



Mandarin Tone Sandhi Realization: Evidence from Large Speech Corpora

Zuoyu Tian¹, Xiao Dong¹, Feier Gao¹, Haining Wang¹, Chien-Jer Charles Lin¹

¹Indiana University Bloomington

{zuoytian, dong1, gaof, hw56, chiclin}@iu.edu

Abstract

There has been a long-standing debate on the acoustical difference between second tone and sandhi third tone in Mandarin Chinese. Except [1], most of the studies focus on the tonal realization in carefully controlled speech, and the relationship between demographic features and tonal realization has not been addressed by existing studies using large-scale speech data. In this paper, we investigate the tonal realization in three large speech corpora of Mandarin Chinese, which contain around 2,300 speakers from different regions of China. The relation between demographic factors (e.g., gender, age, and place of origin) and tonal realization is addressed. The results show that even though the sandhi third tone is close to the second tone in terms of acoustic features, it still displays significant differences in large data. We also report some demographic influences on tonal realization, including age, gender, and places of origin.

Index Terms: tone sandhi, Mandarin, large speech dataset, demographics, phonological alternation

1. Introduction

Tone 3 sandhi is an obligatory and productive phonological alternation rule in Mandarin Chinese, where an initial low-dipping tone (tone 3 or T3) syllable changes into a rising tone (tone 2 or T2) when it is followed by another low-dipping tone (tone 3) syllable. This process, referred to as the Mandarin Tone 3 Sandhi, therefore involves the substitution of one lexical tone by another; namely a tone 3 is replaced by a tone 2 when it is followed by another tone 3 syllable [2]. Previous studies have consistently reported a 100% application of the sandhi rule to both real words and novel words by adult Mandarin speakers [3, 4]. However, it is still under debate whether the tone 2 of a sandhi tone 3 syllable is phonologically identical to a regular tone 2 syllable or whether it is phonetically adjusted from tone 3. How the demographic factors of the speakers (e.g., place of origin, gender, etc.) affect the realization of T3 sandhi also remains to be explored.

By adopting a corpus method, the current study seeks to answer two questions: 1) Are sandhi syllables realized in a categorical way or in a gradient way by native speakers? 2) How is the realization of T3 sandhi in disyllabic words affected by demographic factors (i.e., place of origin, age, and gender) of the speakers?

2. Related Studies

One way to test the categorical vs. gradient view of the application of tone 3 sandhi is to investigate if the realization of a sandhi T3 is acoustically distinct from a T2 syllable that is underlyingly T2. Previous perception studies have consistently reported that Mandarin speakers could not distinguish sandhi tone 3 from lexical tone 2, which provides support to the categorical view (see [5, 6, 7]), but findings from production studies are more complex. To the best of our knowledge, almost all

earlier experimental studies on the production of sandhi words, in which participants' production of T2+T3 and T3+T3 minimal pairs were compared, have found that the sandhi tone 3 produced by Beijing Mandarin speakers differs from underlying tone 2 (see [8, 9, 10, 11]). On the other hand, most studies involving speakers from non-Beijing Mandarin regions as participants found no difference in the F0 data between sandhi T3 and T2, suggesting that tone 3 sandhi may involve categorical mapping from tone 3 to tone 2 (see [12, 6, 2]). Two exceptions are [4] and [7]. Both studies found differences between sandhi T3 and underlying T2. Although [4] argues that these differences might be due to factors like coarticulation and speaker differences, these findings imply that speakers of a variety of Mandarin other than Beijing Mandarin may also process tone sandhi in a gradient fashion. [1] is one of a few studies that used the corpus method to investigate the acoustic characteristics of sandhi T3 relative to underlying T2. They took two corpora as databases, one being made of telephone conversations while the other featuring broadcast news speech. The results indicate that sandhi T3 differs from underlying tone 2 in both the magnitude and the time span of the F0 rising. They also found an effect of word frequency: sandhi T3 in highly frequent words exhibited a smaller F0 rising and therefore were more different from underlying T2, which might suggest that different processes are involved in producing high- vs. low-frequency words involving tone sandhi in Mandarin. However, the possible contributions of speakers' demographic characteristics to the realization of T3 sandhi were not addressed in this study.

In a word, earlier studies on the realization of tone 3 sandhi have showed inconsistent results, and regions where participants are from tend to add variations to this issue. Previous studies mainly recruited participants from a single dialectal region (e.g., Beijing or Taiwan), so the production details of tone 3 sandhi within a broader geographical context remain to be explored. Besides, although it is widely recognized that speech production may differ across other demographic factors (e.g., gender and age) [13, 14], few studies have investigated how these factors affect the realization of T3 sandhi. It is also worth noting that results from previous studies mainly came from reading speech in experimental settings; more evidence regarding how tone 3 sandhi is realized in natural speech is needed. Meanwhile, we notice that many corpora are developed for automatic speech recognition (ASR). Such datasets typically involve large amounts of natural speech and provide high-quality transcriptions. However, only a few studies used these corpora to investigate phonetics issues of Mandarin Chinese.

Therefore, in the present study, we expand the scope of previous tone sandhi studies by conducting acoustic analysis on the data obtained from three large-scale spoken corpora, which contain speech data produced by speakers across mainland China. Details are given in the next section.

3. Data

We used speech data from three large ASR corpora of Mandarin Chinese: Free ST Chinese Mandarin Corpus (STCMDS) [15], aidatang_200zh [16], and MAGICDATA Mandarin Chinese Read Speech Corpus [17]. All three corpora contain three types of demographic information of speakers: age, gender, and place of origin. They also come with high-quality transcripts, which are critical for word segmentation and forced alignment.

3.1. Speech Corpora

In this study, we focus on the realization of two tonal sequences: /T3T3/ and /T2T3/. Following a previous study [1], we chose disyllabic words containing /T3T2/ or /T2T2/ as control groups. Table 1 shows the statistics of our targeted data. It shows that MAGICDATA is much larger than the other two speech corpora, while STCMDS and aidatang_zh200 are similar in size. The speakers' gender ratio is adequately balanced across the three corpora, and the average ages of speakers are all around 22. We see a clear divergence in terms of the speaker's places of origin, but most of the locations (operationalized as provinces) have at least 10 speakers in each corpus. Our processed data is released at <https://github.com/zytian9/Mandarin-tone-sandhi-statistics>.

3.2. Data Processing

Since the word segmentation of each corpus is not unified, we first used LTP [18] to redo the word segmentation for all transcripts. As LTP achieves the state-of-the-art accuracy in Chinese word segmentation, we expected it to reduce the issue of segmentation error. Then, we used g2pM [19] to convert Chinese characters to pinyin. These two steps allowed us to locate the target words.

We used Charsiu [20] to carry out the phonetic segmentation. It is a transformer-based aligner, which aligns the vowel part of each syllable (represented by a character in writing) with high accuracy¹. After the forced alignment, Parcelmouth [21], a Python library for the Praat, was used for extracting the F0 values of 20 evenly split data points in the tone-bearing syllable. We set pitch range as 75-500Hz during extraction.

3.3. Acoustic Measurement

We used LogRange of the F0 rising introduced by [1] to measure the rising magnitude of the tone-bearing syllable. The LogRange of the F0 rising was calculated by first taking the ratio of the F0 of the tone bearing syllable's offset and minimum datum, and then scaling with logarithm with base e. We noticed that the distribution of LogRange of the F0 rising was positively skewed with more than a few zeros. Therefore, we adopted Mood's median test to test the null hypothesis in the following sections.

4. Acoustic Features of Tone 2 and Tone 3

Figure 1 shows the LogRange of the F0 rising in the tone bearing syllable of the first character in the sequence of T3T3 and T2T3. Even though the degrees of the F0 rising vary across

¹Here we directly used the alignment files of all three corpora released by the authors of Charsiu. To make sure that the phonetic segmentation quality is good enough for our study, we sampled 10 files from each corpus to check the segmentation accuracy. We calculated the overlap of rhymes between automatic alignment and human judgments. It showed that 82.6% predicted intervals had more than 80% overlap with the human ones.

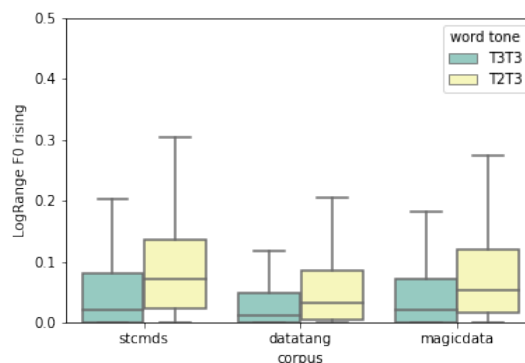


Figure 1: LogRange of the F0 rising in tone-bearing syllable of the first character of T2T3 and T3T3

three corpora, in all three corpora, tone 2 shows a clear positive F0 rising, as 0 is not in the 50 percentile zone in all three box plots. Compared to T2, the F0 rising of sandhi T3 is smaller in scale. Furthermore, unlike the control group T3T2 (Figure 2) where sandhi does not happen, a clear F0 rising is observed in the sandhi tone 3 in T3T3. In all three corpora, 0 is within the 95 percentile zone for T3T2 sequences. The Mood's median test in Table 2 also suggests that T2T3 significantly differs from T3T3.

In the next step, We examined the effect of word frequency (see Table 2). We chose 30 most frequent words which occur more than 100 times in each corpus as high-frequency words and another 30 words from the bottom of the wordlists which occur 10-13 times as low-frequency words. We found that the mean and median of the F0 rising in high-frequency words were lower than that of low-frequency words, and the difference was salient. Across different frequency settings in all three corpora, T3T3 and T2T3 were statistically different.

When we compared with the results of LogRange of the F0 rising from Yuan and Chen's study, we found that more zero F0 rising values occurred in our dataset, and the F0 rising values in our data were also smaller than those of the naturally-produced telephone conversations. In their paper, broadcast news showed much larger F0 rising than telephone conversations. It indicates that informal speech seems to show greater variability and smaller rising scale in terms of F0 rising. In our paper, all of three corpora we used contain large amounts of informal speech. Meanwhile, the corpora are much larger than that of telephone conversation², the varieties of word usage and speakers' diversified background may lead to the smaller F0 rising magnitude.

5. Demographic Factors in Tone Realization

In the previous section, we discussed the acoustic features of tone 2 and sandhi tone 3 across three corpora, and the results indicate that the three corpora, irrespective of the sample size, show a similar trend where a sandhi T3 syllables have a smaller F0 rising scale than an underlying T2 syllable. We further combined these three corpora to investigate the impact of demo-

²There are 3938 T3T3 words in the HKUST Mandarin Telephone corpus, while our smallest corpus (aidatang_zh) has 14995 T3T3 words

Corpus	T3T3	T2T3	T3T2	T2T2	Speakers	Male	Female	Av. Age	Top 5 provinces
STCMDS	15996	12979	10249	15142	844	317	527	22.0	Tianjin, Shanxi, Henan, Hebei, Gansu
aidatang	14995	12942	13172	14343	420	214	206	22.9	Fujian, Guangdong, Beijing, Jiangsu, Hunan
MAGICDATA	46222	38898	31840	43316	1013	476	537	22.8	Anhui, Hebei, Henan, Shandong, Jiangsu

Table 1: Statistics of $T3T3_{word}$, $T2T3_{word}$, $T3T2_{word}$, and $T2T2_{word}$ across three corpora

Corpus	Group	Mean	Median	p value
STCMDS	T3T3 vs T2T3	0.062, 0.1	0.021, 0.07	0***
	$T3T3_{high}$ vs $T2T3_{high}$	0.062, 0.087	0.017, 0.055	< 0.01***
	$T3T3_{low}$ vs $T2T3_{low}$	0.076, 0.129	0.041, 0.102	< 0.01***
aidatang	T3T3 vs T2T3	0.041, 0.067	0.012, 0.033	< 0.01***
	$T3T3_{high}$ vs $T2T3_{high}$	0.036, 0.051	0.01, 0.024	< 0.01***
	$T3T3_{low}$ vs $T2T3_{low}$	0.053, 0.103	0.02, 0.079	< 0.01***
MAGICDATA	T3T3 vs T2T3	0.061, 0.09	0.02, 0.053	0***
	$T3T3_{high}$ vs $T2T3_{high}$	0.053, 0.073	0.017, 0.041	0***
	$T3T3_{low}$ vs $T2T3_{low}$	0.063, 0.131	0.037, 0.101	< 0.01***

Table 2: Results of Mood's median test in $T3T3_{word}$ and $T2T3_{word}$ across corpora and different frequency settings

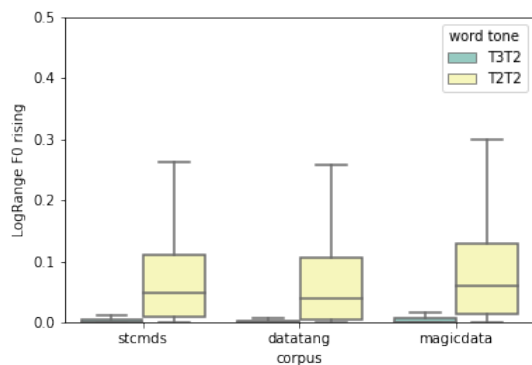


Figure 2: LogRange of the F0 rising in tone-bearing syllable of the first character of $T3T2_{word}$ and $T2T2_{word}$

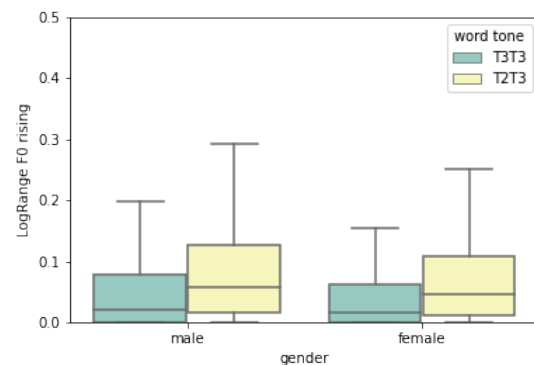


Figure 3: LogRange of the F0 rising in $T3T3_{word}$ and $T2T3_{word}$ across gender

graphic factors on the tonal realization of sandhi tone 3 in this section in order to look into the effect of demographic factors using a larger sample size.

5.1. Gender

We first looked into the role of gender in tone sandhi realization. Figure 3 shows that male and female speakers seem to produce tone 2 and sandhi tone 3 in a similar way where the sandhi T3 syllables are smaller in pitch rising than the T2 syllables. But when we looked into the statistical test results (see Table 3), we found that the tonal realization of sandhi tone 3 differed significantly between male speakers and female speakers. We saw a larger median value in male speakers (0.021 vs 0.016), although the mean values of F0 rising in these two groups were very close (0.056 vs 0.058). Interestingly, we found a very similar trend of F0 rising in T2 syllables as well (mean: 0.0876 vs 0.0876; median: 0.058 vs 0.047), so this gender difference was not restricted to the process of tone sandhi.

5.2. Age

To investigate how the age of speakers will affect tonal realization, we split speakers into four groups by age: <20, 20–29,

30–39, and ≥ 40 . By doing descriptive statistics (see Figure 4) and Mood's median test (see Table 3), we found that there were no differences between speakers under 20 years old and those from 20 to 29 years old in terms of the tonal realization of sandhi tone 3. However, sandhi tone 3 was produced differently in speakers who were between 30 and 39 years old. A larger F0 rising value was found in this group of speakers while we did not observe similar trend of higher F0 rising in speakers who are older than 40.

5.3. Provinces

We investigated the impact of place of origin in terms of provinces. For this analysis, we chose to focus on seven provinces to represent greater dialectal diversity: Tianjin, Henan, Sichuan, Heilongjiang, Guangdong, and Fujian. Several major dialects/languages are included in these seven provinces, such as Beijing Mandarin, Northeastern Mandarin, Zhongyuan Mandarin, Southwestern Mandarin, Min, and Cantonese.

Results presented in Figure 5 shows that speaker's dialectal background poses an influence on the realization of T3 sandhi. Specifically, the speakers from the south (e.g., Fujian and Guangdong) tended to produce smaller LogRange F0 rising on the initial syllable of the disyllabic sandhi word, as opposed

Types of Factor	Group	Mean	Median	p value
Gender	Male vs Female	0.056, 0.058	0.021, 0.016	< 0.01***
Age	<20 vs 20-29	0.053, 0.056	0.017, 0.017	0.69
	20-29 vs 30-39	0.056, 0.078	0.017, 0.03	< 0.01***
	30-39 vs ≥ 40	0.078, 0.06	0.03, 0.015	< 0.01***

Table 3: Results of Mood's median test in $T3T3_{word}$ in gender and age

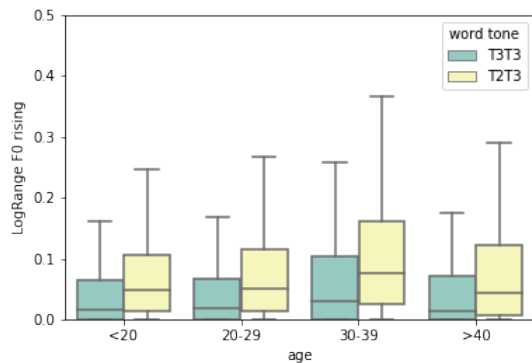


Figure 4: LogRange of the F0 rising in $T3T3_{word}$ and $T2T3_{word}$ across different age groups

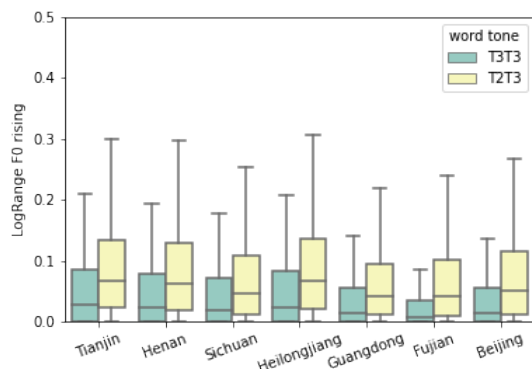


Figure 5: LogRange of the F0 rising in $T3T3_{word}$ and $T2T3_{word}$ based on speakers' province

to the speakers from the north. Specifically, speakers from Fujian (a southern province of China) produced the smallest F0 rising on the sandhi T3 (mean: 0.039, median: 0.006), and speakers from Heilongjiang (a northeastern province) showed the greatest F0 rising (mean: 0.063, median: 0.022). It is also noticeable that Beijing (a northern city) speakers produced smaller F0 rising compared with other northern speakers. However, T2 also presents a similar trend, it may suggest that regional acoustical differences result in varying tonal realization.

6. Discussion and Conclusion

The current study investigates the realization of tone 3 sandhi in three large speech corpora of Mandarin Chinese and analyzes how word frequency and different demographic factors affect its realization. Firstly, our results indicate that in all three corpora, across different frequency ranges and demographic fac-

tors, T3T3 and T2T3 are statistically different. This observation provides further support for the gradient view as was also reported by Yuan and Chen (2014). In other words, Mandarin speakers do not substitute a T3 with a T2 in the sandhi context. Instead, they adjust their articulation of the sandhi tone 3 to make it sound like a tone 2. Note that in our study we found that non-Beijing Mandarin speakers (e.g., people from Fujian) also applied the sandhi rule in a gradient way and were more susceptible to incomplete production than speakers in the northern regions. This result contrasts with those of Peng (2000), which recruited speakers from Taiwan and found categorical productions of sandhi tone 3 by monolingual Mandarin speakers and bilingual speakers of Mandarin and Taiwan Southern Min [7]. We suspect this difference may be due to whether tone 3 sandhi is explicitly described as a categorical mapping between Tone 3 and Tone 2 in school education.

Regarding the effect of word frequency, a similar pattern was found as [1]: sandhi T3 in highly frequent words exhibited a lower mean and median of the LogRange F0 rising and were more distinct from T2 than those in low-frequency words. One possible explanation is that different processes are involved in the encoding of high-frequency versus low-frequency sandhi words. An alternative possibility for this difference may be that syllable durations tend to be shorter in high frequency words in production, and thus leading to the reduction of pitch rising. This latter explanation is consistent with the effect of phonological reduction in high-frequency words as they are easier to retrieve from the lexicon for encoding.

Our study also demonstrates that gender and regional background play a role in the realization of sandhi words. We found that male speakers present a bit larger LogRange F0 rising than female speakers in producing both sandhi tone 3 and tone 2. It indicates that even though some female speakers produce more extreme F0 rising and lead to the two very close mean values, male speakers are more likely to produce a greater F0 rising than female speakers in general. The finding that regional background affects the realization of tone 3 sandhi is interesting. This is the first study that shows one's geographical region (and therefore dialectal experience) leads to differential production of tone 3 sandhi where southern speakers of Mandarin (e.g., from Fujian where Min is the dominant language) tend to reduce the pitch-rising scale in the sandhi tone 3 syllables. Such a reduced production can be taken as an incomplete realization of tone 3 sandhi due to processing difficulty by bilinguals or non-native speakers of Mandarin. While the real cause of the regional differences may be more complex and need to be further explored, our study provides a more nuanced view regarding the realization of tone 3 sandhi. It also suggests that future studies in tonal studies should take these factors into consideration.

7. Acknowledgements

We thank Jian Zhu at the University of Michigan and Chien-Han Hsiao at Indiana University Bloomington for their helpful advice.

8. References

- [1] 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.
- [2] J. Myers and J. Tsay, "Investigating the phonetics of Mandarin tone sandhi," *Taiwan Journal of Linguistics*, vol. 1, no. 1, pp. 29–68, 2003.
- [3] J. Zhang and Y. Lai, "Testing the role of phonetic knowledge in Mandarin tone sandhi," *Phonology*, vol. 27, no. 1, pp. 153–201, 2010.
- [4] C. Zhang and G. Peng, "Productivity of Mandarin third tone sandhi: a wug test," *Eastward flows the great river: Festschrift in honor of Prof. William S.Y. Wang on his 80th birthday*, pp. 256–282, 2013.
- [5] W. S. Wang and K.-P. Li, "Tone 3 in Pekinese," *Journal of speech and hearing research*, vol. 10, no. 3, pp. 629–636, 1967.
- [6] Y.-C. Chang and Y.-C. Su, "La modification tonale du 3ème ton du Mandarin parlé à Taiwan," *Cahiers de Linguistique-Asie Orientale*, vol. 23, no. 1, pp. 39–59, 1994.
- [7] S. Peng, "Lexical versus 'phonological' representations of Mandarin sandhi tones," *Papers in laboratory phonology V: Acquisition and the lexicon*, vol. 5, p. 152, 2000.
- [8] E. Zee, *A spectrographic investigation of Mandarin tone sandhi*, 1980.
- [9] P. Kratochvil, "Phonetic tone sandhi in Beijing dialect stage speech," *Cahiers de linguistique-Asie orientale*, vol. 13, no. 2, pp. 135–174, 1984.
- [10] X.-n. S. Shen, *the prosody of Mandarin Chinese*. University of California Press, 1990, vol. 118.
- [11] Y. Xu, "Contextual tonal variation in Mandarin Chinese," Ph.D. dissertation, University of Connecticut, 1993.
- [12] R. Brotzman, "Progress report on mandarin tone study," 1964.
- [13] L. Liang and X. Meng, "A sociophonetic study on tones of chongqing Mandarin in gender and age difference." in *ICPhS*, 2011, pp. 1230–1233.
- [14] Y. Wang and J. Zhang, "A preliminary study on the gender differences of Mandarin focal accent," in *2021 International Conference on Asian Language Processing (IALP)*. IEEE, 2021, pp. 67–71.
- [15] Surfing Technology Beijing, "STCMD520170001_1, free st Chinese Mandarin corpus," 2017. [Online]. Available: <https://openslr.org/38/>
- [16] Beijing DataTang Technology, "aidatang_200zh," 2018. [Online]. Available: <https://openslr.org/62/>
- [17] MAGIC DATA Technology, "MAGICDATA Mandarin Chinese read speech corpus," 2019. [Online]. Available: <https://openslr.org/68/>
- [18] W. Che, Y. Feng, L. Qin, and T. Liu, "N-ltp: A open-source neural Chinese language technology platform with pretrained models," *arXiv preprint arXiv:2009.11616*, 2020.
- [19] K. Park and S. Lee, "A neural grapheme-to-phoneme conversion package for Mandarin Chinese based on a new open benchmark dataset," *Proc. Interspeech 2020*.
- [20] J. Zhu, C. Zhang, and D. Jurgens, "Phone-to-audio alignment without text: A semi-supervised approach," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
- [21] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.