Vowel Spectral 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 Division of Speech and Hearing Sciences, The University of Hong Kong, China

Abstract

The current study investigated the spectral distribution of vowel cues to Mandarin and English sentence intelligibility. Sentences were segmentally interrupted to preserve various amounts of vowel information. Interruption parameters ensured that similar durations of speech were presented across the two languages. The remaining vowel cues were then either high-pass or low-pass filtered. Results demonstrated significant contributions of information present during vowels to sentence intelligibility. Performance was much higher in most of the conditions for Mandarin speech. In addition, Mandarin vowels contained a higher distribution of cues in the low-frequencies, below 1000 Hz. Results suggest that Mandarin vowels carry additional cues in the low-frequencies, most likely related to lexical tone. This is further supported by the highly similar performance between Mandarin and English sentence intelligibility when spectral cues are confined to frequencies above 2000 Hz.

Index Terms: vowels, lexical tone, interruption

1. Introduction

The partial masking of speech by background noise is relatively common in everyday listening environments. Under such conditions, speech recognition may have to rely primarily on brief spectro-temporal fragments, or glimpses, of speech that remain relatively preserved at favorable signal-to-noise ratios (SNR) [1]. Previous studies have demonstrated that the proportion of the total speech duration preserved cumulatively across these speech glimpses is the primary predictor of speech intelligibility [e.g., 2]. However, it is likely that not all fragments of speech contribute equally to intelligibility. One predominant example of this is observed through the investigation of vowel cues to sentence intelligibility. While in English, vowels preserve less of the total sentence duration than consonants (45% vs. 55%), listeners perform with at least a 2:1 advantage for glimpses constrained to vowels compared to those of consonants [3]. In an account of this vowel advantage, Fogerty recently investigated acoustic explanations to this spectral distribution. In an account of this vowel advantage, vowel syllable structure that varies significantly from the complex syllable structure of English that allows for consonant clusters. This syllable structure results in at least two additional vowel differences (in addition to lexical F0) between Mandarin and English sentences. First, Mandarin sentences result in a larger proportion of the sentence duration presented during vowel segments, 66%, compared to vowels only contributing to 45% of the total sentence duration in English. Second, the statistical probability of accurately identifying a syllable based on the vowel is much higher in Mandarin than it is in English because of the higher number of syllable mutations that can be created in English from the use of consonant clusters [5]. This study investigated these vowels differences between Mandarin and English by examining the spectral distribution of vowels. The findings in the present study will shed light on the language difference in vowel contribution for speech perception, and guide the development of language-specific speech processing (e.g., speech enhancement) algorithms and assistive hearing devices (e.g., hearing aids).

2. Experiment 1: English Sentences

Experiment 1 was designed to investigate the contribution of low and high frequency vowel information to English sentence intelligibility. Sentences were interrupted to preserve primarily vowel cues in the low or high frequencies through low and high pass filtering.

2.1. Listeners

Two groups of eight listeners (M = 21 yrs, 18-30 yrs., 11 female) participated in Experiment 1. All listeners were native speakers of American English and had normal audiograms with octave pure tone thresholds ≤ 20 dB HL. The two groups of listeners were assigned to either a group that received high-pass filtered stimuli or a group that received low-pass filtered stimuli. Testing was completed at the University of South Carolina. One listener in the low-pass filter group failed to complete testing at the longest sentence proportion tested.

2.2. Stimuli

Stimuli consisted of sentences selected from the TIMIT database [6]. Sentences were divided into 10 lists of 18 sentences each according to PRESTO, a test for assessing speech perception with high-variability linguistic and indexical properties using TIMIT sentences [7]. Sentence lists were controlled for keyword frequency, familiarity, and talker gender. No sentences were repeated in any condition. TIMIT sentences have been used in several studies using segmentally interrupted
speech to selectively preserve vowel segments [e.g., 3] as the database provides time markings for segmental boundaries that were confirmed by expert phoneticians.

2.3. Signal Processing

Sentences were processed to segmentally interrupt speech using the segmental boundaries identified in the TIMIT database and adjusted within 1-ms to the nearest local minima (i.e., zero-crossing). As previously conducted [3], boundaries were adjusted to control the duration of speech preserved. All listeners received three different proportion conditions that deleted a portion of the preserved vowel or added a portion of the neighboring consonants by 20% of the duration of the added or deleted segment. This proportional shift is cumulative across the segment, resulting from a 10% shift at the beginning and ending of the segment. Combined with the original vowel duration, three proportions of the total sentence duration were tested (27%, 45%, and 67%), where 45% represents the proportion preserved using the original vowel duration. Larger proportions indicate the addition of neighboring consonant acoustics, while proportions smaller than 45% indicate reduced vowel durations.

Consistent with previous studies using this method [3], sentences were interrupted to preserve the segment durations, but replace the neighboring segments with a low-level speech shaped noise (16 dB SNR) based on the long-term average spectrum of the sentence corpus.

Finally, sentences were processed to limit speech cues to within a certain frequency range. The low-pass filtered group was tested on wideband (0-8000 Hz) stimuli and with a low-pass filter cutoff of 1000 Hz. The high-pass filtered group completed testing using three different cutoff frequencies at 500 Hz, 1000 Hz, and 2000 Hz. Stimulus filtering was completed in MATLAB using a linear-phase FIR filter, where the filter order, \( n \), was defined as:

\[
n = 2 \times 10 \times \frac{f_s}{2 \times c_f},
\]

where \( f_s \) is the sampling rate and \( c_f \) is the cutoff frequency.

2.4. Procedures

All testing was completed in a sound attenuating booth. The participants listened to stimuli presented at a sampling rate of 16 kHz via Sennheiser HD 280 PRO headphones. The level of the speech (prior to replacement) was calibrated to be presented at 70 dB SPL. All sentence materials were presented to listeners in blocks of 18 sentences (76 keywords) according to the test condition. Sentences within each condition were fully randomized and the order of blocks across participants was also randomized. Participants first completed demonstration trials to familiarize them with the stimulus processing and the task prior to completing the experimental conditions.

Listeners were required to repeat each sentence and were encouraged to guess. Sentence presentation was self-paced and a button on the custom interface provided listeners with the opportunity to listen to each sentence up to three times in a row before responding. Listener responses were audio recorded for offline analysis by trained raters who scored the correct keywords repeated. Preliminary testing indicated that keyword and all-word scoring methods resulted in highly similar measures of performance. Keywords were selected for analysis here as PRESTO lists were designed to balance lexical properties of the keywords. Percent correct scores were transformed to rationalized arcsine units (RAU) to stabilize the error variance.

2.5. Results & Discussion

Two ANOVAs were run separately for the low- and high-pass groups (see Figure 1). For the low-pass group, main effects of sentence proportion \([F(2,12) = 250.3, p < .001]\) and cutoff frequency \([F(1,6) = 410.5, p < .001]\) were significant, as well as the interaction \([F(2,12) = 14.1, p < .001]\). A significant effect of filtering was observed at all sentence proportions.

For the high-pass group, results also demonstrated a significant main effect of proportion \([F(2,14) = 249.5, p < .001]\) and cutoff frequency \([F(2,14) = 134.4, p < .001]\), and a significant interaction \([F(4,28) = 3.7, p < .05]\). Contrasts demonstrated a significant difference between 500 Hz and 1000 Hz high-pass filtering only at 27% of the sentence preserved, \(F(1,7) = 4.2, p < .05\), following Bonferroni correction for multiple comparisons.

![Figure 1. English intelligibility for high-pass and low-pass filtered sentences preserving only the vowels at three different proportions. The solid black line indicates maximum performance for the full bandwidth of the speech materials (0-8000 Hz). Open symbols denote high-pass filter conditions. Closed symbols denote low-pass conditions.](image1)

![Figure 2. Cross-over functions for high-pass and low-pass filtered English sentences.](image2)
To examine the difference in the distribution of vowel spectral cues, high- and low-pass groups were plotted in Figure 2 according to the filter cutoff frequency. Inspection of this plot suggests a cross-over frequency within the range of about 2000 Hz, although an exact crossover frequency cannot be calculated due to the sparseness of sampling the cutoff frequencies and possible nonlinearity. However, given the significantly higher performance of the high-pass filter at 1000 Hz, it is clear that the crossover frequency must occur above this frequency range. Thus, the spectral contributions of vowels to English sentence intelligibility appear to be equally divided by low-frequency cues under 2000 Hz, which includes F1 cues, and high frequencies above this cutoff frequency.

3. Experiment 2: Mandarin Sentences

Experiment 2 was designed to investigate the contribution of low and high frequency vowel information to Mandarin sentence intelligibility.

3.1. Listeners

Two groups of eight listeners (M = 22 yrs, 19-25 yrs., 6 female) participated in Experiment 2. All listeners were native speakers of Mandarin Chinese and had normal hearing. The two groups of listeners were assigned to either a group that received high-pass filtered stimuli or a group that received low-pass filtered stimuli. Testing was completed at the University of Hong Kong.

3.2. Stimuli

Sentences were selected from the Mandarin version of the Hearing in Noise Test (MHINT) corpus [8]. Vowel-consonant boundaries (see [5] for more on vowel and consonant classification) for MHINT sentences were labeled manually by an experienced phonetician, and later verified by another experienced phonetician.

3.3. Signal Processing

Signal processing was completed according to the procedures outlined in Experiment 1, with the following exceptions. First, vowels in Mandarin account for 66% of the total sentence duration, which is close to the largest sentence duration (67%) tested in Experiment 1 with English sentences. Thus, all proportion conditions reduced the vowel duration by 40% or 60% of the total vowel duration. This resulted in three sentence durations, which is close to the largest sentence duration (67%) tested in Experiment 1 with English sentences. Segments were replaced with a speech-shaped noise scaled to 16 dB SNR, as in Experiment 1.

Sentences were also high- and low-pass filtered as in Experiment 1. The low-pass filter group completed testing with wideband stimuli, as well as with cutoff frequencies at 1000 Hz and 500 Hz (the later not tested with English listeners due to floor performance). The high-pass filter group completed testing using the same cutoff frequencies as Experiment 1: 500 Hz, 1000 Hz, and 2000 Hz.

3.4. Procedures

Listeners were tested according to the procedures outlined in Experiment 1. They were seated in a sound attenuating booth and listened to sentences presented 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. 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.

3.5. Results & Discussion

Figure 3 displays the results obtained for Mandarin plotted according to the sentence proportion preserved. Two ANOVAs were run separately for the low- and high-pass groups. For the low-pass group, main effects of sentence proportion [F(2,14) = 389.9, p < .001] and cutoff frequency [F(2,14) = 70.5, p < .001] were significant, as well as the interaction [F(4,28) = 15.4, p < .001]. A significant effect of filtering was observed at all sentence proportions.

Consistent with previous work [5], results for the wideband 8000 Hz low-pass filter demonstrate highly preserved intelligibility despite significant reductions in the proportion of speech preserved (by 40% of the overall sentence duration). From 67% to 27% of the sentence duration preserved, intelligibility only decreased by an average of 22 RAU, compared to the 70 RAU for English sentences. This highlights the highly robust cues carried by Mandarin vowels.

It is also worthy to note that no additional improvement in intelligibility was gained in the 500 Hz low-pass filter condition between 40% and 67% (p > .05). This may suggest that vowel cues in this low-frequency band may be mostly present in vowel center, and that 20% at the beginning and end of the vowel contains either redundant or less informative/reliable cues.

For the high-pass group, results also demonstrated a significant main effect of proportion [F(2,14) = 124.5, p < .001] and cutoff frequency [F(2,14) = 242.3, p < .001], but no significant interaction [F(4,28) = 2.3, p > .05]. Contrasts demonstrated a significant difference between 500 Hz and 1000 Hz high-pass filtering only at 40% of the sentence preserved, t(7) = 3.3, p < .05, following Bonferroni correction for multiple comparisons. This may suggest that when high frequencies are availa-
ble, vowel frequencies between 500 and 1000 Hz only appear to make a marginal contribution in addition to the cues already provided below 500 Hz. However, for the low-pass conditions, there was a very large difference between these two cutoff frequencies, suggesting a significant role of 500-1000 Hz band when only low-frequency vowel cues are available. These contrasting results suggest contributions to sentence intelligibility from the Mandarin vowels spectrum may be equally divided at around 1000 Hz. To further investigate this possibility, comparison between the low- and high-pass groups was conducted according to the cutoff frequency, and is displayed in Figure 4. Independent samples t-tests between high- and low-pass filter groups at each proportion for the 1000 Hz cutoff frequency demonstrated no significant differences between conditions at this cutoff frequency (p > .05). Results are highly consistent across all sentence proportions, confirming an approximately equal division of vowel sentence cues at a cross-over frequency near 1000 Hz.

As noted in the introduction, English and Mandarin are different in three important ways: (1) greater duration of the sentence is provided by Mandarin vowels, (2) lexical F0 cues are present in Mandarin, and (3) greater statistical cues for syllable identification are available in Mandarin due to the constrained syllable structure. The methods employed in the current study specifically controlled for (1). In addition, the similarity in performance for this one condition (i.e., 2000 Hz high-pass filtering), demonstrates that the statistical advantage that may be available for Mandarin listeners did not play a role in contributing to better overall performance (at least in this one filtering condition). Therefore, results appear to suggest the greater importance of low-frequency cues (such as lexical F0) in Mandarin compared to English, particularly below 2000 Hz. Indeed, the lower cross-over frequency for Mandarin compared to English (~1000 Hz compared to ~2000 Hz respectively), further highlights a greater weighting of low-frequency information in Mandarin speech. The ultimate example of this is the low, but residual speech intelligibility for 500 Hz low-pass filtered Mandarin speech. In contrast, English listeners could not complete testing at any proportion for frequencies less than 500 Hz.

4. General Discussion

Interestingly, across all of the conditions tested, Mandarin conditions almost always outperformed the complementary English condition, even when examining performance at equal proportions of the sentence preserved. The only exception to this observation was for the 2000 Hz high-pass filtered conditions (see Figure 5). Here, English and Mandarin groups did not perform significantly different at any of the three sentence proportions (p > .05).

5. Conclusions

Overall, results demonstrate the overwhelming importance of vowel cues to Mandarin sentence intelligibility, with limited decrements in intelligibility for sentences interrupted to preserve primary vowel segments. In addition, Mandarin vowels appear to be weighted toward low-frequency cues (under 1000 Hz). This region primarily contains F0 and amplitude modulation cues, which are important for tone identification [9]. In contrast, English vowels, while still carrying important cues for sentence intelligibility, are much less effective than Mandarin vowels. Controlling for the proportion of speech preserved, the additional advantage in Mandarin appears to be due to the additional low-frequency lexical information. Spectral cues for English vowels appear to divide more evenly in higher frequencies around 2000 Hz. This appears to be due to reduced prominence of low-frequency information (i.e., no lexical information conveyed by tone) than to a greater importance of high-frequency vowel information in English. This result is supported by the nearly identical performance between English and Mandarin listeners for high-pass filtered speech at 2000 Hz.

These results support the consistent findings that vowels carry essential cues for sentence intelligibility, now for two diverse languages: English and Mandarin. Additional advantage is provided by low-frequency lexical information unique to Mandarin, and likely other tonal languages.

6. Acknowledgements

This work was supported, in part, by a grant from the National Institutes of Health, NIDCD R03-DC012506 (D.F.) and by Seed Funding for Basic Research, The University of Hong Kong (F.C.).
7. References


