Articulatory Reduction in Mandarin Chinese Words

Jeffrey Berry[1], *Sunjing Ji*[1], *Ian Fasel*[2], *Diana Archangeli*[1]

[1] Department of Linguistics, [2] School of Information: Science, Technology, and Arts,
The University of Arizona, Tucson, Arizona, USA

{jjberry@email, sunjing@email, ianfasel@cs, dba@email}.arizona.edu

Abstract

We investigate the effect of reduction induced by repetition during articulation. Specifically, we report how tongue movement differs between the first mention of Mandarin words and that of later repetitions using ultrasound imaging. Two analyses were carried out in this paper: tongue deformation and timing. We used Dynamic Time Warping to measure the tongue deformation from a neutral position. We used Functional Data Analysis to measure the timing difference between the first and later repetitions. We found that the tongue deviates less from the neutral and moves faster in time for the later repetitions, namely the more reduced ones. Our study shows promise for more thorough investigation of speech reduction from the articulatory perspective, and provides insights for constructing applications for speech synthesis/recognition towards more natural speech.

Index Terms: Speech reduction, ultrasound imaging, Mandarin Chinese, dynamic time warping, functional data analysis

1. Introduction

One well known cause of variation in speech is what is often termed 'speech reduction,' which usually refers to the shortening of utterances. The majority of research on speech reduction has examined only the audio signal, and has focused on differences in duration and spectral effects. Here we expand the investigation to the articulatory domain, using ultrasound images of the tongue. We examine reduction in two different ways: differences in the amount of deformation the tongue undergoes to achieve a constriction, and differences in the timing of the tongue tip gestures. We use dynamic time warping (DTW) to examine deformation and functional data analysis (FDA) to examine timing differences. Using these methods, we find that repeated words show reduction in ways that are similar to variations observed due to differences in prosodic domains [1, 2]. Specifically, we found that the tongue shows less range of motion and more speed-up in later repetitions.

Fowler & Housum [3] first noticed that after a content-word has been introduced by a speaker, subsequent mentions of that word are shortened in time. [4] showed that speakers shorten subsequent mentions regardless of whether their interlocutors heard the first mention. Speakers were also shown to reduce even when the first mention was made by another person. Based on similar findings from the switchboard corpus, even after controlling for word predictability and frequency, [5] argue that a coordinating mechanism exists between conceptual/lexical planning and articulation that keeps the two synchronized. After a word has been mentioned, its lexical activation is temporarily increased, which speeds up subsequent access and articulation in later repetitions of the word.

The observed effects of repetition on the duration of the word raises questions about how articulations change when duration is shortened. Under theories of speech production such as H&H theory of [6], reduced words are expected to show undershoot, or less deformation from a neutral position. [1] examined this type of variation in tongue gestures, and found that the tongue had less contact with the palate in prosodically reduced positions, such as at the end of syllables. In addition to changes in deformation, differences in the gestural timing patterns were found as well. [2] found differences in the overlap of gestures depending on the type of prosodic boundary that the overlap occurred on. Using FDA [7], they showed non-linear time deformations between gestures for different prosodic boundaries.

We believe that it is of strong interest to extend the investigation to the articulatory differences in reduced utterances. Investigation into this question will contribute to the understanding of speech articulation in running conversation. It will also have implications for creating applications such as automatic speech synthesis and recognition for more natural speech. In this study, we use ultrasound imaging to examine the differences in tongue deformation and gestural timing of Mandarin disyllabic words due to repetition reduction.

2. Method

2.1. Design, materials and participant

One way to investigate speech reduction is to study frequent words contrasted with infrequent words. However, the articulatory difference observed in this case may be largely due to segmental differences rather than reduction. To avoid this problem, we therefore use consecutive repetition as our experimental task. The design of the present study is 5 (Repetition) $\times$ 2 (Aspiration). The Repetition factor is designed to induce speech reduction. The Aspiration factor is designed to see if the aspiration of the obstruent at the intervocalic position affects tongue movement. Aspiration has been shown to affect the realization of pitch [8]. The fortis vs. lenis distinction between aspirated and unaspirated consonants may affect tongue posture, with aspirated consonants possibly showing more tongue movement. For the aspiration distinction, we used the Mandarin Chinese coronal obstruents /ʈʂ ʈʂʰ tɕ tɕʰ/ (Pin-yin */zh ch j q/* respectively) in intervocalic positions.

Table 1 shows the list of stimuli used in this study. All the stimuli are of the CV(C)CV(C) form, mimicking typical Chinese names. The second syllables of all the stimuli are controlled for frequency according to the frequency count of syllables with tones provided by [9]. The logarithm of the mean is 12.12 for the voiceless syllables and 12.29 for the voiced respectively.

A single female native speaker of Mandarin Chinese par-

28 – 31 August 2011, Florence, Italy

Table 1: Mandarin Stimuli (Pin-yin)

	Voiced	Voiceless
Palatal (j vs. q)	sūn **j**iān wáng **j**iǎng zhāng **j**ǐng lǐ **j**īng	sūn **q**iān wáng **q**iǎng zhāng **q**ǐng lǐ **q**īng
Alveolar (zh vs. ch)	zhōu **zh**èn liú **zh**í cáo **zh**ù xǔ **zh**ǎn	zhōu **ch**èn liú **ch**í cáo **ch**ù xǔ **ch**ǎn

ticipated in this study. The list of items were randomized and displayed on a screen in Chinese orthography. The items were embedded in the frame sentence *tā pà . . . ma?* 'Is he/she afraid of . . . ?' The subject was asked to repeat the sentence 5 times in succession at a natural pace before moving onto the next item.

2.2. Ultrasound tongue imaging & tongue trace extraction

Ultrasound imaging of the tongue is a popular method for studying articulation of speech, due to its inexpensive and non-invasive nature. Typically, ultrasound is used to view the mid-sagittal tongue surface contour in real time by placing the transducer beneath the chin during speech. The difference in density between the tissues of the tongue and the air above the tongue cause reflections of the ultrasound waves that result in a white band in the image. The lower edge of this band indicates the position of the tongue surface. Studies of speech sounds using ultrasound imaging often make use of a trace of the tongue surface for data analysis [10]. In this study we used the automatic method of [11] to trace the tongue surface in the images. After automatic tracing, each trace was manually inspected and corrected if necessary. The ultrasound video was recorded at 30 frames per second, resulting in a snapshot of the tongue surface every 33ms. On average, the target words (excluding the frame sentence) had durations of 13.2 frames, or 436ms.

2.3. Measuring deformation with dynamic time warping

To analyze the deformation of the tongue, we first selected a set of tongue images that showed the tongue in an inter-speech posture, which is a language-specific posture in which the tongue is ready for speech [12]. We averaged the tongue surface traces from this set to obtain a neutral position from which we can compute the deformation needed to achieve a posture during speech. In order to measure the deformation of the tongue from a neutral position, we applied dynamic time warping (DTW) to each frame using the `dtw` package for R [13].

DTW has been widely used in various speech-related applications such as alignment in speech recognition [14]. Here, the neutral position is the reference pattern, and the tongue posture during articulation is the test pattern. DTW is appropriate here because the reference pattern and the test pattern usually do not have the same length. In DTW, a cost matrix is calculated as the Euclidean distance of every data point in the reference pattern to every point in the test pattern. The alignment of the two patterns is decided based on the path with minimum sum across the matrix. The globally optimal alignment can be obtained through dynamic programming [15]. The sum of all the values along the alignment path is obtained as the distance score for the two patterns. The distance between the neutral and the raw tongue surfaces during articulation is the deformation measurement, referred to as the DTW score.

Figure 1 shows the frame-by-frame DTW score for all 5 repetitions of the word 'zhōu chèn' in time. At each frame, the DTW score represents the cost of aligning the tongue trace from that frame to the averaged inter-speech posture. The patterns in figure 1 are similar for all repetitions of the same word 'zhōu chèn', i.e. there are the same numbers and sequence of peaks and valleys. However, the range of DTW scores (the difference between the maximum and the minimum DTW scores of a word) for the later repetitions is smaller.

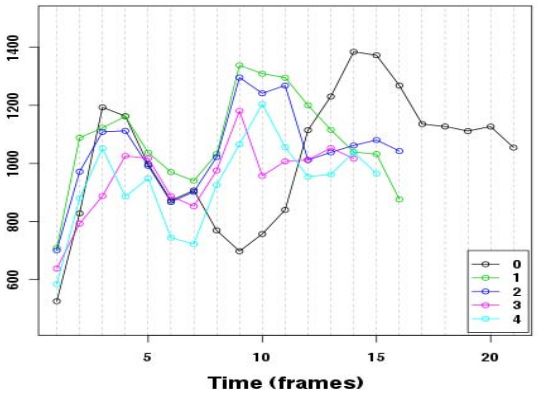

Figure 1: Frame-by-frame DTW scores for 5 consecutive repetitions of the word 'zhōu chèn' uttered in a carrier phrase.

2.4. Measuring timing with functional data analysis

As [2] showed using functional data analysis (FDA) [7], changes in articulation often exhibit non-linear changes in time. We chose a ray in the ultrasound image and tracked the point at which the tongue tip crossed this ray through time using FDA. FDA assumes that the same underlying processes generate each sequence of data. For comparison across words, FDA models each discrete sequence x as a smoothed representation $y(t)$ which is a linear combination of a set of predefined basis functions,

$$y(t) = \sum_{k=1}^{K} c_k \phi_k(t), \quad (1)$$

where c_k is the weight for the basis function $\phi_k(t)$ and K is the number of basis functions. Modeling $y(t)$ is done by finding weights c that minimize the cost function,

$$F(x, y, \lambda) = \sum_j [x_j - y(t_j)]^2 + \lambda \int \left(\frac{d^4 y(t)}{dt^4} \right)^2 dt, \quad (2)$$

where λ is a roughness penalty that controls smoothness. As is typical with non-periodic data, we used a basis of B-splines to smooth the tongue tip trajectories and obtain a functional representation that could be compared across words.

Once the data sequences are modeled as functions, the phase variation between words is calculated by estimating a time-warping function $h_i(t)$ which transforms time t for an individual word i to a reference clock time. When $h_i(t) < t$, the timing of the tongue tip gestures in word i is moving faster than the reference time, and vice versa. Since we were interested

in the difference of later repetitions to the first mention, we set the trajectory of the first mention as the reference time and estimated $h_i(t)$ for each subsequent repetition. The estimation of $h_i(t)$ was done using the time registration method, in which landmarks are chosen that are easily identifiable in each word. To facilitate the identification of the landmarks, we chose to use 0-crossings in the first derivative of the tongue tip position trajectories, which represents instantaneous velocity of the tongue tip. At the landmark locations we require $h_i(t) = t$. $h(t)$ is estimated by minimizing a cost function that is similar to the cost function for estimating $y(t)$,

$$D(\widehat{y}, y, \lambda, w) = \int_0^T [\widehat{y}(h(t)) - y(t)]^2 \, dt + \lambda \int_0^T w(t)^2 dt. \tag{3}$$

$w(t)$ controls the smoothness of $h(t)$ and $\widehat{y}(h(t))$ represents the aligned function values. All smoothing and estimation of time-warping functions was done using the `fda` package in R [16]. Figure 2 shows example velocity trajectories before and after alignment using FDA.

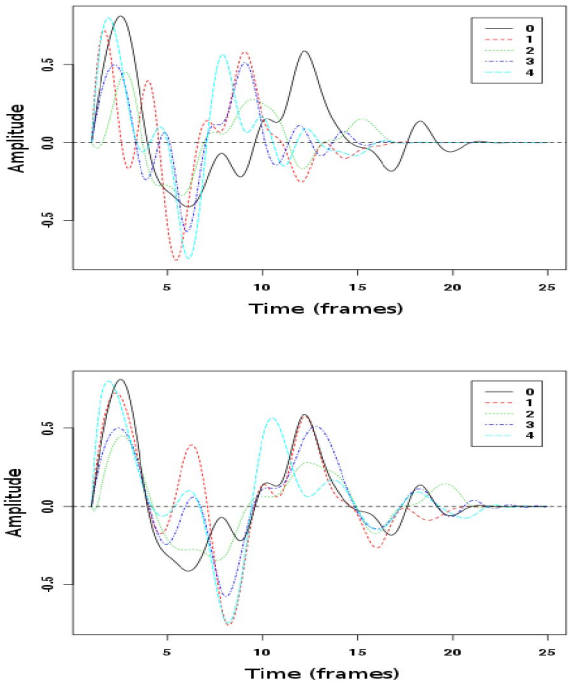

Figure 2: *Top*: unaligned tongue tip velocity trajectories for 5 consecutive repetitions of the word 'zhōu chèn'; *Bottom*: the same trajectories after alignment with FDA.

3. Results

3.1. Deformation

As seen in figure 1, subsequent repetitions appear to have less overall difference in the deformation of the tongue when compared to the first repetition, i.e. there is a smaller range in the scores between peaks and valleys. In order to test this, we computed the range of the DTW scores for each repetition by subtracting the maximum DTW score for that word from the minimum. Figure 3 shows these range scores aggregated by repetition and aspiration. A clear declination in the range of motion is apparent in later repetitions. A two-factor ANOVA (repetition and aspiration as factors) showed a significant main effect of repetition [F(1, 12) = 5.85, $p < 0.033$]. There was no effect of aspiration and no interaction.

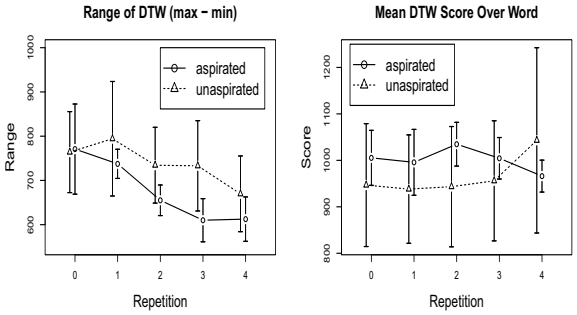

Figure 3: *Left*: Comparison of the difference between maximum and minimum DTW scores by repetition. The effect of repetition is significant. *Right*: Comparison of the mean DTW score by repetition. There are no significant differences.

We also tested the hypothesis that the later repetitions would have lower *overall* deformation from the neutral position. To test this, we compared the mean DTW scores by repetition and aspiration, as shown in figure 3, right. As seen from the figure however, there is no mean difference between repetitions. A two-factor ANOVA confirmed that there were no significant effects of repetition or aspiration, and no interaction.

The results based on the *range* of DTW scores show that there is an effect of repetition on tongue posture deformation. However, the results based on the *mean* DTW scores show no difference across repetitions. These results can be understood by examining Figure 1. In Figure 1, peaks correspond to the extreme deformation of the tongue away from the neutral position, and are aligned in time to the constrictions of the obstruents; valleys occur where the tongue is relatively closer to the neutral position, corresponding in time to the opening of the vocal tract for articulation of the vowels. The non-significant difference in the mean DTW scores suggests that overall, the tongue postures are similar across words. However, the difference in range results suggests that the deformation of the tongue is compressed in later repetitions, i.e. constrictions for the consonants are not as tight, and openings for vowels are not as wide, which is in line with the findings of [1]. When averaging the DTW scores across time, the similar magnitude of peaks and valleys within a word result in a mean score that does not show the difference in the magnitudes across words. The similarity of mean scores across repetitions of the same word suggests that there is a minimum tongue posture deformation from neutral needed to preserve contrast between different segments. The reduction lies in the deviation from this minimum deformation from neutral as shown by the range results.

3.2. Timing

Many studies on speech reduction focus on differences in duration for repetitions [3, 4, 5]. We also found a significant effect of repetition on duration using a two-factor ANOVA [F(1, 12) =

21.30, $p < 0.001$]. Later repetitions were significantly shorter in time than earlier repetitions, as shown in figure 4.

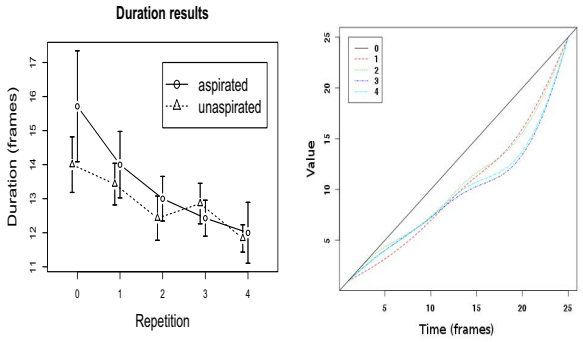

Figure 4: *Left*: Comparison of duration by repetition. The effect of repetition is significant. *Right*: The time-warping functions for the repetitions of 'zhōu chèn'.

The results from FDA show that the differences in duration are non-linear in nature, with more speed-up occurring at the end of the word. The time-warping functions for the trajectories of the repetitions of the word 'zhōu chèn' are shown in figure 4. The time-warping functions show greater distance from the reference trajectory (represented by the diagonal line) towards the end of the word, indicating that the later repetitions are sped up more at the end of the word. In other words, the tongue tip experiences greater accelerations at the later part of the word. These findings are consistent with the results of [2], who found that gestures in more prosodically prominent positions, such as the beginnings of syllables, are slowed in time. Similarly figure 4 shows less speed up at the more prominent beginning of the word.

4. Discussion

These results show that repetition induces speech reduction in both gestural timing and tongue posture. Later repetitions show less range of motion in the tongue and faster execution. These findings are consistent with previous studies on the effects of repetition on duration [3, 4, 5]. Our results differ from these studies in two ways: our results are from direct analysis of tongue movements rather than based on the acoustic signal, and our results were obtained from simple repetitions in a laboratory setting. In contrast, [5] used speech extracted from conversations in the switchboard corpus. This has implications for the design of future studies, which should take effects of repetition into account when analyzing duration of stimuli. Our results on timing differences show non-linearities, with the ends of syllables showing more speed-up in later repetitions. These findings are consistent with results of studies on prosodic effects on articulation [2].

Although we found no significant differences for aspiration, there appear to be differences in the variation between the two groups, with voiced tokens showing more variation in the DTW scores. Future studies with more subjects are needed to investigate effects of aspiration on tongue posture. In summary, these results show that ultrasound images of the tongue show promise for more thorough investigations of speech reduction. Specifically, the development of analyses that make use of the entire tongue surface contour rather than a single point would likely provide new insights into the mechanisms underlying speech production and reduction.

5. Acknowledgements

This work was made possible in part by James S. McDonnell Foundation grant #220020045 BBNB, ONR contract N00014-09-1-065 as part of the "Science of Autonomy" program, the DARPA "CLIME" seedling, contract N10AP20008, and by an Arts, Humanities & Social Sciences (AHSS) grant from the University of Arizona.

6. References

[1] C. Fougeron and P. Keating, "Articulatory strengthening at edges of prosodic domains," *The Journal of the Acoustical Society of America*, vol. 101, no. 6, pp. 3728–3740, 1997.

[2] D. Byrd, S. Lee, and R. Campos-Astorkiza, "Phrase boundary effects on the temporal kinematics of sequential tongue tip consonants," *The Journal of the Acoustical Society of America*, vol. 123, pp. 4456–4465, 2008.

[3] C. Fowler and J. Housum, "Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction," *Journal of Memory and Language*, vol. 26, pp. 489–504, 1987.

[4] E. Bard, A. Anderson, C. Sotillo, M. Aylett, G. Doherty-Sneddon, A. Newlands *et al.*, "Controlling the intelligibility of referring expressions in dialogue," *Journal of Memory and Language*, vol. 42, no. 1, pp. 1–22, 2000.

[5] A. Bell, J. Brenier, M. Gregory, C. Girand, and D. Jurafsky, "Predictability effects on durations of content and function words in conversational English," *Journal of Memory and Language*, vol. 60, no. 1, pp. 92–111, 2009.

[6] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory," *Speech production and speech modelling*, vol. 55, pp. 403–439, 1990.

[7] J. Ramsay and B. Silverman, *Functional data analysis*. Springer, New York, 2005.

[8] J. Hombert, J. Ohala, and W. Ewan, "Phonetic explanations for the development of tones," *Language*, vol. 55, no. 1, pp. 37–58, 1979.

[9] J. Da, "A corpus-based study of character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction," in *The studies on the theory and methodology of the digitized Chinese teaching to foreigners: Proceedings of the 4th International Conference on New Technologies in Teaching and Learning Chinese*. Beijing, The Tsinghua University Press, 2004, pp. 501–511.

[10] J. Mielke, A. Baker, and D. Archangeli, "Variability and homogeneity in American English /ɹ/ allophony and /s/ retraction," in *Laboratory Phonology 10*. Berlin: Mouton de Gruyter, 2010, pp. 699–730.

[11] I. Fasel and J. Berry, "Deep belief networks for real-time extraction of tongue contours from ultrasound during speech," in *20th International Conference on Pattern Recognition*, 2010.

[12] I. Wilson, "Articulatory settings of French and English monolingual and bilingual speakers," Ph.D. dissertation, Citeseer, 2006.

[13] T. Giorgino, "Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package," *Journal of Statistical Software*, vol. 31, no. 7, pp. 1–24, 2009.

[14] L. Rabiner, *Fundamentals of Speech Recognition*. Prentice Hall PTR, Englewood Cliffs, NJ, 1993.

[15] R. Bellman, "On the theory of dynamic programming," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 38, no. 8, p. 716, 1952.

[16] J. Ramsay, H. Wickham, S. Graves, and G. Hooker, *fda: Functional Data Analysis*, 2010, R package version 2.2.5. [Online]. Available: http://CRAN.R-project.org/package=fda