Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Cross-linguistic generalization of the distal rate effect: Speech rate in context affects whether listeners hear a function word in Chinese Mandarin

Wei Lai[1], *Laura Dilley*[2]

[1] Department of Linguistics, University of Pennsylvania

[2] Department of Communicative Sciences and Disorders, Michigan State University

weilai@sas.upenn.edu, ldilley@msu.edu

Abstract

Recent findings show that altering the speech rate of the context several syllables away from a word (i.e., the distal context) can cause the word to disappear in perception in non-tonal Indo-European languages like English [1] and Russian [2]. This study investigated the distal rate effect in Chinese Mandarin, a tonal language belonging to the Sino-Tibetan language family. We examined whether perception of the monosyllabic function word "—" /i/ was affected by the distal rate in casual speech. The results showed that slowing the distal rate caused the function word to be perceived significantly less often than if the distal rate were normal or speeded. The results support a theory of generalized rate normalization, according to which distal speech rate shapes listeners' expectancy towards proximal speaking rate, thereby influencing the number of morphophonological units perceived. This study supports the idea that certain spectro-temporal parameters might be universally tracked for word segmentation across languages and language families, extending prior work on word segmentation to Chinese Mandarin.

Index Terms: distal speech rate, lexical perception, word segmentation, Chinese Mandarin

1. Introduction

It has long remained puzzling how listeners isolate and recognize spoken words from continuous speech. Cues that assist in spoken word segmentation can be categorized as either being derived from knowledge of the language, or else as being derived from the specific acoustic properties of the signal. For the present study, we investigate the role of a specific signal-derived cue – distal speech rate – in word segmentation of casual Mandarin speech. Distal speech rate refers to the rate of speech distant from a given word. Recently, substantial evidence from English has suggested that distal speech rate can have substantial effects on whether or not listeners hear a coarticulated function word [1-3]. For example, in the sentence *Deena didn't have any leisure or time* the function word *or* can be heavily coarticulated and blend spectrally with the preceding syllable, such that, for American English, *leisure or* is pronounced as [liʒɚː]. In [1], when the entire sentence was presented with the original speaking rate, listeners perceived the function word with relatively high accuracy. However, when the distal speech rate preceding and following the function word was slowed down, function word reports dropped from 79% to 33%, even though the function word-containing region was acoustically identical. The distal rate effect also replicates in Russian, for both native Russian listeners and second language Russian learners [2]. These findings indicate the possibility that certain temporal cues could be universally available for language learners in word segmentation and lexical access.

These findings indicate that distal context speech rate influenced the number of morphophonological units (whole syllables or words) that listeners perceived [1-3]. The findings thus contrast with well-known rate effects in phoneme perception, whereby proximal and – to a limited extent – distal speech rate influence perceptual boundaries between phonemic categories along an acoustic continuum [4-7]. Building on such phonemic rate effects, Dilley and Pitt [1] proposed that *generalized rate normalization* might account for this distal rate effect. According to this proposal, listeners entrain to context timing and generate expectations about the typical durations of units like syllables. Encountering a stretch of speech which is relatively fast compared with distal context, such as would occur when context rate is slowed down, would lead to perception of relatively fewer morphophonological units than when the context rate is relatively faster. An alternative possible explanation which was not supported in prior work [1, 2] was that any mismatch in speech rates *per se* between context and the proximal region around the function word distracted listeners' attention in detecting the reduced, heavily coarticulated function words.

It is not clear whether the distal rate effect will occur for Mandarin, since Mandarin differs substantially in phonology, morphology, and other characteristics relative to Indo-European languages such as English and Russian. One phonological difference that would seem relevant in this case is that Mandarin has relatively clear syllable boundaries, such that syllables show hardly any ambisyllabicity or resyllabification [8, 9], in contrast to English. The clarity of syllable boundaries results partly from the strictness of Mandarin syllable structure. Syllables in Mandarin are largely CV, CVV, CVVV, or CVC, allowing only a single consonant in onset position and codas which are restricted to /n/ or /ŋ/ [10]. Also, syllables in Mandarin are available at lexical retrieval and also function as fundamental building blocks in the written system in the forms of characters [9, 11]. The correspondence between a syllable and a morpheme/character leads to speakers' showing higher sensitivity to units of syllables than words in the production process [12]. The

 doi: 10.21437/SpeechProsody.2016-231

higher productive activation of the syllable as a unit, together with the strictness of Mandarin syllable structure, suggest that the syllable might more likely maintain its perceptual integrity under contextual variation. Thus, we might expect that the distal rate effect might be suppressed in Mandarin.

In contrast to the above properties which tend to predict an absence or mitigation of the distal rate in Mandarin, other language facts point toward a different prediction. One such fact is that syllabic reduction, a coarticulation-symbiotic phenomenon documented for both English [13] and Mandarin [10], has been found to occur more frequently in Mandarin than in English [14]. Syllable reduction has been shown to be crucial to triggering the distal rate effect [3], since reduction makes syllables deviate from their canonical forms, ensures smooth spectral transitions at boundaries, and makes syllables less recognizable in running speech. The higher syllable reduction rate in Mandarin is thought to be caused by the fact that Mandarin has a larger proportion of open syllables than English – 75% for Mandarin [10] vs. 40% for English [15] – where open syllables are more prone to reduction [16]. This conjecture was verified in [14], which found that the difference in syllable reduction rate between Mandarin and English became insignificant with syllable type controlled for.

Interestingly, we now have two totally opposite proposals on how syllable type might affect the distal rate effect in Mandarin. The strictness of syllable structure assists syllable isolation with language experience, while the prevalence of open syllables helps blur syllable boundaries with signal casualness. It is difficult to weigh their roles against each other in the generation of distal rate effect.

Another property of Mandarin that cannot be overlooked in word segmentation is that Mandarin is a tonal language. Mandarin has four stressed lexical tones and a neutral tone. Contours of the four stressed tones are phonologically described as high-level, mid-rising, falling-rising and high-falling, which vary greatly from one another. In contrast to the situation in English and Russian, in Mandarin pitch contributes lexical meaning and is expected to provide a particularly strong cue for word segmentation. However, Mandarin sometimes has a contrast between reduced and unreduced syllables in parallel with the contrast between full tone and neutral tone. It is generally accepted that neutral-tone syllables are weakened syllables and can distinguish meaning [17-19], just as reduced syllables do in English. Moreover, heavy tonal coarticulation in running speech will make the tonal contour deviate to a significant extent from its canonical form, such that the tone becomes less recognizable [20]. These properties make it possible that the distal rate effect occurs in real speech in Chinese Mandarin.

To address these diverging predictions, we conducted a perception experiment. The general goal of this study was to evaluate how the tracking of temporal cues in word segmentation interacts with language-specific properties, specifically, in this case, whether and how the distal rate effect previously found in English and Russian may also generalize to Mandarin. A separate goal was to assess the validity of rate mismatch hypothesis versus generalized rate normalization as explanations for the distal rate effect. If a lower function word report rate is observed for any mismatch in speech rate between the context and the target region (either by being slower or faster), then this pattern will tend to support a proposal that rate mismatch leads to reduced function word reports. In contrast, if slowing the context results in a lower proportion of function word reports while speeding the context results in the same or a higher proportion of function word reports, then the generalized rate normalization would be supported. We also expect the results would give us a better understanding of phonological and acoustic cues in word perception and segmentation processes for Chinese Mandarin, issues which have hardly been studied before.

2. Method

2.1. Participants

The participants were 15 native Mandarin speakers (6 male, 9 female), between the age of 23 and 29. All reported that they had normal hearing.

2.2. Materials

We constructed 36 target sentences containing the critical function word "一" /i/. (Hereafter, we will use the PINYIN transcription "yi" instead.) The function word "yi" approximately corresponds to "one" or "a" in English. Although it is counted as a numeral (Num) in most Chinese grammar works, it also expresses indefiniteness in the modified object and can be omitted when the sememe of quantity is not particularly emphasized, especially on informal occasions (e.g., oral conversations).

Each target sentence was embedded with a structure of verb + [numeral + classifier + noun], since this structure was grammatically correct as well without the Num "yi". For example, the sentence "We want to put a pot of plants on the windowsill" can be transliterated as follows, either with or without the function word "yi":

我们 想 摆 **(一)** 盆 花 在 窗台上。
We want (to) put **a** pot(of) plants on (the) windowsill.
| | | | |
V **Num** Cl N

The target sentences were made up of 8-13 syllables. To ensure heavy coarticulation between function word and context, we preposed monosyllabic verbs whose rhymes are either the monophthong /i/ or diphthongs that ended up with reduced /ɪ/ (e.g., /aɪ/, /eɪ/) in front of "yi". These sentences basically bore present and future tense, since past tense requires an additional tense-marker "了" /lə/ between the verb and "yi", and this would interrupt spectral consistency at syllable boundaries.

Materials were recorded at Michigan State University in a professional recording booth by a female Mandarin speaker (the first author). Each sentence was repeated several times, and a single token was selected for each item. We selected tokens in which the function word was pronounced casually, showed shorter duration and evidenced a continuous formant with respect to the preceding syllable rhyme. Specifically, instances of "yi" were selected which were pronounced as /i/ if

the rhyme of the preceding syllable was the monophthong /i/, or else as /ɪ/ if the preceding rhyme was an /ɪ/-final diphthong.

The lexical tone of "yi" can be largely affected by tone coarticulation, yielding flexible tonal realizations in spontaneous speech. Therefore, we capitalized on both tonal sequence constraints and coarticulation to generate acceptable tonal patterns for both one-syllable and two-syllable patterns, in order to avoid strong additional tonal cues for word segmentation, as shown in Fig. 1.

In Fig. 1, the syllables preceding "yi" carried four Chinese stressed lexical tones with features of [level, rising, dipping and falling]. The "yi" evidenced phonetic variability in tonal properties due to tonal coarticulatory effects, so the tonal pitch trend of the first syllable was able to be preserved after including the pitch fragment of "yi", which facilitated hearing one syllables instead of two syllables.

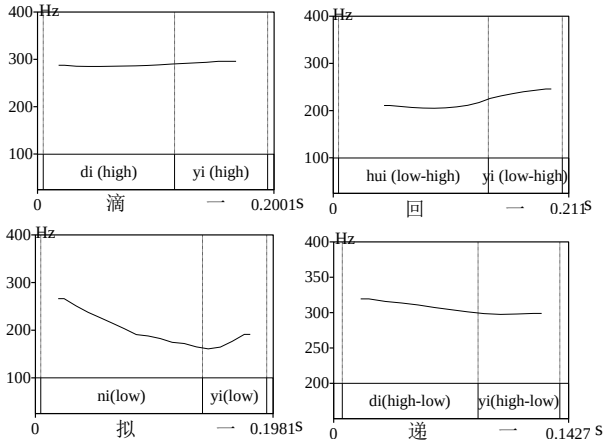

Fig. 1. *The preservation of the original tonal contours on "yi" after tonal coarticulation (Level, Rising, Dipping and Falling).*

72 additional filler sentences were added to increase variety, which contained 5–18 syllables. To maintain consistency with the target sentences, the fillers also mostly used present and future tense and maintained a less formal style of wording and sentence structure.

2.3. Manipulation

Stimuli containing the critical function word "yi" were divided into a target portion and two context portions, by annotated boundaries of the target in Praat [21]. The target corresponded to the critical function word plus the preceding syllable and the following phoneme, as in [1], and the context corresponded to speech fragments both preceding and following the target.

Three Distal Rate conditions were generated in this experiment. In these conditions the target remained unaltered, while the context was respectively: a) slowed through time expansion (Slow condition), b) speeded through time compression (Fast condition), and c) presented at normal speech rate (Normal condition). The time-compression factor was 0.6, and the time-expansion factor was 1.9. These factors had previously been used for English [1] and Russian [2] and shown to be effective in yielding different perceptions of the target. To avoid potential signal mismatch or missing material at target boundaries associated with splicing, rate was altered using the PSOLA algorithm in Praat [21] and its graphical user interface for duration interpolation. We set aside a region of 0.01 sec both before and after the target as transitions where the speech rate slowed down or speeded up gradually. A schematic of the manipulated samples is shown in Fig. 2.

Likewise, one-third of the fillers were manipulated by the temporal factor of 0.6, another one-third by 1.9, and the remaining preserved their original rates. The fillers did not contain target regions, so the speech rate was consistent throughout the whole sentence.

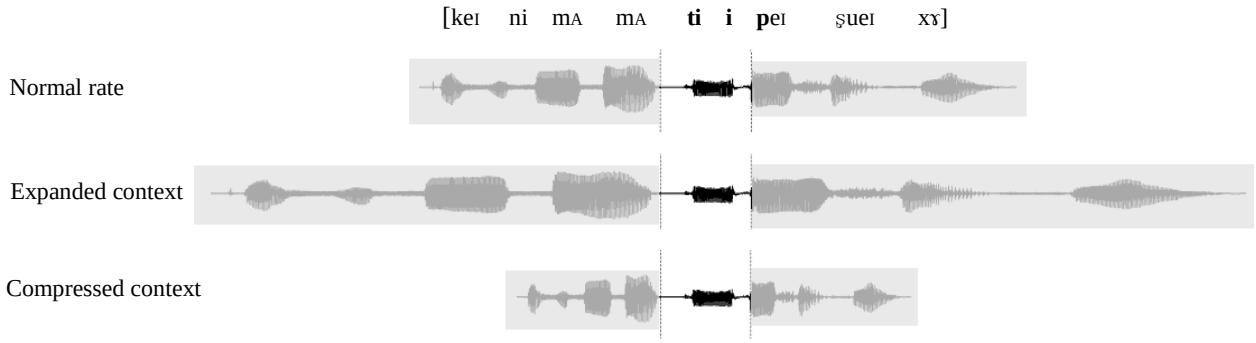

Fig. 2. *Waveforms of a sample time-altered stimulus across the three conditions. The sections of the waveform without background shading (and in boldface in the phonetic transcription.) corresponded to the target region, and those with shading (and not bolded in the phonetic transcription) corresponded to the context.*

2.4. Procedure

Three lists were constructed from 36 experimental stimuli and 70 filler items. The 36 experimental stimuli in each list were made up of 12 items with fast context, 12 items with slow context and 12 items with normal context. The pairing of items with conditions was counterbalanced across the three lists, so that each participant heard each experimental stimulus only once in one of the three speech rates. For each list, the first 5 trials were filler sentences, and the remaining 101 items occurred in quasi-random order, with the constraint that stimuli were all separated apart by one or more fillers, and no more than two items of the same rate (Slow/Normal/Fast) occurred in a row.

Participants were randomly assigned to one of three lists, with an equal number of participants (N = 5) hearing each list. They were instructed to listen carefully to each item for as many times as they liked and to fill in blanks on an answer sheet, producing a veridical transcription of what they heard and typing their responses using a computer keyboard. The blank space corresponded to the function word, the monosyllabic verb preceding the function word, and two extra syllables either prior to or after the target, with the constraint that words were not split apart by the blank. Three or four characters/syllables were expected in each blank depending on whether the participant perceived the function word "yi". Accordingly, blanks contained 3 syllables in one half of the fillers and 4 syllables in the other half contained 4 syllables, so that for both target sentences and filler sentences either 3 or 4 syllables were to be filled.

3. Results

Fig. 3 shows the function word report rate for the three Distal Rate conditions (slow, normal, fast).

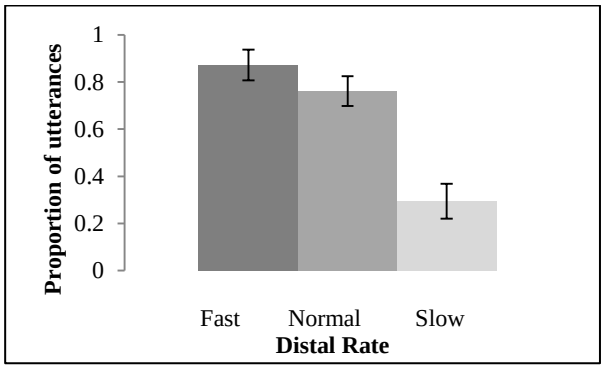

Fig. 3. *Proportion of utterances containing a function word for the three distal speech rate conditions. Error bars show +/- 1 standard error.*

As shown in Fig. 3, in the normal-rate condition (i.e., the baseline), the function word was perceived 76% of the time. That the function word was perceived less than 100% is consistent with other studies of function word perception [1, 22, 23] and is likely due to the reduction and heavy coarticulation of this word. Notably, when the context rate was slowed, function word identification dropped by more than half to 29%, even though the target region remained identical in the acoustic signal. This dramatic reduction of function word reports was not observed for the other kind of rate mismatch, i.e., where the context was fast; on the contrary, the function word report rate increased moderately to 87%. A repeated measures one-way ANOVA showed a significant effect of distal speech rate [by-subjects: $F_1(2,28) = 47.92$, $p < .001$, $\eta^2 = .77$; by-items: $F_2(2,70) = 86.05$, $p < .001$, $\eta^2 = .71$; min-$F'(2,60) = 30.78$, $p < .001$]. Paired-samples t-tests with Bonferroni correction revealed that, in both by-subjects and by-items analyses, all conditions were significantly different from one another ($p < .01$).

It can be observed that function word report rates reduced only when the context was slowed down, not when just any rate mismatch occurred (i.e., when the context was speeded). This finding plausibly lent support to the theory of *generalized rate normalization* against *rate mismatch hypothesis*, indicating that making the duration of a stretch of speech containing a function word faster than its context affected the number of morphophonological units perceived. Through entrainment to context timing, listeners were induced to expect slower speaking rate of the target region and therefore heard fewer morphophonological units than that was originally spoken.

4. Discussion

Little prior perception research has tested how acoustic information affects word segmentation and lexical recognition in Mandarin. It was unclear whether the distal rate effect on lexical segmentation and recognition previously reported for English and Russian [1-3] would generalize to a non-Indo-European tone language such as Mandarin Chinese. An experiment was conducted in which the distal context speech rate in Mandarin sentences containing the coarticulated function word "yi" was altered to be fast, slow, or a normal rate.

The pattern of results was consistent with a generalized rate normalization account. According to this proposal, timing cues in distal context lead to temporal entrainment associated with general auditory perceptual processes; this leads listeners to form timing expectations which affect perception of the number of units (syllables or words) in the target region [1]. By contrast, the results were not consistent with a rate mismatch account, since not all rate-mismatching conditions produced a lowering in function word report rate from the normal-rate baseline (e.g., a decrement in function word reports). (See also [2].) The extent to which domain- and/or language-specific knowledge plays a role in timing expectations in the distal rate effect remains unclear. However, recent results by Pitt et al. [23] and Baese-Berk et al., this volume, suggest that language-specific knowledge affects function word recognition. Future work can address these issues.

5. Conclusions

In summary, the present experiment showed that distal context speech rate can indeed cause a monosyllabic function word to appear or disappear perceptually in Mandarin. These results show that the distal rate effect can be generated in tonal languages as well as in non-tonal languages. In spite of the strict phonological structure of syllables in Mandarin, substantial reduction does occur [14]. The present study showed that when function words are reduced and coarticulated in Mandarin, that distal context speech rate contributes substantially to whether they are heard.

6. Acknowledgements

We gratefully acknowledge the support of NSF Grant BCS #1431063 to LCD. We also would like to thank Yang Wang for his helpful comments on this paper.

7. References

[1] Dilley, L.C., Pitt, M., "Altering context speech rate can cause words to appear or disappear." *Psychol. Sci.*, 21, pp. 1664-1670, 2010.

[2] Dilley, L.C., Morrill, T., Banzina, E., "New tests of the distal speech rate effect: Examining cross-linguistic generalizability." *Frontiers in Language Sciences*, 4, pp. 1-13, 2013.

[3] Heffner, C., Dilley, L.C., McAuley, J.D., Pitt, M., "When cues combine: how distal and proximal acoustic cues are integrated in word segmentation." *Language and Cognitive Processes*, 28, pp. 1275-1302, 2013.

[4] Liberman, A.M., Delattre, P., Gerstman, L., Cooper, F.S., "Tempo of frequency change as a cue for distinguishing classes of speech sounds." *J. Exp. Psychol.*, 52, pp. 127-137, 1956.

[5] Newman, R.S., Sawusch, J.R., "Perceptual normalization for speaking rate: effects of temporal distance." *P&P*, 58, pp. 540-560, 1996.

[6] Kidd, G.R., "Articulatory rate-context effects in phoneme identification." *J. Exp. Psychol. Hum. Percept. Perform.*, 15, pp. 736-748, 1989.

[7] Summerfield, Q., "Articulatory rate and perceptual constancy in phonetic perception." *J. Exp. Psychol. Hum. Percept. Perform.*, 7, pp. 1074-1095, 1981.

[8] Kuo, F.L., "Aspects of segmental phonology and Chinese syllable structure." Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1994.

[9] Chen, T.M., Dell, G.S., Chen, J.Y., "A cross-linguistic study of phonological units: Syllables emerge from the statistics of Mandarin Chinese, but not from the statistics of English." In *Proceedings of the Cognitive Science Society*, pp. 216–220, 2004.

[10] Tseng, S.C., "Syllable contractions in a Mandarin conversational dialogue corpus." *International Journal of Corpus Linguistics*, 10, pp. 63-83, 2005.

[11] Levelt, W.J.M., "Spoken word production: A theory of lexical access." *Proceedings of the National Academy of Sciences*, 98, pp. 13464–13471, 2001.

[12] O'Seaghdha, P.G., Chen, J.Y., Chen, T.M., "Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English." *Cognition*, 115, pp. 282-302, 2010.

[13] Johnson, K., "Massive reduction in conversational American English." In *Spontaneous speech: Data and analysis. Proceedings of the 1st Session of the 10th International Symposium*, pp. 29-54, 2004.

[14] Burchfield, L.A., Bradlow, A., "Syllabic reduction in Mandarin and English speech." *The Journal of the Acoustical Society of America*, 135, pp. EL270-EL276, 2014.

[15] Delattre, P., Olsen, C., "Syllabic features and phonic impression in English, German, French and Spanish." *Lingua*, 22, pp. 160–175, 1969.

[16] Cheng, C., Xu, Y., "Extreme reductions: Contraction of disyllables into monosyllables in Taiwan Mandarin." In *Proceedings of Interspeech*, pp. 456-459, 2009.

[17] Chao, Y.R., A grammar of modern spoken Chinese. Univ of California Press. 1968.

[18] Duanmu, S., Syllabic weight and syllabic duration: A correlation between phonology and phonetics. Phonology, 11(01), pp.1-24. 1994.

[19] Chen, Y. and Xu, Y., Production of weak elements in speech–Evidence from f0 patterns of neutral tone in standard Chinese. Phonetica,63(1), pp.47-75. 2006.

[20] Xu, Y., "Production and perception of coarticulated tones." *The Journal of the Acoustical Society of America*, 95, pp. 2240-2253, 1994.

[21] Boersma, P., Weenink, D., "Praat: Doing phonetics by computer [Computer program]." Software and manual available online at www.praat.org. 2012.

[22] Baese-Berk, M.M., Heffner, C.C., Dilley, L.C., Pitt, M.A., Morrill, T.H., McAuley, J.D., "Long-term temporal tracking of speech rate affects spoken-word recognition." *Psychol. Sci.*, 25, pp. 1546-1553, 2014.

[23] Pitt, M., Szostak, C., Dilley, L., "Rate-dependent speech processing can be speech-specific: Evidence from the perceptual disappearance of words under changes in context speech rate." *Attention, Perception, & Psychophysics*, pp. in press.