INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Vowel Fundamental and Formant Frequency Contributions to English and Mandarin Sentence Intelligibility

Daniel Fogerty [1], *Fei Chen* [2]

[1] Department of Communication Sciences and Disorders, University of South Carolina, USA
[2] Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China
fogerty@sc.edu, fchen@sustc.edu.cn

Abstract

The current study investigated spectral components of vowels that contribute to Mandarin and English sentence intelligibility. Sentences were processed to preserve various amounts of vowel information. Processing parameters ensured similar proportions of speech preserved between the two languages. In the first experiment, speech segments, primarily containing vocalic cues, were processed to flatten fundamental frequency (F0) cues. In the second experiment, sine-wave speech synthesis was used to coarsely code speech to retain only amplitude and frequency variation associated with the first three formants. Results demonstrated remarkable similarity between Mandarin and English sentence intelligibility with flattened F0 sentences. In contrast, the intelligibility of English sentences surpassed that of Mandarin sentences for sine-wave speech. Combined with earlier reports of superior intelligibility of Mandarin sentences with full spectrum vowels, these results highlight significant contributions of Mandarin F0 information, likely related to lexical tone. In contrast, English listeners may rely more on frequency and/or amplitude variation of the formants.

Index Terms: speech recognition, vowels, lexical tone, interruption.

1. Introduction

Previous studies have indicated that acoustic information present during vowel segments provides significant contributions to Mandarin and English sentence intelligibility [1-3]. However, it is not currently clear whether the acoustic contributions from vowels are the same between the two languages, or whether listeners of one language weight some acoustic features more than listeners of the other language. Such a language comparison will assist in defining language-specific and language-general processes for how speech information useful for sentence intelligibility is distributed across the complex acoustic parameters of speech.

The comparison between English and Mandarin Chinese is informative due to a number of acoustic-phonetic differences between the languages. First, as Mandarin is a tone language, lexical information is conveyed by the fundamental frequency (F0). This may result in different vowel contributions to intelligibility compared to English, where F0 is also important [e.g., 4], but does not directly convey lexical meaning. In addition, Mandarin has a sparse vowel system compared to English. This difference, combined with the phonological structure of the language, is likely related to larger vowel inherent spectral changes (VISC) that have been observed for Mandarin vowels compared to English vowels [5]. VISC reflects the slow varying changes in the vowel formants that play an important role for vowel perception [6]. The current study was designed to specifically investigate language differences that occur as a result of relative differences in the way vowel F0 and VISC contribute to sentence intelligibility. Toward this end, two experiments were conducted to independently assess these two acoustic features. Experiment 1 tested sentence intelligibility for F0 flattened sentences as a way of indexing differences in the way vowel F0 contour contributes to overall sentence intelligibility for the two languages. Experiment 2 was designed to assess differences between the two languages in the contribution of amplitude and frequency variations in the first three formants by using sinewave speech to coarsely represent speech according to only these acoustic features. In this way, differences in overall intelligibility between Mandarin and English could be attributed to how well listeners were able to extract meaning from the preserved acoustic cues. This study extends the literature on cross-linguistic vowel differences to examine how specific acoustic differences determine sentence intelligibility.

In addition to differences in F0 and VISC between the two languages, Mandarin and English also have different syllabic structures. Mandarin has a consonant-vowel syllable structure that varies significantly from the complex syllable structure of English that allows for consonant clusters. This difference results in a greater proportion of the sentence accounted for by vowel acoustics in Mandarin compared to English. To control for this durational difference, the total proportion of speech information presented was equated between English and Mandarin testing by examining performance at different preserved proportions of the vowel. Initial testing of the Mandarin listeners was previously reported [7] and is included here to investigate the cross-language comparison with new data from the English-speaking listeners.

2. Experiment 1: Vowel F0

Experiment 1 was designed to investigate the contribution of the F0 contour to English and Mandarin sentence intelligibility. Vowel contributions were isolated by interrupting sentences to preserve primarily vowel cues with F0 contours flattened to the mean sentence level. Consonant segments were replaced with a low-level speech-shaped noise.

2.1. Listeners

Two groups of listeners participated in Experiment 1. The first group of listeners (N=18) consisted of native speakers of American English who were tested with the English sentences. Testing for this group was completed at the University of South Carolina. The second group of listeners (N=20) were native speakers of Mandarin Chinese and were tested with the Mandarin sentences. Testing for this group was completed at the University of Hong Kong. All listeners had normal audiograms with octave pure tone thresholds ≤ 20 dB HL.

http://dx.doi.org/10.21437/Interspeech.2016-28

2.2. Stimuli

English sentences were selected from the Hearing in Noise Test (HINT) [8]. Vowel boundaries were identified using FAVE [9], an automatic vowel alignment and extraction program, followed by manual verification by two trained phoneticians. Mandarin sentences were selected from the Mandarin version of the Hearing in Noise Test (MHINT) [10]. These sentences were previously coded for segmental boundaries [2]. The two sentence corpora were chosen because of similar semantic and syntactic complexity and were each spoken by a single talker. Ten sentences were presented in each condition with an average of 53 total words. All stimuli were presented at with a sampling rate of 16 kHz.

2.3. Signal Processing

Sentences were first processed to flatten the F0 contour. This was accomplished by extracting the F0 contour and replacing this contour with the mean F0 value across the sentence. This re-synthesis was carried out in Praat [11] using the Pitch Synchronous Overlap and Add method (PSOLA). Sentences were further processed to only preserve vowel acoustic information by deleting and replacing consonant segments with a low-level speech-shaped noise at 16 dB signal-to-noise ratio (SNR) based on the long-term average of the sentence corpus. Vowel boundaries marked by the sentence corpora were adjusted to within 1-ms of the nearest local minima (i.e., zero-crossing). To control for differences in the total duration of speech presented between English and Mandarin presentations, vowel boundaries were adjusted at the beginning and ending of each vowel by various proportions of the vowel duration. Boundaries for Mandarin and English were shifted in 10% increments. That is, the boundary at the start and end of the vowel were both adjusted inward (or outward) by 10% of the vowel duration. Overlapping conditions between the two languages resulted in average total sentence proportions of 65%, 53%, 41%, and 26% for English and Mandarin. As vowels in Mandarin account for larger proportions of the sentence duration compared to English vowels (66% versus 41%, respectively, for these sentence materials), longer sentence proportions for English contain additional consonantal information. Thus, these conditions should be viewed as a reflection of the total speech information provided, centered primarily on the vocalic cues within the sentence. Conditions were also examined that equated the total preserved proportion of individual vowels at 100%, 80%, and 60%. For Mandarin sentences, these preserved vowel proportions corresponded to 65%, 53%, and 41% sentence proportion conditions respectively; while for English sentences the corresponding sentence proportions were 41%, 33% (a new condition), and 26%.

2.4. Procedures

Listeners were seated in a sound attenuating booth and listened to sentences presented over circumaural headphones presented at a comfortable listening level (~70 dB SPL). Listeners first completed a practice session to familiarize them with the stimulus conditions and the task. No feedback was provided. Sentence conditions were randomized across listeners. Participants were allowed to listen to each sentence up to three times and were instructed to repeat aloud all of the words in the sentence. Responses were scored by trained raters according to a strict scoring procedure (e.g., no missing or extra suffixes). The proportion of words correct for each condition was transformed into rationalized arcsine units (RAU) to stabilize the error variance.

2.5. Results & Discussion

A mixed model analysis of variance (ANOVA) was conducted with group as the between subjects variable and sentence proportion as a within subjects variable. Results demonstrated a main effect of sentence proportion, $F(3,108) = 297.8$, $p<.001$. No significant main effect of language group was observed. However, there was a significant interaction following Greenhouse-Geisser correction, $F(3,108) = 4.4$, $p<.05$. This interaction occurred due to better English performance at the smallest sentence proportion, and therefore shortest preserved vowel duration, $t(36) = 2.2$, $p<.05$.

Results for this Experiment are plotted in Figure 1. For comparison, the dotted lines are provided which indicate average performance by different groups of listeners for Mandarin [2] and English sentences with full spectrum vowels for these same sentence materials. As indicated in the figure, previous data indicate substantial differences in overall intelligibility for Mandarin and English sentences that preserve predominantly vocalic information. Performance is significantly better with Mandarin speech, even though English materials at the longest sentence proportions contain additional information from the consonants. In sharp contrast to these earlier results, the current study with flattened F0 contours demonstrates similar performance for Mandarin and English sentences. These results strongly suggest that previous language differences are likely the result of dynamic F0 information present during vowels that provides additional acoustic information to aid in lexical processing for Mandarin, but not English, materials. This contrasts with other work that has demonstrated no difference in Mandarin performance for flattened F0 sentences in quiet, but markedly poorer performance in babble [12-13]. Thus, additional consonantal acoustics appear essential for Mandarin listeners to compensate for the loss of lexical F0 cues. In English, greater consonantal cues may have been preserved within the vowel conditions, partially due to greater formant dynamics, examined in Experiment 2.

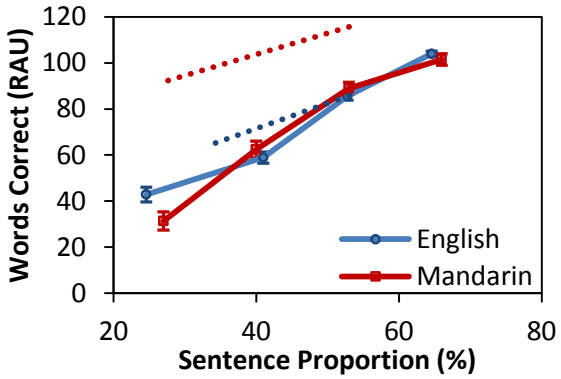

Figure 1. *English (blue) and Mandarin (red) intelligibility for F0 flattened sentences preserving primarily vocalic segments at four different proportions of the total sentence duration. Dotted lines indicate baseline performance for sentences with unprocessed vowels that preserved the natural F0 contour for a different group of participants tested on the same sentence materials. Error bars = standard error of the mean.*

Analysis for F0 flattened sentences was also conducted for conditions that matched the preserved proportion of individual vowels. From Figure 2 it is clear that intelligibility of Mandarin sentences is better than that for English sentences when individual vowels within the sentence are equally preserved. A mixed model ANOVA confirmed main effects of group [$F(1,36) = 77.4$, $p<.001$] and vowel proportion [$F(2,72) = 92.9$, $p<.001$] as well as a significant interaction [$F(2,72) = 1271.4$, $p<.001$]. The earlier analysis clearly demonstrated that lexical tone appears to account for main differences in intelligibility associated with the vowels of Mandarin and English sentences. This analysis demonstrates that listeners still obtain higher sentence recognition performance levels for Mandarin sentences when individual vowels between the two languages are equally preserved. This language difference is largely accounted for by the higher proportion of the sentence occupied by Mandarin vowels compared to English vowels.

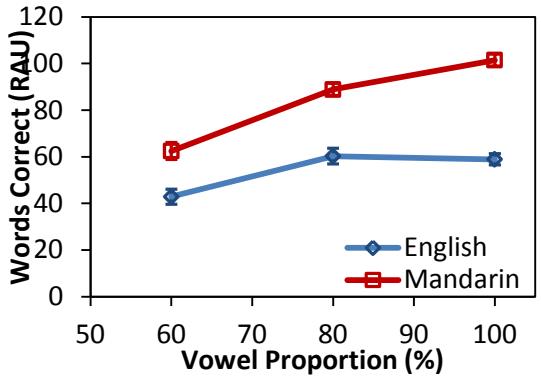

Figure 2. *English (blue) and Mandarin (red) intelligibility for F0 flattened sentences preserving primarily vocalic segments at three conditions that equated the average preserved proportion of individual vowels within the sentence. Error bars = standard error of the mean.*

3. Experiment 2: Vowel Formants

Experiment 2 was designed to investigate the contribution of the vowel formants to the intelligibility of English and Mandarin sentences. This was accomplished by coarsely coding the speech to preserve only the amplitude and frequency variation of the first three formants via sinewave synthesis.

3.1. Methods

The same listeners from Experiment 1 participated in the experimental conditions for Experiment 2. The same sentence corpora from Experiment 1 were used in this experiment. No sentences were repeated between the two experiments.

Sentences were first processed in Praat using sinewave speech synthesis scripts provided by Chris Darwin [14]. This algorithm estimates the formant frequencies using LPC. Formant amplitudes are then picked from a wideband FFT spectrum. Following sinewave synthesis, sentences were interrupted as in Experiment 1 to replace target consonant intervals with low-level speech-shaped noise (16 dB SNR). The same sentence proportions tested in Experiment 1 were also examined here.

Listeners were tested according to the procedures outlined in Experiment 1. They were seated in a sound attenuating booth and listened to sentences over circumaural headphones presented at a comfortable listening level. Listeners first completed a practice session to familiarize them with the stimulus conditions and the task. No feedback was provided. Sentence conditions were randomized across listeners. Participants were again able to listen to each sentence up to three times and were instructed to repeat all of the words in the sentence.

3.2. Results & Discussion

Figure 3 displays the results obtained for English and Mandarin sinewave speech plotted according to the sentence proportion preserved. Surprisingly, as can be readily observed from the figure, English listeners performed better than Mandarin listeners across most sentence proportions tested. This demonstrates a clear difference in performance between the two languages than that obtained with natural, full-spectrum vowels (see dotted lines in Figure 3).

These results were quantified using a mixed model ANOVA with group as the between subjects variable and sentence proportion as a within subjects variable. Results demonstrated a main effect of sentence proportion [$F(3,108) = 57.4$, $p<.001$] and of group [$F(1,36) = 30.1$, $p <.001$]. A significant interaction was also observed [$F(3, 108) = 11.9$, $p<.001$]. Independent samples t-tests were conducted between the two language groups at each of the four sentence proportions to define the interaction. Results indicated significant differences between the two groups at all proportions ($p<.001$), except at the condition that preserved the greatest amount of speech information: 66% ($p>.05$).

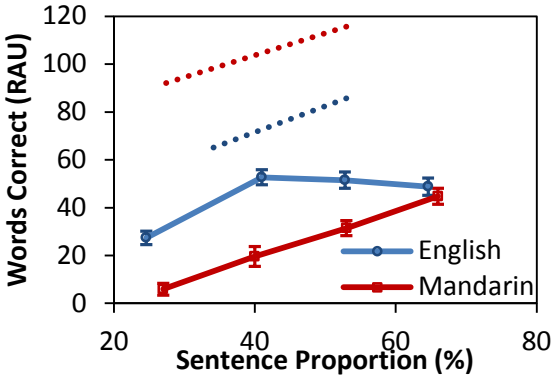

Figure 3. *English (blue) and Mandarin (red) intelligibility for sentences processed using sinewave speech synthesis. Sentences preserved primarily vocalic segments at four different proportions of the total sentence duration. Dotted lines indicate baseline performance for sentences with unprocessed, full-spectrum vowels for a different group of participants tested on the same sentence materials. Error bars = standard error of the mean.*

From Figure 3 it is also clear that there are two very different types of trends for the two languages across the sentence proportions tested. While Mandarin performance increased linearly, English performance appears to reach asymptotic performance near 51 RAU. This dissimilarity is likely due to the different distribution of consonants and vowels in the two languages. Mandarin is heavily dominated by vowel segments, which account for 66% of the total sentence

duration. In contrast, the numerous consonant clusters that occur in English result in less of the total sentence duration provided by vowel segments (41% for the sentence materials tested here). In order to equalize sentence proportions between the two languages, some consonant information was added to the English vowels at the longer sentence proportions tested. When Figure 3 is examined in this context, the asymptotic portion of the English function is explained by these conditions that increase the sentence proportion by increasing the preservation of consonant intervals. As can be observed, for sinewave speech, these preserved portions of neighboring consonant segments provide little-to-no additional information for speech intelligibility based on the formants. This occurs even though some of these neighboring segments are likely semivowels that would have clear formant structure.

At the shorter sentence proportions tested that only preserved the traditionally defined vowel, significant advantages are observed for English over Mandarin materials. This language difference may be due to several reasons. First, the sinewave speech synthesis conditions tested here preserved vowel formant frequencies, but did not present vowel F0 contours that are known to be an important component of Mandarin vowel acoustics. However, when F0 contour information was removed in Experiment 1, the two language groups performed similarly across equal sentence proportions (Figure 1). Thus, the better performance for the English listeners in the current experiment is not likely due to the removal of F0 information alone. Another difference between vowel acoustics of the two languages, as outlined in the introduction, is differences in VISC. The Mandarin vowel system is more sparse compared to English, and therefore allows greater formant variability within individual vowels. This is accounted for by having larger VISC distance as defined by larger differences in F1 x F2 vowel space from 20% to 80% of the vowel duration [5]. At smaller sentence proportions (i.e., 25% and 41%), individual Mandarin vowels would have been reduced more than individual English vowels in order to equate the total duration of the sentence. This could have resulted in poorer access to VISC information from these truncated vowels. Therefore, a second analysis was conducted that compared Mandarin and English conditions at equal levels of vowel preservation. Results of this analysis are displayed in Figure 4. The mixed model ANOVA demonstrated a significant main effect of vowel proportion [$F(2, 72) = 52.6$, $p<.001$]. A clear trend toward better performance across vowel proportions is observed for English compared to Mandarin sentences, and was demonstrated by a marginally significant main effect of group with a medium effect size [$F(1,36) = 4.0$, $p = .05$, $\eta^2 = .10$]. There was no significant interaction.

The results of this analysis are surprising given that when conditions between the languages are equated for the preserved proportion of individual vowels, total sentence duration is better preserved for Mandarin. This underscores the significant role VISC plays in English for sentence intelligibility as listeners still performed better than Mandarin listeners when compared at equal vowel proportions. Furthermore, in addition to VISC, English sentence performance may also be augmented by the preserved amplitude modulation of the vowels, which sinewave speech synthesis also preserves to some degree. A number of previous studies have now documented an essential contribution of amplitude modulations from vowel segments to sentence intelligibility that is observed even in the absence of vowel formant information [3, 15-16]. Thus, the relative contribution of vowel frequency and amplitude modulation of the formants still needs to be characterized between English and Mandarin to more comprehensively explain the acoustics behind this observed effect.

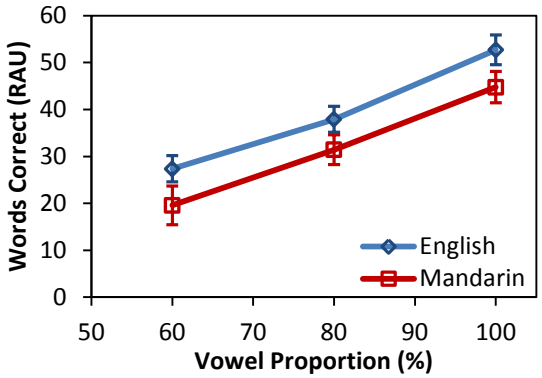

Figure 4. *Performance for Mandarin and English sinewave speech sentences at equal preserved vowel proportions. Error bars = standard error of the mean.*

4. Summary and Conclusions

The results from Experiments 1 and 2 demonstrate that there are significant differences between Mandarin and English languages regarding how the acoustic properties specified by vowels contribute to sentence intelligibility. The results of Experiment 1 suggest that vowel F0 contour information plays a significant role for Mandarin, likely due to the tonal structure of the language where F0 contour information provides direct lexical information [17-18]. When examining performance at equal proportions of the sentence, the prior observed effect of superior Mandarin intelligibility based primarily on vowel acoustics is absent, with performance largely equated between the two languages. When comparing the two languages at equal vowel proportions, a clear advantage for Mandarin is observed, likely due to greater preservation of the entire sentence. In sharp contrast, performance in Experiment 2 was marked by significantly better performance for English compared to Mandarin sinewave speech. This effect was observed when comparing performance across both equal preservation of the sentence duration and of the vowel duration. The latter is further remarkable given the significantly lower preservation of the sentence duration for English. The results of Experiment 2 indicate that frequency and amplitude variation of the vowel formants are differentially important cues for English sentence intelligibility. This may reflect the reduced vowel contrasts necessary in Mandarin due to the comparatively sparse vowel system. Overall, these results suggest that dynamic F0 cues are more important in Mandarin and dynamic formant cues (i.e., frequency and/or amplitude) are comparatively more important for English sentence intelligibility. These results suggest that speech processing technologies may result in greater intelligibility from focusing on language-specific acoustic characteristics, particularly when only partial speech information is available.

5. Acknowledgement

This work was supported, in part, by a grant from the National Institutes of Health, NIDCD R03-DC012506 (D.F.).

6. References

[1] F. Chen, L.L. Wong, and E.Y.W. Wong, "Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. EL178–EL185, 2013.

[2] D. Fogerty and F. Chen, "Vowel spectral contributions to English and Mandarin sentence intelligibility," in *Proceedings of 15th Annual Conference of the International Speech Communication Association (InterSpeech)*, Singapore, 2014, pp. 499–503, 2014.

[3] D. Fogerty, "Acoustic predictors of intelligibility for segmentally interrupted speech: Temporal envelope, voicing, and duration," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1402–1408, 2013.

[4] J.S. Laures, and G. Weismer, "The effects of a flattened fundamental frequency on intelligibility at the sentence level," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 5, pp. 1148–1156, 1999.

[5] S.H. Jin and C. Liu, "The vowel inherent spectral change of English vowels spoken by native and non-native speakers," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. EL363–EL369, 2013.

[6] T. M. Nearey and P. F. Assmann, "Modeling the role of vowel inherent spectral change in vowel identification," *The Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1297–1308, 1986.

[7] F. Chen, S.W.K. Wong, and L.L. Wong, "Effect of spectral degradation to the intelligibility of vowel sentences," in *Proceedings of 15th Annual Conference of the International Speech Communication Association (InterSpeech)*, Singapore, 2014, pp. 2002–2005, 2014.

[8] M. Nilsson, S.D. Soli, and J.A. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 1085–99. 1994.

[9] I. Rosenfelder, J. Fruehwald, K. Evanini, and Y. Jiahong *FAVE (Forced Alignment and Vowel Extraction) Program Suite,* http://fave.ling.upenn.edu, 2001.

[10] L.L. Wong, S.D. Soli, S. Liu, N. Han, and M.W. Huang, "Development of the Mandarin hearing in noise test (MHINT)," *Ear and hearing*, vol. *28*, no. 2, pp. 70S–74S. 2007.

[11] P. Boersma, and D. Weenink, "Praat: doing phonetics by computer" [Computer program], Version 5.3.80, retrieved 29 February 2014 from http://www.praat.org/, 2014.

[12] A.D. Patel, Y. Xu, and B. Wang, "The role of F0 variation in the intelligibility of Mandarin sentences," in *Proc. Speech Prosody* Chicago, IL, 2010.

[13] J. Wang, H. Shu, L. Zhang, Z. Liu, and Y. Zhang, "The roles of fundamental frequency contours and sentence context in Mandarin Chinese speech intelligibility. *The Journal of the Acoustical Society of America,* vol. 134, no. 1, EL91-EL97, 2013.

[14] http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/SWS

[15] D. Fogerty, "Importance of envelope modulations during consonants and vowels in segmentally interrupted sentences," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1568–1576, 2014.

[16] D. Fogerty, "Indexical properties influence time-varying amplitude and fundamental frequency contributions of vowels to sentence intelligibility," *Journal of Phonetics*, vol. 52, pp. 89–104, 2015.

[17] J.M. Howie, *Acoustical studies of Mandarin vowels and tones*. Cambridge: Cambridge University Press, 1976.

[18] D. H. Whalen, and Y. Xu, "Information for Mandarin tones in the amplitude contour and in brief segments," *Phonetica*, vol. 49, no.1, pp. 25–47, 1992.