9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Audio-visual Analyses of Differences Between Natural and Teaching Styles for Mandarin Tone Production

Yueqiao Han, Martijn Goudbeek, Maria Mos and Marc Swerts

Tilburg Center for Cognition and Communication, Tilburg University, the Netherlands
y.han@uvt.nl

Abstract

In order to examine the acoustic and visual information that is used by speakers to facilitate the perception of Mandarin tones, 4 Chinese native speakers were video-taped and asked to produce Mandarin tones/words in natural and teaching speaking styles as the experimental materials for a tone identification task. Acoustic and visual analyses of the produced materials were conducted to evaluate the characteristics of the tones/words produced in both speaking styles. Acoustic results showed that as compared to natural style, speakers in teaching style produce Mandarin tones in a more "exaggerated" way, represented mainly by prolonging the pronunciation of the tones. Visual analyses revealed that speakers in teaching mode signaled more visual information/facial motions than in the natural style. Furthermore, the types of facial motions (horizontal and vertical movement) displayed are associated with tone variations.

Index Terms: Mandarin tones, speaking style, audio-visual analyses, acoustic characteristics, facial motions

1. Introduction

It has been widely acknowledged that speech is an audio-visual event: whenever visual information is available, observers use it to decode what they hear [1, 2, 3, 4]. The classic McGurk effect demonstrates how perceivers are influenced by what they see: observers perceive an auditory [ba] paired with a visual [ga] as 'da' or 'tha' [5]. Therefore, access to visual information about the source of speech can have clear effects on speech perception, as it alters the perception of speech. Visual information is provided by movements of the lips, face, head and neck. In order to be understood correctly, speakers can be assumed to strive to provide optimal acoustic and/or visual information to meet the demands of the target audience or the communicative situation [6]. This paper aims to investigate the acoustic and visual characteristics of the speech/tones produced by Mandarin Chinese speakers in two different communicative situations: natural style (the way native speakers speak in their daily interactions), and teaching style (the hyperarticulated way in which teachers/native speakers address non-native speakers in a teaching context).

Many scholars have found that in order to make themselves more intelligible to listeners, speakers usually articulate in a more "exaggerated" manner: they maximize phonetic contrast, speak more slowly, more loudly and more clearly. These modifications in speaking style have been discussed extensively as "clear speech" [7, 8, 9]. Clear speech modification aims at providing more salient acoustic cues in the speech signal for the listeners to enhance their access to and comprehension of the message. In the field of second language learning, "clear speech" is commonly associated with "teacher talk" that teachers use when addressing second language (L2) learners in the classroom [10, 11], anticipating learners' needs for assistance in their attempts at comprehension.

An interesting question regards the extent to which a teaching style also affects the production of tones. Tones in Mandarin Chinese serve to distinguish meanings at the lexical level; tone can be viewed as a phonemic distinction that is attached to the syllable at a suprasegmental level [12]. The consensus is that fundamental frequency (F0) is the most dominant phonetic cue for Mandarin Chinese tones [13, 14, 15]. Based on F0 patterns (both height and contour) and the direction of pitch, tone 1 has been described as high level (5-5), tone 2 as mid-rising (or mid-high-rising; 3-5), tone 3 as low-dipping (also low-falling-rising or mid-falling-rising; 2-1-4), and tone 4 as high-falling (5-1) [16]. In order to make the acoustic difference among tones more salient for the non-native listeners, native speakers commonly apply the teaching style when they produce Mandarin Chinese. Therefore, acoustic analysis should reveal longer duration, and an exaggerated pitch range. The tone fidelity, however, should not be impacted by the speaking style, since pitch is closely related to the lexical meaning of the word [17].

In addition, many studies have shown that there is visual speech information for lexical tones [2, 18, 19, 20, 21, 22, 23, 24]. However, some scholars ([2] for instance) argued that for tone perception the extra facial information may actually distract the listeners, since when acoustic sources are available and reliable, listeners are reluctant to use the visual information. Regardless of how effective the visual cues on the perception for the listeners, it is a fact that speakers' faces signal information when they articulate the speech. Especially when the speakers tend to transfer the tone knowledge (the pitch contour for instance), facial cues (along with gestures) are the common resource they resort to alongside the acoustic information.

Physiological studies [25] have suggested that visual information may result from certain restrictions with respect to the coordination of the laryngeal and articulatory systems [19]. Put another way, the way the tones are acoustically realized is also visually signaled because our mouths and faces need to move in a certain way to produce a given tone. For instance, vowel duration tends to be the longest for tone 3 and shortest for tone 4; amplitude tends to be lowest for tone 3 and highest for tone 4 [26, 27]. These acoustic differences may have visual correlates, for instance in the amplitude and the length of the visible articulations. Therefore, when speakers employ a "teaching style", specifically geared to non-native listeners, or a more natural speaking style, geared towards fellow native speakers, the amount of visual information displayed on speakers' face can be different. Since speakers spend more energy to exaggerate the tone information in the teaching style,

10.21437/SpeechProsody.2018-148

we expect more facial movements to be generated in the accompanying articulatory process.

The central purpose of this paper is to evaluate the acoustic and visual characteristics of the produced tones/words by native Mandarin Chinese speakers in teaching and natural styles. Considering there are variations among speakers and tones and some speakers differ in the clarity of their articulation and visual information [28], and the fact that some tones also demand more explicit visual cues than other tones, we also looked into variation between speakers and between tones in this study.

2. Method

Acoustic and visual analyses were conducted to assess the differences between two speaking styles (teaching and natural) for the produced Mandarin tones. Relevant acoustic parameters of the materials were measured (duration, average pitch and pitch range). Total facial motions were measured for video recordings of the tones in teaching and natural conditions, including the motions on the horizontal and vertical axis. Variations among speakers and tones were also examined.

2.1. Subjects

Four adult native Mandarin-Chinese speakers (two male and two female) were recruited from Tilburg University: two females (ages 29 and 35) who had been in the Netherlands for about 1 year, and two males (ages 29 and 28) who had been in the Netherlands for about 3 years. All speakers were born and raised in China and had come to the Netherlands for their graduate studies.

2.2. Material

2.2.1. Stimulus construction

A word list with 10 monosyllables (e.g., ma, ying …) was constructed (selection based on [14, 23] see Appendix). Each of these syllables was chosen such that the four tones would generate four different meanings resulting in 40 (10 syllables × 4 tones) different existing words in Mandarin Chinese.

2.2.2. Material recording

Speakers were all instructed to produce the 40 words in two different scenarios in sequence: a natural mode ("pronounce these words as if you were talking to a Chinese speaker") and a teaching mode ("as if you were talking to someone who is not a Chinese speaker"). In both conditions, they were explicitly told not to use any hand motions, but there were no other instructions or constraints imposed on the way they should produce the stimuli. There was a 20-minute break between the two recordings, with the recording of the natural stimuli preceding the recording of the teaching style stimuli.

Eye-catcher (version 3.5.1) and Windows Movie Maker (2012) were used to record the speakers' images and sounds. One of the advantages of Eye-catcher is that the camera is located behind the computer screen, which is convenient for capturing the full-frontal images of speakers' faces unobtrusively, similar to what listeners see in a face-to-face setting.

Two sets of 160 video stimuli (10 syllables × 4 tones × 4 speakers in teaching mode and natural mode each) were generated. These two sets of video clips (normal speaking style and teaching style) were segmented into chunks of individual tokens, with each token containing one stimulus. Format Factory (version 3.9.5) was used to extract the sound from each video.

2.3. Procedure

2.3.1. Acoustic analysis

Praat 6.0.33 [29] was used to measure the acoustic parameters of each of the 320 materials (4 speakers x 40 words x 2 speaking styles). In order to automatically extract the vocal parameters from the speech segments, designated Praat scripts were written. A textgrid command firstly was used to get the actual sounding parts from the audio segments. We manually checked all the processed segments to make sure all the sounding parts were captured fully and correctly. After that, another script was used to track the acoustic characteristics of the proceeded materials. The defaults parameter settings for speech analysis in Praat were used (for instance the standard pitch range of 75 - 500 hertz). The relevant acoustic parameters of each stimulus were measured by Praat: the mean of the pitch, the minimum and maximum of the pitch, the range of the pitch (pitch_max – pitch_min) and the duration of the sounding segments.

2.3.2. Visual analysis

Flow Analyzer[1] was employed to track the amount of facial movements present in the videos. Total amount of motion was measured for each of the 320 video segments. Meanwhile, the motions displayed in horizontal (X) and vertical (Y) directions were also measured (for more detailed information, please see [30]).

3. Results

3.1. Acoustic analysis

Figure 1 provides an illustrative example of the difference between the two speaking styles for the four tones. The figure clearly shows the expected rising and falling patterns, which are more pronounced (especially in their duration) in the teaching style. A repeated-measures ANOVA, with tone and speaking style as the within-subject factors and speaker as between-subject factor, revealed that speaking style had a significant effect on the duration of the stimuli, F (1, 9) = 63.3, $p < .001$, $\eta_p^2 = .876$. In line with our expectation, the average duration of the stimuli in the teaching style (M = 0.54 ms, SE = 0.02) was longer than in the natural style (M = 0.48 ms, SE = 0.02), t (9) = 7.75, $p < .001$ (see Figure 2).

[1] FlowAnalyzer is a piece of software, based on Optical Flow Analysis, for extracting motion from 2D video sequences. Optical flow computes pixel displacements between consecutive frames in the video.

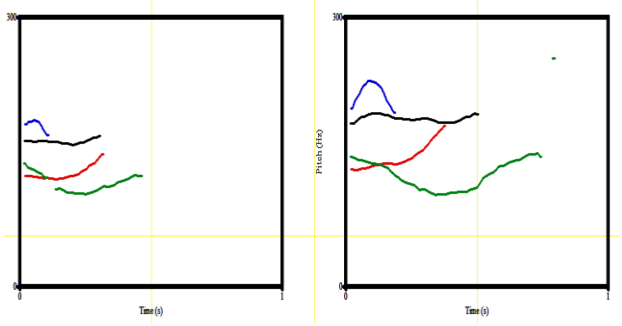

Figure 1: *Plots of tone contours for natural (Left) and teaching style (Right). Figure based on 1 male speaker producing syllable /ma/. To illustrate differences in duration and pitch between the two speaking styles, the scale of the x and y axis is kept identical: time (0-1s) on the x-axis and pitch (0-300 Hz) on the y-axis. Tone 1-black; Tone 2-red; Tone 3-Green; Tone 4-blue.*

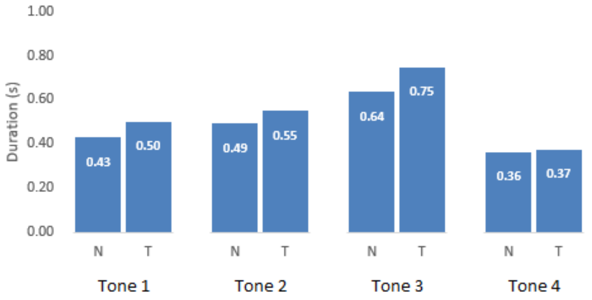

Figure 2: *Durations of the tones in teaching (T) and natural (N) styles*

The average pitch was not influenced by speaking style, F (1, 9) = 2.09, p = .183, $ŋ_p^2$ = .188. That does not surprise us, since tone fidelity should not be influenced by speaking styles. Figure 3 depicts the pitch ranges of the tones in teaching and natural styles. Speaking style in general does not have a significant effect on the average pitch range between teaching and natural style, F (1, 9) = .54, p = .48, $ŋ_p^2$ = .056, which is not in line with our hypothesis. Tone contributed a main effect on the pitch range: F (3, 27) = 18.32, p < .001, $ŋ_p^2$ = .671.

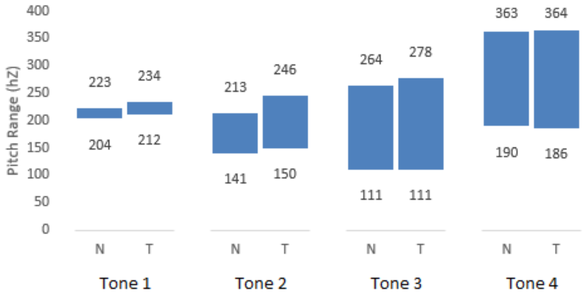

Figure 3: *Pitch ranges of the tones in teaching (T) and natural (N) styles*

Speaker differences accounted for a decent amount of variation in the average duration of the experimental stimuli, F (3, 27) = 8.01, p = .001, $ŋ_p^2$ = .471, and for even more variation in average pitch, F (3, 27) = 28.7, p < .001, $ŋ_p^2$ = .761. Tones accounted for a large amount of variation between the two speaking styles in duration: F (3, 27) = 399.8, p < .001, $ŋ_p^2$ = .978 and in pitch: F (3, 27) = 66.4, p < .001, $ŋ_p^2$ = .881.Thus, our global acoustic analysis reveals strong effects of individual speakers and tones, while small differences between the two speaking styles also emerge.

3.2. Visual analysis

A repeated-measures ANOVA showed that speaking style had a main effect on the total amount of motion, F (1, 9) = 115, p < .001, $ŋ_p^2$ = .928. In teaching style (M = 0.25, SE = 0.01), speakers tended to use more visual cues than in the natural style (M = 0.15, SE = 0.003), t (9) = 10.78, p < .001, which is in line with the idea of hyperarticulation. Individual speakers also differed significantly in their amount of movement, F (3, 27) = 19.56, p < .001, $ŋ_p^2$ = .685. Pairwise comparisons (using Bonferroni adjustment) showed that speaker 1 and speaker 3 provided the most visual movement information and that there is no difference between speaker 1 (M = 0.27, SE = 0.023) and speaker 3 (M = 0.21, SE = 0.008). Speaker 4 (M = 0.18, SE = 0.003) displayed significantly less facial information than speaker 1 and speaker 3 (p = .009 and p = .018 respectively). Speaker 2 (M = 0.13, SE = 0.009) signaled the least visual information. The different tones did not affect the amount of movement, F (3, 27) = 2.27, p = .103, $ŋ_p^2$ = .202.

Except for the total amount of the facial motions, the Flow Analyzer also measured the motions displayed in the horizontal direction (X) and the vertical direction (Y), which can give a clearer picture of the directionality or type of motions among different tones. Tone variations have a significant effect on the amount of facial motions present on the horizontal (F (3, 27) = 3.82, p = .021, $ŋ_p^2$ = .298) and vertical directions (F (3, 27) = 29.74, p < .001, $ŋ_p^2$ = .768). As shown in Figure 4, tone 1 has the least amount of vertical movement (M =0.029, SE = 0.001) and the most horizontal movement (M = 0.037, SE = 0.005), which is in line with a level tone. For tone 2 (the rising tone), tone 3 (the dipping tone) and tone 4 (the falling tone), on the other hand, there is more vertical than horizontal motion.

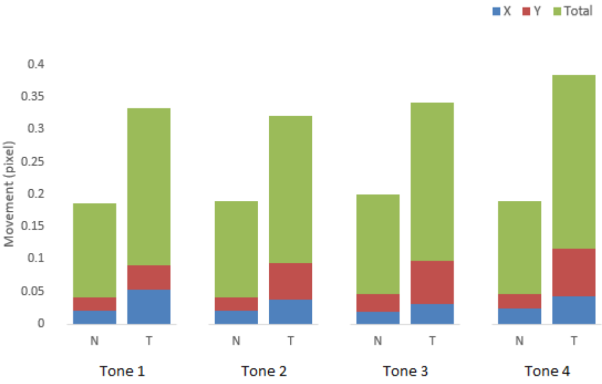

Figure 4: *Facial motions among tones displayed in horizontal (X) and vertical (Y) directions, as well as the total motions, in teaching (T) and natural (N) styles.*

4. Discussion and Conclusions

Our acoustic and visual analyses show important differences between natural and teaching style in both the auditory and visual domain, though the differences between the two speaking styles are admittedly small. Also, the audio-visual analyses reveal strong effects of individual speakers and tones.

In line with our hypotheses, speaking style has a significant effect on the duration of tones. In teaching style, speakers tend to prolong the pronunciation of the tones, possibly to make it facilitate understanding by (non-native) listeners. Notably, not all tones can be effectively prolonged: tone 4 is lengthened the least, even in the inducing "teaching style", because of its intrinsically short vowel duration.

The teaching style has no significant effect on the average pitch of the tones, which is in line with our expectations since pitch is associated with the meaning of the word and cannot be altered without changes in semantics. However, we considered the possibility that the range of the slope of the pitch has been increased as a result from the exaggerated style, while the average level of the pitch range is not influenced by the teaching style. So, we expected that speaking style would have an influence on pitch *range*, with this effect being stronger for some tones than for others. Tone 1 is a level tone, making it difficult to exaggerate the pitch range. Consequently, there are barely any range changes between teaching and natural style for this tone. Tone 3 is a dipping (falling-rising) tone, so the average change of the range caused by hyperarticulation could be subtle. Tone 4 is too short to generate a noticeable change in pitch range. The most promising tone in terms of potential for increased pitch range in exaggerated articulation is tone 2 (the rising tone). However, we did not find a signifant change in range for this tone. Pitch contour and direction are the main acoustic features for speakers to change to accommodate perception by listeners, but modification of these acoustic features is limited because of their relation with the lexical meaning of the uttered words.

With respect to the visual analyses, the findings are all in line with our expectations. More facial motions are signalled by speakers when they produce Mandarin tones in the teaching style. Even though Optical Flow is a relatively straightforward and coarse method to summarize the directionality of facial movements, we still found a general pattern between tone contour and the type of facial movements: speakers produce more horizontal movements when they articulate a level tone (tone 1). More vertical movements can be found when speakers produce contour tones (tone 2, 3 and 4).

A possible explanation for the small effects of speaking style lies in the way the experimental stimuli were generated: as single, isolated words both in the natural and in the speaking style. Even in the natural style, producing separate words/tones may well lead to hyperarticulation to some extent (reflecting the difference between read and natural speech [31]. For future studies, extracting words from continuous speech, for example, by asking participants to read out complete sentences or extracting words from spontaneous productions, could result in a relatively authentic natural speaking style.

In addition, future research could provide more in depth analysis of the relationship between the visual and auditory channel. It remains to be seen whether speakers that produce clear tones also provide abundant visual information with respect to tone identity or whether there is a trade-off between providing visual and auditory cues to tone identity. Of course, such a study first needs to identify the relevant acoustic properties that differentiate between a natural speaking style and a teaching style. For this, we plan to analyse more fine grained acoustic features (e.g., jitter, shimmer, harmonics to noise ratio) as well as a more detailed analysis of prosodic contours [32].

In sum, when speakers try to produce Mandarin tones in a teaching style, the main acoustic information which is affected is the duration of the tones. The average pitch and the pitch range are not influenced by an "exaggerated" pronunciation. Consciously or unconsciously, speakers give out more facial information when they apply teaching style. By being aware of such acoustic and visual information, listeners may benefit from the speakers for Mandarin tone perception.

5. Acknowledgements

This project is supported by China Scholarship Council (CSC).

6. References

[1] Bailly, G., Perrier, P., & Vatikiotis-Bateson, E. (Eds.). (2012). *Audiovisual speech processing*. New York, NY: Cambridge University Press.

[2] Burnham, D., Lau, S., Tam, H., & Schoknecht, C. (2001). "Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers", in *AVSP 2001-International Conference on Auditory-Visual Speech Processing, Aalborg, Denmark.*

[3] Calvert, G., Spence, C., & Stein, B. (2004). *The handbook of multisensory processes*. Cambridge, Mass: MIT Press.

[4] Campbell, R., Dodd, B., & Burnham, D. (Eds.) (1998). *Hearing by Eye II*. Hove, ES: Psychology Press Ltd.

[5] McGurk, H., & MacDonald, J. (1976). "Hearing lips and seeing voices". *Nature, 264*, 746-748.

[6] Skowronski, M. D., & Harris, J. G. (2006). "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments". *Speech Communication*, 48(5), 549-558.

[7] Ferguson, S. H., & Kewley-Port, D. (2007). "Talker differences in clear and conversational speech: Acoustic characteristics of vowels". *Journal of Speech, Language, and Hearing Research*, 50(5), 1241-1255.

[8] Smiljanić, R., & Bradlow, A. R. (2009). "Speaking and hearing clearly: Talker and listener factors in speaking style changes." *Language and linguistics compass, 3*(1), 236-264.

[9] Uchanski, R. M. (2005). *Clear speech*. In D. B. Pisoni & R. Remez (Eds.). *The handbook of speech perception*. Malden, MA/Oxford, UK: Blackwell.

[10] Ferguson, C. A. (1975). "Toward a characterization of English foreigner talk". *Anthropological linguistics*, 1-14.

[11] Ferguson, C. A. (1981). "'Foreigner talk' as the name of a simplified register". *International Journal of the Sociology of Language*, 1981(28), 9-18.

[12] Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

[13] Kong, Q. M. (1987). "Influence of tones upon vowel duration in Cantonese". *Language and Speech*, 30(4), 387-399.

[14] Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). "Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers". *Journal of Phonetics*, 36(2), 268-294.

[15] Hallé, P. A., Chang, Y. C., & Best, C. T. (2004). "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners". *Journal of Phonetics*, 32(3), 395-421.

[16] Chao, Y. R. (1948). *Mandarin Primer*. Cambridge, Mass: Harvard University Press.

[17] Xu, N., & Burnham, D. (2010). "Tone hyperarticulation and intonation in Cantonese infant directed speech". *In Speech Prosody 2010-Fifth International Conference, Chicago, United States.*

[18] Burnham, D., Reynolds, J., Vatikiotis-Bateson, E., Yehia, H., Ciocca, V., Morris, R. H. ... Jones, C. (2006). "The perception and production of phones and tones: The role of rigid and non-rigid face and head motion." *Proceedings of the 7th International Seminar on Speech Production, Ubatuba, Brazil.*

[19] Mixdorff, H., Charnvivit, P., & Burnham, D. K. (2005a). "Auditory–visual perception of syllabic tones in Thai". In E. Vatikiotis-Bateson, D. Burnham, S. Fels (Eds.), *Proceedings of AVSP 2005, International Conference on Auditory–Visual Speech Processing (pp. 3–8). Adelaide, Australia: Causal Productions.*

[20] *Mixdorff, H., Hu, Y., & Burnham, D. (2005b).* "Visual cues in Mandarin tone perception." *In Proceedings of Eurospeech 2005 (InterSpeech-2005): Lisbon, Portugal, 405-408.*

[21] Mixdorff, H., Lirong, M. C., Nguyen, D. T., & Burnham, D. (2006). "Syllabic tone perception in Vietnamese". *Proceedings of International Symposium on Tonal Aspects of Languages 2006, La Rochelle, France, pp. 137–142.*

[22] *Mixdorff, H., & Charnvivit, P. (2004).* "Visual cues in Thai tone recognition". *In International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing, China.*

[23] Chen, T. H., & Massaro, D. W. (2008). "Seeing pitch: Visual information for lexical tones of Mandarin-Chinese." *The Journal of the Acoustical Society of America*, 123(4), 2356-2366.

[24] Reid, A., Burnham, D., Kasisopa, B., Reilly, R., Attina, V., Rattanasone, N. X., & Best, C. T. (2015). "Perceptual assimilation of lexical tone: The roles of language experience and visual information". *Attention, Perception, & Psychophysics, 77*(2), 571-591.

[25] Xu, Y., & Sun, X. (2002). "Maximum speed of pitch change and how it may relate to speech". *The Journal of the Acoustical Society of America, 111*(3), 1399-1413.

[26] Tseng, C. Y., (1981). *An acoustic phonetic study on tones in Mandarin Chinese*. PhD dissertation. Brown University, Providence, RI.

[27] Tseng, C. Y., Massaro, D. W., & Cohen, M. M. (1986). "Lexical tone perception in Mandarin Chinese: Evaluation and integration of acoustic features". In H. S. R. Kao & R. Hoosain (Eds.), *Linguistics, psychology, and the Chinese language* (pp.91-104). Centre of Asian Studies, University of Hong Kong.

[28] Gagné, J. P., Masterson, V., Munhall, K. G., Bilida, N., & Querengesser, C. (1994). "Across talker variability in auditory, visual, and audio-visual speech intelligibility for conversational and clear speech". *Journal-Academy of Rehabilitative Audiology, 27*, 135-158.

[29] Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.33, retrieved from http://www.praat.org/.

[30] Beauchemin, S. S., & Barron, J. L. (1995). The computation of optical flow. *ACM computing surveys (CSUR)*, *27*(3), 433-466.

[31] Lieberman, P., Katz, W., Jongman, A., Zimmerman, R., & Miller, M. (1985). "Measures of the sentence intonation of read and spontaneous speech in American English". *The Journal of the Acoustical Society of America*, 77(2), 649-657.

[32] Ladd, D. R. (2008). *Intonational phonology.* Cambridge University Press.

Appendix: *List of words used for producing the stimuli.*

mā	má	mă	mà
yī	yí	yǐ	yì
xiē	xié	xiě	xiè
shē	shé	shě	shè
shī	shí	shǐ	shì
yōu	yóu	yŏu	yòu
fēn	fén	fĕn	fèn
fū	fú	fŭ	fù
pō	pó	pŏ	pò
yīng	yíng	yǐng	yìng