Beijing Language and Culture University
Beijing, China
boroooo@163.com

Linkai Peng
Youdao
NetEase
Beijing, China
penglinkai96@gmail.com

Yingming Gao
School of Artificial Intelligence
University of Posts and Telecommunications
Beijing, China
yingming.gao@bupt.edu.cn

Jinsong Zhang
School of information Science
Beijing Language and Culture University
Beijing, China
jingsong.zhang@blcu.edu.cn

***Abstract*—The missing information of lexical tone of Chinese has been proved to have no inhibitory effect on understanding in quiet environment in previous study. The current study set out to examine the importance of tone for speech comprehension when the contextual information is incomplete. The first experiment examined the correlation between second language (L2) speech comprehensibility and tone error ratio through 180 L2 speech samples with severe segmental and suprasegmental errors. The second experiment investigated the correlation between functional load of missing tone and word choices of sentences with tone-caused ambiguity pinyin sequences without tone information, and 30 native Mandarin speakers were asked to dictate the 14 ambiguous level-tone sentences. The results showed that tone error ratio strongly correlated with comprehensibility, and the word choice is also significantly correlated with the FL of the tones, suggesting that tone is not a redundant information in communication, and its ambiguity resolution function can be measured through FL model.**

***Index Terms*—tone information, functional load**

I. Introduction

Mandarin Chinese has five lexical tones, namely, high level (T1), mid rising (T2), low dipping (T3), high falling (T4) and neutral (T5), which play important roles in distinguishing lexical meanings. The importance of tone in speech recognition has been measured by reduction of word uncertainty [1], and the result showed that conditioning on tone information can reduce word uncertainty in conversational speech by 11%-20%. An entropy-based functional load (an information-theoretic measure that computes phonological contrast's contribution, hereafter referred to as FL) study has shown that lexical tone contrast has a comparable FL to that of vowels for Mandarin [2], [3]. Another FL study based on mutual information of Chinese text and phonemes has found FLs of some tone contrasts are much larger than that of phoneme pairs [4]. In some other studies, lexical tone was shown to have no significant inhibitory effect on comprehension in quiet environments [5], [6]. Since understanding spoken language is often thought of as an interaction procedure between top-down information of linguistic context and bottom-up information of perceptual input [7], listeners can guess the content through segmental information. To examine whether tone is important for comprehension when the contextual information is severely damaged, the first goal of the current study is to examine the contribution of tone errors to the understanding of L2 speech with severe segment and tone errors.

Another extreme condition is that in certain context, missing of tone information might cause ambiguity. When the phonetic input is ambiguous, listeners get the most likely meaning the speakers intend to express from the contexts. It has been proved that contextually appropriate meanings of ambiguous words are activated more strongly than inappropriate ones [8], suggesting some interactive processing in lexical ambiguity resolution. To quantify "appropriateness" on the sentence level, a subjective rating was conducted in the previous study [9]. The sentences were rated on the scale of 1-7 in terms of semantic plausibility. When bottom-up information cannot resolve the conflict alone, then sentential plausibility of speech comes into play. More "appropriate" in a context can also be interpreted as more frequent in this context. In the Neighborhood Activation Model (NAM), the critical assumption that words' frequency are combined with bottom-up word recognition [8], [10], and increased frequency of the components of spoken stimuli facilitates speech processing [11], which means when there are more than one lexical

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.

candidates, listeners tends to choose the more frequent one. In this paper, we intend to examine a new method, namely, FL model, to objectively quantify the "appropriateness" of the possible sentence candidates for ambiguous speech and investigate its relation between listeners' decisions.

The traditional methods for calculating FL were based on word-level frequency and entropy [2], which cannot sufficiently model the top-down process of word identification. The FL measurement based on mutual information between text and its phonetic transcription [4] offers the possibility to model the word-level and sentence-level context effects. The contribution of a phonemic contrast for a specific sentence can be computed using this model. We predict that if an ambiguity in a context is caused by a missing phoneme, the phoneme with the least FL score is the most "appropriate" candidate, because lower FL represents higher frequency in this context. To evaluate the the importance of tone information for ambiguity resolution, FL model is investigated on tonal contrasts in Mandarin Chinese in this paper.

If Mandarin loses its tonal information, words that share same phonemes but different tones would definitely cause confusion, but listeners would still come up with an answer on which the tone information is supplemented with the help of their language experience and linguistic knowledge. We intend to simulate this mechanism of spoken language processing using the FL theory. We choose sentences which will cause ambiguity without tone. To illustrate this point with a simple example, consider when people hear the word "gu shi" with flat tone and guess what the word is. Most of them would recognize the word as "故事gu4 shi5 (story)" rather than "股市gu3 shi4 (stock market)" or "古诗gu3 shi1 (ancient poem)", since "故事gu4 shi5 (story)" is more common than the others. And we predict the FL of the tone of "故事(story)" is less than the other two, which means tone is less important for recognizing this word. Therefore, listeners need the least effort to recognize it as "故事(story)".

The purpose of the current paper is to assess the importance of tone for speech perception from two perspectives: first, examining whether tone errors in L2 speech are correlated with listener-based judgement of comprehensibility. Second, investigating whether FL model is able to predict human's decision for ambiguous speech caused by absence of tone information.

II. Experiment 1

A. Materials

The speech materials used in present study were selected from BLCU-SAIT corpus [12], which is an interlanguage speech corpus of L2 learners of Chinese. 180 read speech of simple sentences from 20 Urdu-speaking learners (12 male and 8 female) from Pakistan were selected and annotated in the present study. All the speakers were students from Beijing Language and Culture University, and 9 sentences were selected from each speaker. All the initial, final and tone errors and their substitutions were annotated by a professional linguistic student.

B. 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 [13], 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. All the samples used in the familiarizing tasks were beyond the current study. 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, all 180 speech samples were rated. They 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 beyond the 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.

C. Results

Pearson's correlation was conducted on comprehensibility judgement with tone error ratio and segment error ratio. The results showed that tone error ratio ($r = -0.554, p < 0.001$) have a stronger correlation with comprehensibility judgement compared to segment error ratio ($r = -0.471, p < 0.001$).

TABLE I
Correlation between Comprehensibility Judgements and Two Features

Features	*Corr.*	*Sig.*
Tone error ratio	-0.554	***
Segment error ratio	-0.471	***

III. Experiment 2

A. Calculation of FLs

FL is used to evaluate the importance of phonetic contrasts [4]. In this study, FL is calculated using the model based on mutual information (MI) between text corpus and phoneme transcription [4]. The FL of a phonemic contrast is defined

as the relative change of mutual information of text and phonemic transcription after the language loses this contrast. For example, if we want to calculate FL of the phonemic contrast x and y, we merge all x and y in the corpus and replace them with the same symbol x. It makes the variation of the phonemic system decrease, causing more words sharing same pronunciations, in other words, more text sequences share the same phonemic transcriptions, and the MI changes. The relative change of MI is the loss of information caused by the merger of phoneme x and y. After the merger, the mutual information between the W and F will decrease. The reduction of mutual information reflects the loss of information caused by the merger of the phonemic contrast, and thus it is an optimal method to quantify information contribution of phonemic contrasts. The FL of phonemic contrast x and y is computed as:

$$\mathrm{FL}(x,y) = \frac{\mathrm{MI(W,F)} - \mathrm{MI(W,F}_{x,y})}{\mathrm{MI(W,F)}}, \tag{1}$$

$$\mathrm{MI(W,F)} = \lim_{n\to\infty} -\frac{1}{n}\log\sum_{i=1}^{m}\mathrm{P(W}_i^{'}) \tag{2}$$

In (1), $FL(x,y)$ represents the FL of phonemic contrast x and y; $MI(W,F)$ represents the mutual information of text and original phonemic transcription; MI $(W,F_{x,y})$ represents the MI of text and phonemic transcription after merging the x and y. In (2), $W_1', W_2', ..., W_m'$ are word sequences sharing the same pinyin transcription F. The probability of word sequence $P(W_i')$ is computed using bi-gram and tri-gram language model. The language model is trained using pinyin transcription of Chinese TV programs of 2007. The FL of tone in a specific sentence is calculated as:

$$\mathrm{FL}(T1,T2,T3,T4,T5) = \frac{\mathrm{MI(W,F)} - \mathrm{MI(W,F}_{T1-5})}{\mathrm{MI(W,F)}} \tag{3}$$

In (3), $FL(T1,T2,T3,T4,T5)$ represents the functional load of five tones in the sentence. F is phoneme sequence, and $MI(\mathrm{W,F_{T1-5}})$ represents the mutual information of the pinyin transcription and word sequence when five tones are merged in this language.

B. Materials

14 target sentences were selected (see in Table II), all of which cause ambiguity in the absence of tone information. In other words, after removing all the tone marks of the pinyin sequence of a selected sentence, it can represent more than one grammatically smooth sentences. The average length of the sentences is 5.6 syllables (SD = 0.9). 15 filler sentences were designed as syntactically and semantically similar to the target sentences. Besides, three random sentences were selected as the practice sentences in the experiment.

Each sentence was spoken by a female native Mandarin speaker who was graded first class B (an excellent speaker) in Putonghua Proficiency Test. In order to fully remove tone information. We decided to use natural speech instead of resynthesized speech, because the latter cannot fully remove tone information by flattening f0 contours only [14], [15]. Tone information not only exists in f0, but also amplitude envelope and duration [16].

To avoid any tone information in the stimuli, the speaker was given a list of the pinyin sequences of the sentences, all of which had been marked as tone 1 (level tone). The speaker was asked to read the list without knowing the lexical meaning of the sentences. After the recording, the duration of each syllable was normalized to 386ms (average score of all the syllables' durations) using Praat script. A sample of the normalized speech can be seen in Fig. 1.

The Pinyin sequence "zhe ge hua ti hen hao" in Fig. 1 can either be understood as "zhe4 ge5 hua4 ti2 hen3 hao3" meaning "this topic is very good" or "zhe4 ge5 hua2 ti1 hen3 hao3" meaning "this slide is very good".

TABLE II
PINYIN SEQUENCES SELECTED IN THE EXPERIMENT.

Sequence 1	wo xi huan zhe ge gu shi
Sequence 2	lao ma xi huan shui jiao
Sequence 3	ta xi huan shi zi
Sequence 4	wo yao bao shi
Sequence 5	wo xi huan ta de mei mao
Sequence 6	ta han yu shuo de hao
Sequence 7	dui bei jing zuo jie shao
Sequence 8	ta kan le kan yan se
Sequence 9	zhe shi wo de tu di
Sequence 10	wo xi huan hua xue
Sequence 11	zhe ge hua ti hen hao
Sequence 12	ta shi zhu li
Sequence 13	wo tao yan wei qi
Sequence 14	wo rang ta ban zou

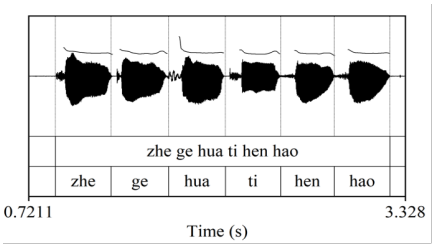

Fig. 1. A sample pinyin sequence without tone information.

C. Participants

Participants were 30 native Mandarin speakers (16 females, mean age = 23.2 yr, SD = 1.95) and were all undergraduate or graduate students. All of them reported normal hearing and no history of language disorder.

D. Procedures

The experiment was designed and carried out on the Finding Five platform [17], and the dictation task was played through headphones. First, participants were asked to listen to 3 practice sentences to familiarize the experiments. Then the formal trials were played randomly. Participants were asked to write down the sentences on a paper sheet, and one answer only for each trial. After writing down a sentence, participants

clicked "next" button to listen to the next sentence. When the experiment was over, the answer sheets were collected.

E. Data Analysis

The orthographic outputs of the participants were analyzed statistically. Answers with segmental errors and grammatical errors were excluded from the analysis. The frequency of each possible answer's occurrence and FL of tone contrast (T1-T5) in each answer sentence based on Bigram and Trigram were calculated. Besides, the frequency of each target word was also calculated. The correlations between FL, word frequency and the frequency of answer's occurrence were analyzed by R studio (Version 1.4.1106) [18].

F. Results

Part of the results of the dictation task can be seen in Table III. The FL value of each possible answer was computed and ranked. The answers with same tones are lumped together due to the limitation of the FL model. In most cases, listeners tend to recognize ambiguous sequence as the sentence with lower tone FL in most cases, eg. "wo xi huan zhe ge gu shi" has been recognized as "I like this story" and "I like this ancient poem", and the former answer occurred 29 times, the latter only occurred once. The FL of the tone in the former sentence is much lower than latter one, which means tone is less important in the former sentence. The most ambiguous sequence was recognized as five different sentences (Sequence 4). However, some answer didn't show a relation with FL, such as sequence 9. The answer with "徒弟(apprentice)" occurred more than "土地(land)", which is probably because of the syntactic structure of this sentence suits "徒弟(apprentice)" more than "土地(land)", and the FL is low probably because the word "徒弟(apprentice)" has few occurrences in the training corpus, i.e., TV program speech.

Pearson correlation coefficient was used to quantify the degree of the relation between FL and answers' frequency. A significant negative correlation between FL (based on bigram and trigram) and frequency of answer's occurrence was observed ($r = 0.46$, $p < 0.01$ for Bigram, $r = 0.50$, $p < 0.01$ for Trigram, see in Table IV), which means when listening to an ambiguous sentence without tone information, listeners tend to recognize it as the one with lower FL tones. Word frequency of the ambiguous part of the sentence showed similar correlation with the answers' frequency ($r = 0.48$, $p < 0.01$). Listeners tend to select the more frequent word as the answer. The FL model based on Trigram showed slightly stronger correlation than word frequency.

TABLE III
PART OF THE RESULTS OF THE DICTATION TASK (FAO: FREQUENCY OF ANSWER'S OCCURRENCE, R: RANK OF FL)

Seq.	Sentence	FAO	FL of tone (bi-gram)	FL of tone (tri-gram)	R
1	我喜欢这个古诗	1	0.212423	0.212675	1
	我喜欢这个故事	29	0.0354665	0.0354724	2
2	老妈喜欢水饺	1	0.319602	0.274547	1
	老妈喜欢睡觉	16	0.19628	0.143051	2
	老马喜欢睡觉	13	0.166315	0.129739	3
3	他喜欢柿子	2	0.101686	0.101999	1
	他喜欢识字	8	0.0803496	0.0808796	2
	他喜欢狮子	20	0.0333407	0.0341947	3
4	我要报失/报诗	3	0.208108	0.209306	1
	我要保湿	8	0.175284	0.169542	2
	我要保释/保试	5	0.100174	0.0956029	3
	我要暴食/报时	4	0.0881178	0.0896066	4
	我要宝石	10	0.0699603	0.0658836	5
5	我喜欢她的美貌	3	0.0655989	0.0687145	1
	我喜欢她的眉毛	27	0.0229971	0.0230068	2
6	对背景做介绍	2	0.119911	0.120393	1
	对北京做介绍	28	0.0013591	0.00141433	2
9	这是我的徒弟	23	0.114892	0.127433	1
	这是我的土地	7	0.0052898	0.00351844	2

TABLE IV
CORRELATION BETWEEN FAO, FL AND WORD FREQUENCY (SIGNIFICANCE LEVEL: $***\leq 0.001, **\leq 0.01, *\leq 0.05$)

	FL(Bigram)	***FL(Trigram)***	***Word Freq.***
FAO	-0.46**	-0.50**	0.48**
FL(Bigram)		0.98***	-0.26
FL(Trigram)			-0.28

IV. DISCUSSION AND CONCLUSION

The first result of this study is that tone error ratio have higher correlation with L2 speech comprehensibility than segment error ratio, indicating that tone is an important component for Chinese speech comprehension, and the importance arises when the segmental information is incomplete. The second result is that the resolution of ambiguity that caused by the absence of tone information has significant correlation with FL of missing tones. Listeners tend to recognize the pinyin sequences as the sentence with the tones with the least amount of FL. The less the FL is, the less the missing information is. Therefore, the result can be interpreted as listeners need the least effort to recognize the sentences with the lowest tone FL.

The main purpose of this study is: (1) to find out whether tone is important for speech comprehension, (2) to prove whether FL can simulate tone ambiguity processing of human. Based on the present study, following conclusions can be drawn: First, tone is important for comprehension in L2 speech. Second, tone is not a redundant information for Mandarin. Loss of tone information causes a certain degree of ambiguity. Third, FL can predict the human's decision to ambiguous speech to a certain extent. The tendency of the recognition of the speech without tone information is related to the FL of the tone in its context. To be specific, when a sentence loses its tone information, listeners tend to recognize the sequence as the sentence with the lowest tone FL value, where tone is not important. When a sentence with high FL tone loses its tone information, it would be very difficult for the listeners to recognize it correctly. The corpus used in present study is not diverse enough to predict human's decision, and a better language model can be used to compute the probability of word sequence in the future study.

REFERENCES

[1] T. Ng, M. Siu, and M. Ostendorf, "A quantitative assessment of the importance of tone in mandarin speech recognition," *IEEE Signal Processing Letters*, vol. 12, no. 12, pp. 867–870, 2005.

[2] D. Surendran and G.-A. Levow, "The functional load of tone in mandarin is as high as that of vowels," in *Speech Prosody 2004, International Conference*, 2004.

[3] Y. M. Oh, F. Pellegrino, C. Coupé, and E. Marsico, "Cross-language comparison of functional load for vowels, consonants, and tones." in *Interspeech*, 2013, pp. 3032–3036.

[4] J. Zhang, W. Li, Y. Hou, W. Cao, and Z. Xiong, "A study on functional loads of phonetic contrasts under context based on mutual information of chinese text and phonemes," in *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 194–198.

[5] 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.

[6] 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.

[7] W. D. Marslen-Wilson, "Functional parallelism in spoken word-recognition," *Cognition*, vol. 25, no. 1-2, pp. 71–102, 1987.

[8] M. Lucas, "Context effects in lexical access: A meta-analysis," *Memory & Cognition*, vol. 27, no. 3, pp. 385–398, 1999.

[9] M. G. Gaskell and W. D. Marslen-Wilson, "Lexical ambiguity resolution and spoken word recognition: Bridging the gap," *Journal of Memory and Language*, vol. 44, no. 3, pp. 325–349, 2001.

[10] P. A. Luce, "A computational analysis of uniqueness points in auditory word recognition," *Perception & Psychophysics*, vol. 39, no. 3, pp. 155–158, 1986.

[11] M. A. Pitt and A. G. Samuel, "Lexical and sublexical feedback in auditory word recognition," *Cognitive psychology*, vol. 29, no. 2, pp. 149–188, 1995.

[12] 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.

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

[14] J. Chen, H. Yang, X. Wu, and B. C. Moore, "The effect of f0 contour on the intelligibility of speech in the presence of interfering sounds for mandarin chinese," *The Journal of the Acoustical Society of America*, vol. 143, no. 2, pp. 864–877, 2018.

[15] F. Chen, "Mandarin tone identification with f0-flattening processed single-vowels."

[16] Q.-J. Fu and F.-G. Zeng, "Identification of temporal envelope cues in chinese tone recognition," *Asia Pacific Journal of Speech, Language and Hearing*, vol. 5, no. 1, pp. 45–57, 2000.

[17] FindingFive Team, *FindingFive: A web platform for creating, running, and managing your studies in one place*, FindingFive Corporation (nonprofit), NJ, USA, 2019. [Online]. Available: https://www.findingfive.com

[18] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA, 2020. [Online]. Available: http://www.rstudio.com/