REALIZATION OF LOW TONE SEQUENCES IN DISYLLABIC WORDS IN A LARGE MANDARIN SPEECH CORPUS

Yaru Wu[1,2,3], Yiya Chen[4,5], Lori Lamel[3]
[1]CRISCO/EA4255, Université de Caen Normandie, 14000 Caen, France
[3]Laboratoire de Phonétique et Phonologie (UMR7018, CNRS-Sorbonne Nouvelle), France
[3]LISN, Univ. Paris-Saclay, 91405 Orsay cedex, France
[4]Leiden University Centre for Linguistics (LUCL)
[5]Leiden Institute for Brain and Cognition (LIBC)
yaru.wu@unicaen.fr, yiya.chen@hum.leidenuniv.nl, lori.lamel@lisn.fr

ABSTRACT

This study investigated the acoustic realization of disyllabic Low tone (T3T3) words in Mandarin. A large corpus of 850 hours of journalistic speech was used. The segmentation of the continuous Mandarin speech was carried out using the LISN (former LIMSI) speech transcription system in forced alignment mode. The full corpus was aligned twice using two different pronunciation dictionaries, one with and one without systematic tone variants. We analyzed the tonal representations of the underlying T3T3 sequences that the LISN system produced with the two alignment strategies. We further investigated the effect of different factors, namely prosodic position, word frequency, tonal contexts, and parts of speech, on conditioning the surface realization of the underlying T3T3 sequences. The LISN outputs suggest that the first T3 of disyllabic T3T3 words is not always realized as T2 in continuous speech and a range of factors conditions its acoustic realizations.

Keywords: Mandarin, tone realization, large corpora, variation factors

1. INTRODUCTION

Over the last decades, there has been great progress in both the construction and availability of large speech corpora (e.g., LDC) and automatic speech recognition techniques that are capable of learning powerful speech representations in acoustic audio signals (e.g., [1]). In this paper, we capitalize on this progress and investigate the acoustic realizations of Low tone sequences (T3T3) in disyllabic words in a large Mandarin speech corpus.

It is well-known that Low tone sequences in Mandarin can change through sandhi, where a Low tone is pronounced as a rising pitch before another Low tone [2] (known as Low tone sandhi or T3 sandhi). Many factors can influence the application of T3 sandhi (e.g., [3]). While there is debate over how to characterize the factors that condition T3 sandhi and whether the rising pitch of the Low tone sandhi variant is completely neutralized with the rising lexical tone (T2), it has been generally accepted, based on both impressionistic observations and lab speech, that T3 undergoes the sandhi process in disyllabic words. Taking a corpus approach, Yuan & Chen [4] investigated the f0 realization of T3 in disyllabic words (T3T3) using the LDC corpus [5] (i.e. LDC2005S15 & LDC98S73) and compared them to disyllabic words with T2T3 sequences. They found that the T3 sandhi variant differs from T2 in terms of the magnitude and time span of the f0 rise.

In this study, we used the LISN (former LIMSI) speech transcription system to test the categorization of different T3 realizations in disyllabic words with T3T3 sequences. The output of the LISN system enables self-supervised learning which has been shown to be effective in situations with limited labeled data. It can be used to discover structural patterns in a significantly larger amount of unlabeled data by exploiting multi-dimensional contextual information in the corpus. The goals of the study was to (1) gain a better understanding of the realization of Low tone in a sandhi context from naturally occurring corpus speech, (2) learn how it varies according to different factors, including prosodic position, word frequency, tonal contexts, and parts of speech, and (3) determine the impact of these variation factors on the acoustic realizations of disyllabic T3T3 sequences in standard Mandarin.

2. METHOD

2.1. Corpus and alignments

A Mandarin journalistic corpus, distributed by LDC [6, 7], was used. It contains about 850 hours of continuous Mandarin speech (e.g., [8, 9]), with 9M word-tokens. The segmentation at the

word and phone levels was carried out using the speech transcription system from LISN in forced alignment mode [10, 11, 12]. The pronunciation lexicon provides phone level representations for each word (tones included). The best matching pronunciations among potential realizations were selected automatically during the forced alignment. Pauses, hesitations, and breath were also detected automatically by the system. Neutral tone was not included in the pronunciation lexicon. The minimal phone segment duration is 30 ms given the acoustic modeling technical constraints of a 3-state model and a 10 ms (frame) step [13, 14].

The full corpus was aligned twice using two different pronunciation dictionaries. The first set of alignments used a base pronunciation lexicon, which includes almost no variants (alignments V0). Words with pronunciation variants (less than 4% of all word-tokens) in the first set of alignments were not concerned by our analyses. We then used an expanded pronunciation dictionary for the second set of alignments (V1), in which tone variants were introduced for the first and last syllable of each word. This means words' first and last syllables can be aligned with any of the four tones. In this study, we decided to focus on T3T3 disyllabic words. Therefore, both syllables of the words can be aligned with any of the four tones the system sees fit. Using this method, we investigated the differences between V0 (reference tone) and V1 (surface realizations of tones) based on the LISN alignments. For instance, 打理(/da3li3/, "manage") can only be aligned with T3T3 in V0; in V1, however, it can be aligned as any of the 16 possibilities ($T_{1-4}T_{1-4}$). These two sets of alignments, i.e., without (V0) and with tone variants (V1), made it possible to quantify the realization of tones with respect to different variation factors.

2.2. Investigated factors

We analyzed the realization of the first T3 in T3T3 disyllabic words according to the following factors: prosodic position, left tonal context, relative word frequency, and part of speech. To better understand the realization of the first T3 in T3T3 disyllabic words, we decided to exclude T3T3 words immediately following another T3 syllable.

The two prosodic positions concerning the first T3 of disyllabic words are phrase-initial and word-initial. Phrase-initial tone is defined as the tone of the first syllable of a disyllabic word immediately following a pause (including silence, hesitation and breath) $\geq$ 100 ms. This threshold was empirically determined given that pause durations of 50, 100 and 200 ms gave similar results with respect to prosodic position. Word-initial tone refers to tones in the first syllable of disyllabic words that do not appear at a phrase-initial position. The left tonal context concerns the tone of the syllable that immediately precedes the T3T3 word. We used the relative word frequency values provided in SUBTLEX-CH for our analyses on the word frequency factor [15]. The SUBTLEX-CH provides word and character frequency measures based on a corpus of film subtitles, covering 33.5 million words or 46.8 million characters. The relative frequency ranges from 0 to 100. The part of speech (POS) of each word-token of the corpus was automatically annotated using Stanza [16]. This supplementary annotation allows us to examine the link between tone realization and POS. Words without matching POS were excluded from our analyses. Given that few word tokens were concerned for "conjunction" (CCONJ) and for "particle" (PART), we did not include these two categories in our analyses. The grammatical words, i.e., words that are adposition (ADP), auxiliary (AUX), determiner (DET), numeral (NUM), pronoun (PRON), were categorized into the same group named GRAM. Details on the concerned level of each variation factor can be found in Tab. 1.

Factors	Levels	Examples
Left tonal context	Tone1#_	T1#打(/da3/)理
	Tone2#_	T2#打(/da3/)理
	Tone3#_	T3#打(/da3/)理
	Tone4#_	T4#打(/da3/)理
Prosodic position	Phrase-initial	##打(/da3/)理(*manage*)
	Word-initial	#打(/da3/)理
Relative frequency	Scaled to the range of 0 to 100	
Parts of speech	ADJ (adjective), ADP (adposition), ADV (adverb), AUX (auxiliary), CCONJ (coordinating conjunction), DET (determiner), NOUN (noun), NUM(numeral), PART (particle), PRON (pronoun), PROPN (proper noun), VERB (verb)	

Table 1: Factors investigated for the first T3 of T3T3 disyllabic words. # stands for word boundary; ## for phrase boundary.

2.3. Statistical analyses and random forest model

We carried out the statistical analyses using MCMCglmm package in R [17], allowing us to fit Generalized Linear Mixed Models using Markov chain Monte Carlo techniques. The results of the statistical analyses are reported with a 95% highest posterior density (HPD) interval for each coefficient.

We also carried out a random forest classifier on the prediction of the realization of the first T3 in T3T3 words. The factors mentioned in Table 1 were included as independent variables: relative

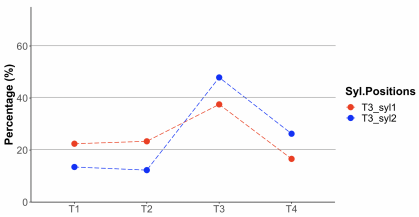

Figure 1: Realization of the first and the second T3 of T3T3 disyllabic words. The four types of realizations are shown on the x-axis. The first T3 is shown in red and second in blue.

word frequency, left tonal context, part of speech, and prosodic position. In addition to this list, we also included the part of speech of the preceding and the following word as independent variables. The contribution of each factor to the model was then evaluated. The data were randomly selected and divided into two parts: (1) 70% of the data was used for training and (2) the remaining 30% of the data was reserved as test data. The 10-fold cross-validation method was applied to extract the best combination of hyperparameters.

3. RESULTS

Fig. 1 presents the percentage of the first (in red) and the second (in blue) T3 of T3T3 disyllabic words realized as T1, T2, T3 or T4. For both T3s, the percentages of the four realizations add up to 100%. The figures show that the first T3 of T3T3 words is realized less as T3 than the second T3 of T3T3 words.

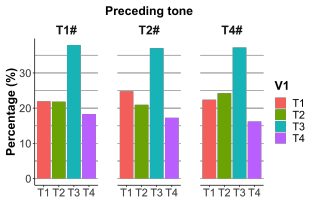

Figure 2: Realization of the first T3 in T3T3 words according to the preceding tone.

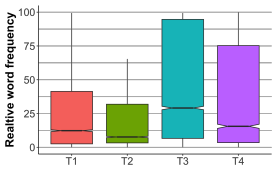

Figure 3: Relative word frequency range for each realization of the first T3 of T3T3 words.

Fig. 2 shows the realization of the first T3 in T3T3 words as a function of the left tonal context (post-lexical tonal context) of the word in question.

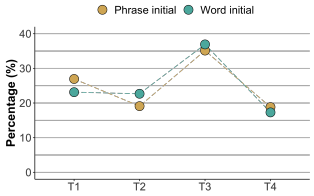

Figure 4: Realization of the first T3 in T3T3 words according to the prosodic position of the syllable carrying the tone.

The first T3 of T3T3 words is more likely to be realised as T1 when it is preceded by T2 than when it is preceded by T4 ($T1 \times PrecededByT2$ = 0.19372, 95%HPD [0.06056, 0.28655], p< 0.01). It is less likely to be realised as T2 when it is preceded by T1 or T2 than when it is preceded by T4 ($T2 \times PrecededByT1$ = -0.10165, 95%HPD [-0.21333, -0.01836], p< 0.05; $T2 \times PrecededByT2$ = -0.13801, 95%HPD [-0.21717, -0.05662], p< 0.01). Interestingly, when the T3 is preceded by T2, it is more likely to be realized as T1 than as T2, probably due to the delayed high f0 peak realization from the preceding T2 [18]: T2(3-**5**)#T1(**5**-5)T3.

Fig. 3 shows the relationship between relative word frequency and the realization of the first T3 in T3T3 words. The relative word frequency ranges from 0 to 100 (see more details in [15]). Results based on the multinominal logistic confirmed that the first T3 of T3T3 words is less likely to be realized as T1, T2 or T4 in high frequency words ($T1 \times frequency$ = -0.00364, 95% HPD $[-0.00400, -0.00326]$, p < 0.01; $T2 \times frequency$ = -0.00198, 95% HPD $[-0.00232, -0.00161]$, p < 0.01; $T4 \times frequency$ = -0.00202, 95% HPD $[-0.00250, -0.00151]$, p < 0.01).

Fig. 4 presents the realization of the first T3 in T3T3 words according to the prosodic position of the word (i.e. IP initial *vs* word-initial positions). Similar patterns are found for both positions as far as the realization of the first T3 is concerned. The T3 in question is observed to be realized as T1 slightly more often than as T2 at the IP-initial position.

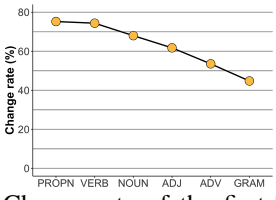

Figure 5: Change rate of the first T3 in T3T3 words according to different parts of speech.

Fig. 5 shows the change rate of the first T3 in T3T3 disyllabic words as a function of the parts

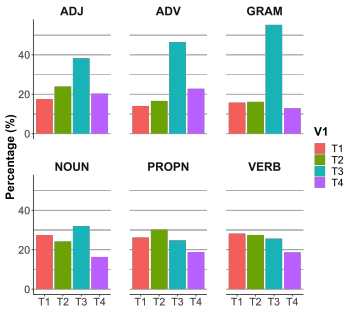

Figure 6: Realization of the first T3 in T3T3 words for the concerned parts of speech.

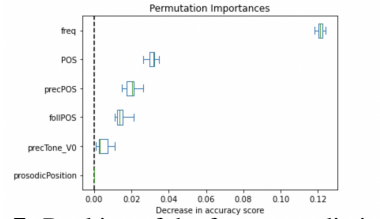

Figure 7: Ranking of the factors predicting of the realization of the first T3 in T3T3 words.

Comparisons	95% HPD	p value
T1 × ADJ = 0.60152	[0.38861, 0.80995]	p < 0.01
T1 × ADV = 0.51933	[0.39077, 0.65743]	p < 0.01
T1 × NOUN = 1.33442	[1.19102, 1.43624]	p < 0.01
T1 × PROPN = 1.07729	[0.66248, 1.45905]	p < 0.01
T1 × VERB = 1.73975	[1.57341, 1.83614]	p < 0.01
T2 × ADJ = 0.79490	[0.56019, 1.04017]	p < 0.01
T2 × ADV = 0.48989	[0.38881, 0.59259]	p < 0.01
T2 × NOUN = 1.16182	[1.07119, 1.27842]	p < 0.01
T2 × PROPN = 1.54294	[1.23952, 1.79054]	p < 0.01
T2 × VERB = 1.66008	[1.54706, 1.74707]	p < 0.01
T4 × ADJ = 1.12799	[0.94772, 1.29447]	p < 0.01
T4 × ADV = 1.07407	[0.94964, 1.19133]	p < 0.01
T4 × NOUN = 1.01376	[0.83692, 1.16253]	p < 0.01
T4 × PROPN = 1.52619	[1.26715, 1.80068]	p < 0.01
T4 × VERB = 1.44156	[1.31836, 1.55031]	p < 0.01

Table 2: Results of the multinominal logistic model concerning parts of speech.

of speech of the words. The first T3 of the T3T3 words tends to be realized as other tones the least in grammatical words (GRAM) and the most in proper nouns (PROPN). Fig. 6 illustrates the realization of the first T3 as a function of the parts of speech. The results indicate that the realization patterns of T3 in T3T3 words differ among different parts of speech categories. The first T3 of a T3T3 word is more likely to be realized as T1 when the word in question is an adjective (ADJ), an adverb (ADV), a noun (NOUN) a proper noun (PROPN), or a verb (VERB) than when it is a grammatical word/GRAM ($p < 0.01$ for all related comparisons; see details in Table 2). Similar trends are found for the first T3 of T3T3 words realized as T2 or T4 (cf. Tab. 2).

The ranking of the factors is presented in Fig. 7 according to the random forest model. Relative word frequency is observed to be the factor that contributed the most to the prediction of the realization of the first T3 in T3T3 words, followed by part of speech of the word containing the T3 in question (POS), part of speech of the preceding word (precPOS), part of speech of the (follPOS) and the lexical tone of the preceding syllable. Prosodic position, however, does not seem to contribute much to the prediction concerning the realization of the first T3 in T3T3 words, which is likely due to the limited levels (i.e. phrase initial vs. word initial) we were able to code and to the relative low occurrence in speech of phrase-initial T3s in T3T3 words.

4. DISCUSSION AND CONCLUSIONS

This study aimed to gain a better understanding of the acoustic realization of Low tone in a tone sandhi context (T3T3) in continuous Standard Chinese speech. To do so, the speech data was segmented twice in forced alignment mode using the LISN speech transcription system: once without and once with systematic tone variants. The output of the first set of alignments can be interpreted as the reference pronunciation of words. The output of the second set of alignments is related to the surface realization of words. This offered us an opportunity to compare directly between the underlying T3T3 sequences and their acoustic realizations, classified as one of the four lexical tones, according to the LISN transcription system. The system learns about the categorization of the four tones without supervision. In this paper, we focused on the tonal representations of the first T3 in the underlying T3T3 sequences of disyllabic words. The results showed that the realization of the first T3 varied as a function of a set of factors which included prosodic position, word frequency, tonal contexts, and part of speech. The results of the random forest algorithm suggest that the impact of these factors on T3 realization is unequal, as indexed by the ranking of the factors based on their contribution to the prediction of the T3 realizations. Among the factors investigated, the relative word frequency tends to contribute the most to predicting the realization of the first T3 in T3T3 words. It would be interesting to explore including neutral tone as a category, in addition to the four lexical tones. The application of T3 sandhi in bi-syllabic words is known to be non-optional and consistent in the literature. The varied realizations of T3 in the LISN output beg the question of how acoustic modeling systems such as that of LISN learn and classify the tones. For future research, it is important to compare native listeners' classification to LISN's output to gain further insights.

5. REFERENCES

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[2] M. Y. Chen, *Tone sandhi: Patterns across Chinese dialects*. Cambridge University Press, 2000, vol. 92.

[3] C. Shih, "Mandarin third tone sandhi and prosodic structure," *Linguistic Models*, vol. 20, pp. 81–124, 1997.

[4] J. Yuan and Y. Chen, "3rd tone sandhi in standard chinese: A corpus approach," *Journal of Chinese Linguistics*, vol. 42, no. 1, pp. 218–237, 2014.

[5] P. Fung, S. Huang, and D. Graff, "Hkust mandarin telephone speech/transcripts, part 1 ldc2005s15/t32," 2006.

[6] S. M. Strassel, C. Cieri, A. Cole, D. DiPersio, M. Liberman, X. Ma, M. Maamouri, and K. Maeda, "Integrated linguistic resources for language exploitation technologies." in *LREC*, 2006, pp. 185–190.

[7] A. Morris, B. Antonishek, X. Li, and S. Strassel, *HAVIC MED Progress Test–Videos, Metadata and Annotation*. Linguistic Data Consortium, University of Pennsylvania, 2019.

[8] K. Walker, C. Caruso, K. Maeda, D. DiPersio, and S. Strassel, "Gale phase 3 chinese broadcast news speech ldc2015s13," *Web Download. Philadelphia: Linguistic Data Consortium*, 2015.

[9] M. Glenn, H. Lee, S. Strassel, and K. Maeda, "Gale phase 3 chinese broadcast news transcripts ldc2015t25," *Web Download. Philadelphia: Linguistic Data Consortium*, 2015.

[10] J.-L. Gauvain, L. Lamel, and G. Adda, "The limsi broadcast news transcription system," *Speech communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[11] M. Adda-Decker and L. Lamel, "The use of lexica in automatic speech recognition," in *Lexicon Development for Speech and Language Processing*. Springer, 2000, pp. 235–266.

[12] S. Huang, J. Liu, X. Wu, L. Wu, Y. Yan, and Z. Qin, "Mandarin broadcast news speech (hub4-ne)," *Linguistic Data Consortium*, 1998.

[13] L. Lamel, J.-L. Gauvain, V. B. Le, I. Oparin, and S. Meng, "Improved models for mandarin speech-to-text transcription," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4660–4663.

[14] L. Chen, L. Lamel, G. Adda, and J.-L. Gauvain, "Broadcast news transcription in mandarin," in *Sixth International Conference on Spoken Language Processing*, 2000.

[15] Q. Cai and M. Brysbaert, "Subtlex-ch: Chinese word and character frequencies based on film subtitles," *PloS one*, vol. 5, no. 6, p. e10729, 2010.

[16] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," *arXiv preprint arXiv:2003.07082*, 2020.

[17] J. Hadfield, "Mcmcglmm: Markov chain monte carlo methods for generalised linear mixed models," *Tutorial for MCMCglmm package in R*, vol. 125, 2010.

[18] Y. Xu, "Fundamental frequency peak delay in mandarin," *Phonetica*, vol. 58, no. 1-2, pp. 26–52, 2001.