Mandarin Lexical Tones: A Corpus-Based Study of Word Length, Syllable Position and Prosodic Position on Duration

Yaru Wu, Martine Adda-Decker, Lori Lamel

Mandarin lexical tones:
a corpus-based study of word length, syllable position and prosodic position on duration

Yaru Wu[1,3]*, Martine Adda-Decker*[1,2]*, Lori Lamel*[2]

[1]Laboratoire de Phonétique et Phonologie (UMR7018, CNRS-Sorbonne Nouvelle), France
[2]LIMSI, CNRS, Universite Paris-Saclay, Rue John von Neumann, F-91405 Orsay Cedex, France
[3]Modèles, Dynamiques, Corpus (MoDyCo), UMR 7114, CNRS

yaru.wu@sorbonne-nouvelle.fr, madda@limsi.fr, lamel@limsi.fr

Abstract

The present study aims to increase our knowledge of Mandarin lexical tones in fluent speech, more specifically their occurrence frequency distributions and their duration patterns. First, the occurrence frequency of each lexical tone was computed in a large speech corpus ($\sim$ 220 hours). Then the duration of each lexical tone, as well as the impact of word length, syllable position and the prosodic position were investigated. Overall, results show that Tone 3 tends to have the longest duration among all lexical tones. Nonetheless, the factors word length, syllable position and prosodic position are found to impact tone duration. Monosyllabic words exhibit tone durations closer to those of word-final syllables (especially for disyllabic words) than to other syllable positions. Moreover, tone duration tends to be the longest at word's right boundary in Mandarin, regardless of word length. An effect of prosodic position is also found: the duration of Mandarin lexical tones tends to increase with higher prosodic level. Tone durations are the longest in phrase-final position, followed by word-final position and word-medial position, regardless of the tone nature.

Index Terms: Mandarin, tone duration, word length, prosodic categories, large corpora, continuous speech.

1. Introduction

Mandarin Chinese (Standard Chinese) is a tonal language with four lexical tones: high (Tone 1), rising (Tone 2), low-dipping (Tone 3) and falling (Tone 4). Word meaning can depend on tone, for example, "ma" (/ma/) could mean "mother" in Tone 1, "linen" in Tone 2, "horse" in Tone 3 and "scold" in Tone 4. Whereas most words in Mandarin carry a lexical tone, there is also a "neutral tone" in Mandarin (Tone 0) which occurs in some words in unstressed syllables.

According to an early publication on isolated spoken words [1], Tone 3 has the longest and Tone 4 the shortest duration among the four lexical tones. More recent studies report some diverging results on tone duration in continuous speech. In line with [1], Tseng [2] found Tone 3 to be the longest among the four lexical tones in continuous Mandarin. Chang [3] however, reported Tone 2 to be the longest and Tone 3 the shortest in sentence-final position. Yang et al. [4] found that the four lexical tones tend to have a similar duration in continuous speech.

This study attempts to clarify the question of Mandarin lexical tone duration. Unlike many previous studies on tone duration, which either focused on monosyllabic words in an isolated setting or investigated tone duration in read sentences, this study makes use of a large corpus of continuous speech of spoken Mandarin. Moreover, this paper aims to provide a better understanding of tone duration and its variation as a function of different factors, namely, word length (here, number of syllables in word), syllable position and prosodic position.

One question we ask is whether or not the most frequent tone correlates with the shortest duration, and if the least frequent tone generally has a longer duration in continuous speech. According to [5] word length can have an impact on tone duration. The position of the syllable in the word was found to influence vowel duration in French [6]. We investigate whether tone durations tend to vary according to their syllable positions in words in Mandarin. Prosodic structure classifies segments into prosodic levels with hierarchically layered categories [7, 8, 9]. It is found to correlate with different acoustical [10] or articulatory changes [11, 12]. With regard to segment duration, the influence of prosodic position is found in English [13]. In this study, we look at the influence of prosodic position on the duration of Mandarin lexical tones. We expect to find longer tone duration at higher prosodic levels, as often observed in other languages (see [14] for English and [15] for French).

2. Method

2.1. Corpus and alignment

A large collection of manually transcribed journalistic Mandarin speech ($\sim$ 220 hours of LDC data [16, 17]) containing journalistic Mandarin was used ([18, 19], to cite a few). The speech data was automatically segmented at the word and phone levels using the LIMSI speech transcription system [20, 21, 22] in forced alignment mode. A Mandarin pronunciation lexicon provides phone-level representations for each word including tones. The orthographic transcriptions guide the forced alignment which selects the best matching pronunciation among the alternatives provided in the lexicon. The pronunciation lexicon includes a limited number of pronunciation variants with an average of 1.05 pronunciations per word. The variants may include changes in the phone (e.g. /b,p/) or in the lexical tone (e.g. all 4 tones are allowed for the character "啊", *particle showing affirmation*). For words with pronunciations variants differing in tone (less than 4% of the entries), the lexical tones are those selected during the forced alignment process. Neutral tone was not integrated in the lexicon and is hence not used by the forced alignment system. Pauses (including hesitations and breath) are also automatically detected. The minimum phone segment duration is 30 ms due to the acoustic modelling technical constraints arising from a 3-state acoustic model and a 10 ms (frame) step [23, 24].

2.2. Phonetic analyses

In the following, tone duration is considered as the duration of

the segment starting from the vowel onset through the end of the syllable offset [25]. Prior to looking into tone duration and its variation, we will take advantage of the corpus to quantify the relative frequency of n-syllabic words in speech, Mandarin being known as a language with many monosyllabic and short words. The occurrence frequency distribution and the duration of Mandarin lexical tones are first analysed globally and then according to word length (i.e. number of syllables in word) and syllable position.

We are interested in the influence of different factors on the four Mandarin lexical tones. In fluent speech, the median segment duration is around 70 ms and a very long duration may indicate an alignment error. Acoustic segments of length above 200 ms typically correspond to pauses or hesitations. In the following analyses, only tone segments with a duration no more than 400 ms are considered[1].

The tone segments were normalized to z-scores for the duration analyses. Analyses of the influence of prosodic position on tone duration concern only polysyllabic words since "word medial" and "word final" are not meaningful for monosyllabic words.

Tone segments in three prosodic positions were selected: word-medial, word-final and phrase-final positions, as shown in the lower part of Table 1. Phrase-final tone segments concern the last syllable of polysyllabic words immediately preceding a pause ($\geq$ 100 ms) [2]. Word-final tone segments refer to tones in the last syllable of words (phrase-final tone segments excluded). Word-medial tones concern tones that are not in word-boundary syllables.

2.3. Statistical analyses

Linear mixed models (LMM) using the *lme4* packages [26] in R [27] were carried out for the statistical analyses of tone duration as a function of different variation factors: tone nature, word length, position of the syllable in the word, and prosodic position. Two LMMs were used in this study. The first LMM was used to test the effect of tone nature, word length and syllable position on tone duration. The fixed effects considered were: tone nature (Tone 1, Tone 2, Tone 3 and Tone 4, reference: Tone 1), word length (1 to 4 syllables in word, reference: 1-syllable words), and syllable position (S_n to S_{n-3}, reference: syllable S_{n-3},). As random effects, intercepts were included for subject and word-form. The second model was used to investigate the effect of prosodic position. In this case, the data set is much smaller than the first one, since only specific prosodic positions were selected. Here prosodic position (word-medial, word-final and phrase-final; reference: word-medial) was included as the fixed effect. Tone nature and word length were included as control variables. Intercepts for subject and word-form were considered as random effects. Post-hoc tests based on each model were performed to obtain information on each level of the fixed effects. Table 1 provides details of the investigated variation factors.

3. Analyses and results

In this section, we first present the occurrence frequencies of words (grouped by word length) and of lexical tones in our 220 hour Mandarin speech corpus. Then, we present the effect of

Table 1: *Examined factors for tone duration. # stands for word boundary; ## for phrase boundary. Syllables bearing the relevant tones are circled in the rows of "prosodic position".*

Factors	Levels	Examples
Tone nature	Tone 1	妈 (/ma1/, *mother*)
	Tone 2	麻 (/ma2/, *linen*)
	Tone 3	马 (/ma3/, *horse*)
	Tone 4	骂 (/ma4/, *scold*)
Word length	1 syllable	九 (*nine*)
	2 syllables	国家 (*country*)
	3 syllables	消费者 (*consumer*)
	4 syllables	画蛇添足 (*gild the lily*)
Syllable position in word	Last syllable (S_n)	画蛇添足 消费者 国家 九
	Last syl.−1 (S_{n-1})	画蛇添足 消费者 国家
	Last syl.−2 (S_{n-2})	画蛇添足 消费者
	Last syl.−3 (S_{n-3})	画蛇添足
Prosodic position	Word-medial	消费者 (*consumer*)
	Word-final	消费者#
	Phrase-final	消费者##

tone nature, word length, syllable position and prosodic position on tone duration.

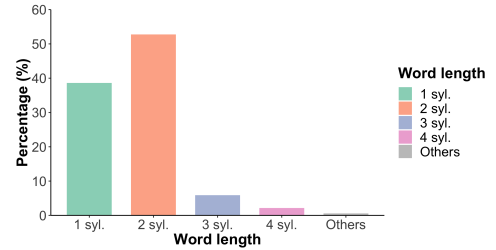### 3.1. Word and lexical tone occurrence frequencies

Figure 1: *Distribution of word length in Mandarin (220h corpus).*

Figure 1 shows the proportion of n-syllable words as found in our data. Words presenting more than four syllables are pooled together in the category "Others". These counts clearly confirm that the most frequently used Mandarin words are short, with a just under 40% being monosyllabic (38.6%, 753794 word-tokens) and slightly more than half disyllabic (52.8%, 1029489 word-tokens). That is to say, less than 10% of the word-tokens in corpus have more than 2 syllables with about 6% and 2% respectively for 3-syllable (5.9%, 115082 tokens) and 4-syllable words (2.1%, 41411 tokens). Words longer than 4 syllables represent less than 1% of all occurrences. Long words such as "印度尼西亚" (*Indonesia*) may exist in Mandarin Chinese. However, they are observed to be rather marginal.

We now turn to occurrence frequencies of the different lexical tones in Mandarin speech. Figure 2 illustrates the proportion of each lexical tone in spoken Mandarin. Interestingly, it turns out that the falling Tone 4 is the most frequently used tone (34.9%, 1172916 tone segments), followed by the stable high Tone 1 (24.8%, 833257 tone segments) and rising

---

[1]Fewer than 1% of the tone segments are excluded with this filter.

[2]This value was empirically determined after tests with pause durations of 50ms, 100ms and 200ms gave similar results on the influence of prosodic position.

Tone 2 (23.9%, 803218 tone segments). The more complex low-dipping Tone 3 is the least frequent tone appearing in our continuous speech corpus (16.4%, 551751 tone segments).
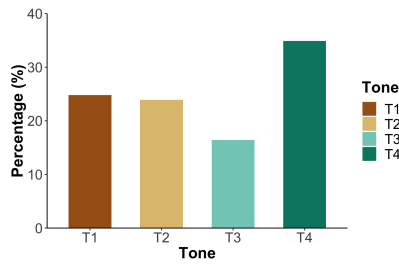

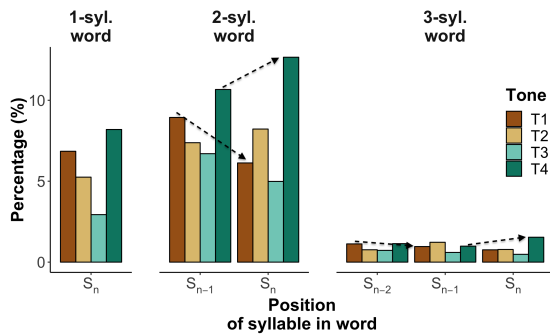Figure 2: *Overall relative occurrence frequencies of Mandarin lexical tones.*

Figure 3: *Occurrence frequency of Mandarin lexical tones as a function of word length and syllable position. The 3 graphs display the tone distribution for each syllable in words with n=1,2,and 3 syllables.*

Figure 3 displays the proportion of Mandarin lexical tones as a function of word length and syllable position. For the sake of space, 4-syllable words were not included in this figure since there are only very few such word-tokens in the corpus (2.1%). Unsurprisingly, Tone 4 is the most frequently observed tone, and this is true for almost all syllable positions and words with different numbers of syllables. Tone 3 is always the least frequently observed tone regardless of its syllabic position in the word. With respect to polysyllabic words, it is interesting to note, that Tone 1 (high tone) occurs more frequently in word-initial position than subsequent positions and that Tone 4 (falling tone) is always most frequent in word-final position. This global word level pattern of more high tones at the start and more falling tones at the end of words may evoke a local "declination line" pattern which is often observed in speech productions [28].

3.2. Duration of Mandarin lexical tones

The durations of Mandarin lexical tones are analysed in this section. We first present average durations for the four lexical tones before going into more detail.

Table 2 shows the means and standard deviations of Mandarin lexical tone durations (in milliseconds). Tone 1, Tone 2 and Tone 4 have very similar mean tone durations. The low-dipping Tone 3 has the longest mean duration. This suggests that the least frequent tone has longer segment durations than

Table 2: *Mean and standard deviation (sd) of the duration (in milliseconds) of Mandarin lexical tones.*

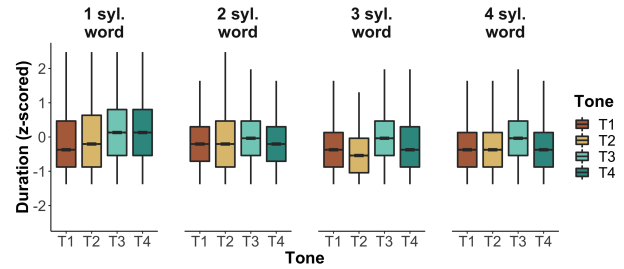| Tone	Mean	sd
Tone 1	110.22	57.57
Tone 2	109.71	64.07
Tone 3	120.06	56.02
Tone 4	111.60	59.18

Figure 4: *Duration of Mandarin lexical tones (normalized to z-score), grouped by word length.*

the other tones (Tones 1, 2 and 4), but the most frequent Tone 4 is not necessarily shorter than all other lexical tones. These results highlight that frequency alone is certainly not a sufficient indicator of segment duration and that more sophisticated analyses may help provide additional insight.

Figure 4 shows the normalized durations of Mandarin lexical tones as a function of the word length in terms of number of syllables. Tone 3 is seen to have the longest duration for polysyllabic words. For monosyllabic words, Tone 3 and Tone 4 have the longest duration, followed by Tone 2 and then Tone 1. For trisyllabic words, Tone 2 has the shortest duration and Tones 1 and 4 have similar durations. Interestingly, Tones 1, 2 and 4 appear to have similar durations in disyllabic and tetrasyllabic words. LMM results confirm the influence of tone nature on duration, while also considering the effect of word length and syllable position on tone duration. The duration for Tone 3 is significantly longer than that of Tone 1 [$\beta = 0.126830$; t = 55.013; SE = 0.002305]. The normalized durations of Tone 2 [$\beta = -0.069645$; t = -35.994; SE = 0.001935] and Tone 4 [$\beta = -0.050923$; t = -26.597 ; SE = 0.001915] are shown to decrease, compared to that observed in Tone 1 (considering also word length and syllable position). Post-hoc results based on the LMM model show that all pairwise-comparisons of the four tones are significant (p < 0.001).

3.3. Effect of word length and syllable position on tone duration

Here, we look at how tone duration depends on word length and how it varies according to syllable position.

Figure 5 shows a comparative view of the normalized tone duration for words of different lengths and at different syllable positions within a word. In this analysis, all tones are pooled. First, it can be seen that the monosyllabic words exhibit tone durations closer to those of word-final syllables (especially for disyllabic words) than to other positions. Overall, the tone duration of the word-final syllables appear to be the longest among all positions. The LMM results suggest that word length and syllable position affect tone duration (see details in 2.3). Normalized tone duration is shorter in disyllabic
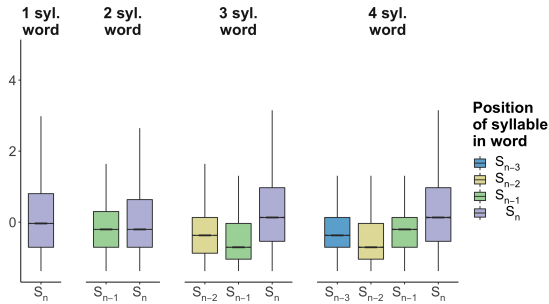
Figure 5: *Normalized tone segment durations (all tones pooled) as a function of word length and syllable position.*

words [β = -0.071847; t = -7.253; SE = 0.009906], trisyllabic words [β = -0.165999; t = -14.103; SE = 0.011770] and tetrasyllabic words [β = -0.135576; t = -10.475; SE = 0.012943], compared to that observed in monosyllabic words (considering also tone nature and syllable position). Post-hoc tests with the concerned LMM model show that words of different length are significantly different from each other as far as tone duration is concerned (trisyllabic vs. tetrasyllabic: p < 0.05; all other pairwise comparisons: p < 0.001). The effect of syllable position is found. Tone duration is significantly shorter in S_{n-2} than in S_{n-3} [β = -0.123477; t = -22.633; SE = 0.005456]. However, tones have significantly longer duration in S_{n-1} [β = 0.075804; t = 14.365; SE = 0.005277] and S_n [β = 0.386141; t = 73.065; SE = 0.005285] than in S_{n-3}. Post-hoc tests show that all pairwise comparisons among the four positions of syllable in word are significant (p < 0.001).

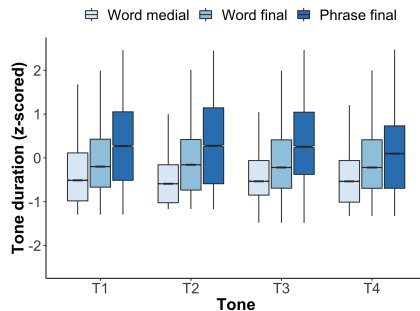### 3.4. Effect of prosodic position on tone duration

Figure 6: *Tone duration as a function of prosodic position: word-medial, word-final vs. phrase-final tone segments.*

The influence of prosodic position on tone duration is shown in Figure 6. The phrase-final level features the longest durations among the three prosodic levels, and this is true for all four tones. Word-medial positions are seen to consistently be the shortest. The figure clearly illustrates that the duration of tone segments increases with higher prosodic levels. LMM results confirm that the duration of tones in word-final position is significantly longer than that observed in word-medial position [β = 0.778868; t = 231.643; SE = 0.003362]. The duration in phrase-final position is also significantly longer than in word-final position [β = 1.140279; t = 241.688; SE = 0.004718]. There is indeed an effect of prosodic position on tone duration

(p < 0.001), with longer tone segments as the level increases (word-medial vs. word-final : p < 0.001; word-medial vs. word-final: p < 0.001; word-final vs. phrase-final: p < 0.001).

4. Discussion and Conclusions

This study focuses on tone duration in Mandarin, defined here as the duration of the segment starting from the vowel onset through the end of the syllable offset. A large journalistic speech corpus of ∼ 220 hours of fluent Mandarin speech was exploited to explore tone durations as a function of tone nature, word length, syllable position and prosodic position. A first step quantified length distribution of words and the occurrence frequency of lexical tones in the speech collection. More than 90% of the word-tokens in spoken Mandarin are either mono- (38.6%) or di-syllabic (52.8%). With regard to lexical tone, the falling Tone 4 is the most frequent (34.9%), followed by the stable high Tone 1 (24.8%) and rising Tone 2 (23.9%). The low-dipping Tone 3 is least frequent tone in our corpus (16.4%). Concerning the proportion of lexical tones as a function of word length and syllable position, we found that Tone 1 (high tone) is more frequent in word-initial position than in subsequent positions and that Tone 4 (falling tone) is always the most frequent tone in word-final position.

Tone duration varies as a function of tone nature for words of different lengths (i.e. number of syllables in word). Tone 3 is found to be longer than the other three lexical tones in disyllabic, trisyllabic and tetrasyllabic words. In monosyllabic words, Tone 3 and Tone 4 are the longest lexical tones. Interestingly, Tone 1, Tone 2 and Tone 4 tend to have similar duration in disyllabic and tetrasyllabic words. Although tone occurrence frequency appears to impact duration for Tone 3, we did not find any convincing evidence of a simple link between the occurrence frequency and tone duration. Indeed, the most frequent tone, Tone 4, was not necessarily shorter than other lexical tones. Our results on Mandarin tone duration are different from those reported in [4] where the authors found similar durations for all four lexical tones in read speech (< 1 hour) and for conversational speech (∼ 1 - 2 hours). The differences may be due to the use of different corpora (size, types of speech) in the two studies and, perhaps most importantly, that tone durations were grouped by both tone nature and word length in our analysis.

Tone duration co-varies with syllable position and word length (1 to 4 syllables). Overall, and independently of word length, tone duration of the word final syllable appears to be the longest. Duration is thus the highest at the word's right boundary. Interestingly, this finding is in line with vowel duration patterns as a function of syllable position observed in French [6]. Similar results of word-final lengthening have been reported for Mandarin [5].

Finally, the influence of three prosodic levels was investigated in this study: word-medial, word-final and phrase-final positions. The results clearly confirmed increasing segment durations with higher prosodic levels, in line with other findings of the influence of prosodic position (see [10, 11, 13]).

5. Acknowledgements

6. References

[1] A. T. Ho, "The acoustic variation of mandarin tones," *Phonetica*, vol. 33, no. 5, pp. 353–367, 1976.

[2] C.-Y. Tseng, *An acoustic phonetic study on tones in Mandarin Chinese*. Brown University, 1981.

[3] C.-Y. Chang, "Dialect differences in the production and perception of mandarin chinese tones," Ph.D. dissertation, The Ohio State University, 2010.

[4] J. Yang, Y. Zhang, A. Li, and L. Xu, "On the duration of mandarin tones." in *INTERSPEECH*, 2017, pp. 1407–1411.

[5] C. Lai, Y. Sui, and J. Yuan, "A corpus study of the prosody of polysyllabicwords in mandarin chinese," in *Speech Prosody 2010-Fifth International Conference*, 2010.

[6] M. Adda-Decker, C. Gendrot, N. Snoeren, and N. Nguyen, "Apport du traitement automatique à l'étude des voyelles," 2013.

[7] M. Nespor and I. Vogel, "Prosodic phonology, foris, dordrecht," *NesporProsodic Phonology1986*, 1986.

[8] D. R. Ladd, "Intonational phrasing: the case for recursive prosodic structure," *Phonology*, vol. 3, pp. 311–340, 1986.

[9] T. Cho, "Prosodic boundary strengthening in the phonetics–prosody interface," *Language and Linguistics Compass*, vol. 10, no. 3, pp. 120–141, 2016.

[10] J. Pierrehumbert and D. Talkin, "Lenition of/h/and glottal stop," *Papers in laboratory phonology II: Gesture, segment, prosody*, pp. 90–117, 1992.

[11] C. Fougeron and P. A. Keating, "Demarcating prosodic groups with articulation," *The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3384–3384, 1995.

[12] C. Fougeron and P. A. Keating, "Articulatory strengthening in prosodic domain-initial position," *UCLA Working Papers in Phonetics*, vol. 97, pp. 61–87, 1996.

[13] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1707–1717, 1992.

[14] T.-J. Yoon, J. Cole, and M. Hasegawa-Johnson, "On the edge: Acoustic cues to layered prosodic domains," in *Proceedings of ICPhS*, vol. 16. Citeseer, 2007, pp. 1264–1267.

[15] C. Fougeron and S.-A. Jun, "Rate effects on french intonation: Prosodic organization and phonetic realization," *Journal of Phonetics*, vol. 26, no. 1, pp. 45–69, 1998.

[16] 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.

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

[18] 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.

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

[20] 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.

[21] 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.

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

[23] 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.

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

[25] J. M. Howie and J. M. Howie, *Acoustical studies of Mandarin vowels and tones*. Cambridge University Press, 1976, vol. 18.

[26] D. Bates, M. Maechler, B. Bolker, S. Walker, R. Christensen, H. Singmann, B. Dai, and G. Grothendieck, "Package 'lme4'," *Linear mixed-effects models using S4 classes. R package version*, vol. 1, no. 6, 2011.

[27] R Development Core Team, *R: A La nguage and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

[28] D. R. Ladd, "Declination.: a review and some hypotheses," *Phonology*, vol. 1, pp. 53–74, 1984.