Segmental Contribution to the Intelligibility of Ideal Binary-masked Sentences

Fei Chen¹, Alexander Siu Tai Kwok²

¹ Dept. of Electrical & Electronic Engineering, South Univ. of Science and Technology of China
² Division of Speech and Hearing Sciences, The University of Hong Kong

fchen@sustc.edu.cn, alexk724@hku.hk

Abstract

Many studies have shown the advantage of using ideal binary mask (IdBM) to improve the intelligibility of speech masked by interfering voices. The present work further investigated the segmental contribution to the intelligibility of the IdBM-processed sentences. Three types (i.e., vowel-only, consonant-only, and vowel-consonant transition) of Mandarin IdBM-processed stimuli were generated by using a noise-replacement paradigm to preserve the selected segments and replace the rest with noise. Normal-hearing subjects participated in listening experiments to recognize the IdBM-processed sentences. Experiment results showed that the recognition score of the vowel-only IdBM-processed sentences was significantly higher than that of the consonant-only IdBM-processed sentences, suggesting a greater contribution of vowels than consonants to the intelligibility of the IdBM-processed sentences. Vowel centers contained rich intelligibility information, while consonants and vowel-onset carried little intelligibility information, adding a small proportion of vowel-consonant transition significantly improved the recognition score of the consonant-only and vowel-onset IdBM-processed sentences.

Index Terms: Speech intelligibility, segmental contribution, ideal binary masking.

1. Introduction

Recent studies using noise-replacement paradigm showed that vowel-only (consonants replaced by noise) sentences have a 2:1 sentence intelligibility advantage against consonant-only (vowels replaced by noise) sentences in English [1]. Similarly, this vowel to consonant advantage in sentence intelligibility was also found in Mandarin, even having an advantage higher than 2:1 [2]. This more remarkable advantage in Mandarin than in English is probably due to three reasons. First, the lexical tone information in Mandarin is probably conveyed by vowel segments [3-4]. Since Mandarin is a tonal language while English is not, the greater importance of Mandarin vowels in sentence intelligibility than English vowels is because of the additional role of carrying lexical tone. Second, there are more vowels than consonants in Mandarin phonemic inventory. In Mandarin, there are 35 vowels but only 21 consonants [5-6], while in English, there are 20 vowels but 32 consonants [e.g., 4]. Having more vowels than consonants in Mandarin, Mandarin vowels seem to have a greater importance in phonemic contrast than English vowels. Third, vowel durations are predominantly longer than consonant durations in Mandarin sentences. Vowels occupied 66.3% of the sentence duration but consonants only occupied 25.9% of the sentence duration in Mandarin [2]. However, in English vocalic intervals are smaller than consonantal interval with around 10% [4]. This difference between vowel and consonant durations in Mandarin also accounts for the large contribution to sentence intelligibility by Mandarin vowels.

Although there are many studies showing the segmental contribution to the intelligibility of clean and intact sentences (i.e., speech without processing) [e.g., 2, 4], there is a lack of to date study on the segmental intelligibility contribution in processed [e.g., by speech enhancement, or ideal binary mask (IdBM) processing in this study] speech. Since listeners may employ a different set of acoustic cues that help them to understand the sentences under clean, noisy and noise-suppressed listening situations [7], there is a need to investigate whether the segmental contribution in clean and intact speech is also present in processed (e.g., IdBM-processed) sentences.

Many studies indicated that ideal binary mask processing, which is a state-of-the-art speech enhancement technique, could improve the intelligibility of speech under interfering noises [7-11]. IdBM uses the time-frequency (T-F) representation of a mixture signal and then compares the local ideal (presumably available from clean and noise signals) signal-to-noise-ratio (SNR) of each T-F units with a present threshold (e.g., 0 dB SNR). Afterwards, T-F units that have a SNR lower than the threshold are eliminated, while the others (with SNR larger than the threshold) are preserved [7]. However, so far little has been done to examine how IdBM processing improves the speech intelligibility at segmental level. As Mandarin and English have different features from each other as mentioned earlier, it is hypothesized that there should also be a notable difference of segmental contribution between Mandarin and English IdBM-processed speech.

Investigation in this study on how IdBM processing improves Mandarin speech intelligibility in noise at segmental level will provide information on this comparative study between Mandarin and English IdBM-processed speech. Since listeners may employ a different set of acoustic cues that help them to understand the sentences under clean, noisy and noise-suppressed listening situations [7], there is a need to investigate whether the segmental contribution in clean and intact speech is also present in processed (e.g., IdBM-processed) sentences.

The present work would investigate the contributions of three segments in a sentence, which are vowel (V) segments, consonant (C) segments and vowel-consonant (VC) transitional segments. It is hypothesized that the sentence intelligibility in the vowel-only sentences, which consist of only V-segments, would be much greater than that of the consonant-only sentences, which consist of only C-segments. VC transitional segments around C-V boundaries are expected to carry important acoustic information for improving the intelligibility of the IdBM-processed sentences.
2. Methods

2.1 Subjects and materials
Seventeen (18-29 yrs, and 8 female) normal-hearing (NH) native Mandarin speakers were recruited from the University of Hong Kong. Air-conduction pure tone assessment was conducted as a hearing screening to ensure that all participants had normal hearing ability and the results of this study were not affected by hearing loss of the participants.

The sentence materials were taken from the Mandarin version of the Hearing in Noise Test (MHINT) [12]. The MHINT database consisted of 24 lists of sentences and each list has 10 ten-syllable Mandarin sentences. All the sentences were produced by a male speaker, with fundamental frequency ranging from 75 to 180 Hz. A speech-shaped noise was used to corrupt test sentences at two SNR levels of -10 and -12 dB.

2.2 Signal processing
In this study, the synthesis of the IdBM-processed stimuli followed the process used in [7]. A 24-channel sinewave-excited vocoder was used to process the speech materials with the utilizing of IdBM processing. During the implementation of the sinewave-excited vocoder, signals were processed through a pre-emphasis filter (2000 Hz cutoff) first, with a 3 dB/octave rolloff, and then bandpass into 24 frequency bands using sixth-order Butterworth filters. Mel filter spacing was used to determine the cutoff frequencies of bandpass filters. Full-wave rectification and second-order Butterworth lowpass filtering (with a 400 Hz cutoff frequency) were used to extract the envelope of the band-passed signal. Afterwards, the masker signal was scaled to obtain the desired SNR level.

Then, the target and masker signals were bandpass filtered independently into the 24 channels, and the envelopes were extracted by low-pass filtering the full-wave rectified waveforms. The filtered target and masker signals were used to estimate the (true) instantaneous SNR in each channel by computing the ratio of the root-mean-square (RMS) energies of the target and masker envelope signals every 4 ms. If the SNR level in a given channel was found to be greater than threshold 0 dB, the mixture envelope of that channel was reserved; otherwise, the mixture envelope of that channel was eliminated. Following the reservation/elimination of the mixture envelopes in each channel, the signal was synthesized as a sum of sinewaves with amplitudes set to the RMS energy of the envelopes with positive SNR values and frequencies set to the center frequencies equal to the center frequencies of the corresponding bandpass filters. The level of the synthesized speech segment was adjusted to have the same RMS value as that of the original speech segment.

After obtaining the IdBM-processed stimuli, three types of sentences were created.
1) Vowel-only sentences were created by preserving the targeted vowel segments and replacing the remaining segments of the sentence by noise [13].
2) Consonant-only sentences were created by preserving the targeted consonant segments and replacing the remaining segments of the sentence by noise.
3) Three types of VC transition sentences with different durations were created, including vowel onset portion only (or condition V-onset, with \( p = 20 \) and 40 %), consonant offset portion (pC) plus vowel onset portion (pV) (or condition pC+pV, with p = 20 and 40 %), and full consonant portion plus vowel onset portion (pV) (or condition C+pV, with p = 20 and 40 %).

Together with two controls (i.e., two noisy situations without IdBM processing) and two IdBM-processing conditions containing all vowel and consonant segments, there were 24 testing conditions in total.

2.3 Procedure
The listening experiments were conducted in a sound-proof booth. Stimuli were played to the participants through a circumaural headphone binaurally at a comfortable listening level. Forty noise-replaced IdBM-processed sentences were played to the participants as a practice session (i.e., with feedback) before the experiment to familiarize the participants with the experiment procedure. Participants participated in all the 24 testing conditions (\( \times 2 \) SNR levels \( \times 12 \) conditions). The order of the 24 testing conditions was randomized across participants. Participants were allowed to listen to the sentences 3 times at most, and then were instructed to verbally repeat all the words that they could recognize.
3. Results and Discussion

3.1. Contributions of vowels and consonants to recognition score
Mean recognition scores for the (unprocessed) noisy, IdBM-full, IdBM-V-only and IdBM-C-only sentences are shown in Fig. 2 (a). Statistical significance was determined by using the percent recognition score as the dependent variable, and SNR level and signal processing condition as the two within-subject factors. The scores were first converted to rational arcsine units (RAU) using the rationalized arcsine transform [15]. Two-way analysis of variance (ANOVA) with repeated measures indicated a significant effect ($F[1, 16]=49.89$, $p<.001$) of SNR level, signal processing condition ($F[3, 48]=470.24$, $p<.001$), and a significant interaction ($F[3, 48]=25.13$, $p<.001$) between SNR level and signal processing condition.

As expected, Fig. 2 (a) shows that the intelligibility of the IdBM-processed (i.e., condition IdBM-full) sentences is significantly higher than that of noisy sentences (i.e., condition Noisy), clearly showing the advantage of using IdBM processing for intelligibility improvement in noise. Furthermore, the score of the vowel-only IdBM-processed sentences (i.e., IdBM-V-only) is smaller than that of noisy sentences (i.e., 49.1% vs. 54.4%) at -10 dB SNR, but higher than that of noisy sentences (i.e., 42.2 % vs. 33.1 %) at -12 dB SNR. This accounts for the above-mentioned significant interaction between signal processing condition and SNR level at 2-way ANOVA analysis. This also suggests that at low SNR level (i.e., -12 dB), the vowel-only IdBM-processed sentences carry more intelligibility information than noisy sentences. However, statistical analysis also showed that the recognition score of the vowel-only IdBM-processed sentences is significantly lower than that of the IdBM-full sentences at -12 dB in Fig. 2 (a).

It is also seen in Fig. 2 (a) that the C-only IdBM-processed sentences carry little intelligibility information at the two SNR conditions. Hence, compared with the high recognition score of the V-only IdBM-processed sentences, a notable intelligibility advantage of vowels against consonants is found in this study. It is known that vowels feature their harmonic structure and large energy. Vowels segments are less affected by energy masking, and more masks labelled with ‘1’ may be obtained in vowel segments than in consonant segments. In this regard, it is not surprising that the V-only IdBM-processed sentences are more intelligible than the C-only IdBM-processed sentences. It is also noted that the intelligibility of the C-only IdBM-processed sentences is almost zero. Hence, there is not the 2:1 intelligibility ratio (i.e., observed from clean and intact speech [4]) between the V-only and C-only IdBM-processed sentences.

3.2. Contributions of vowel centers to recognition score
Mean recognition scores for the three types of the V-only IdBM-processed sentences are shown in Fig. 2 (b), where the duration of vowel centers is controlled by a proportion factor p. Statistical significance was determined by using the percent recognition score as the dependent variable, and SNR level and vowel center duration as the two within-subject factors. The scores were first converted to rational arcsine units (RAU) using the rationalized arcsine transform. Two-way ANOVA with repeated measures indicated a significant effect ($F[1, 16]=20.12$, $p<.001$) of SNR level, vowel center duration ($F[2, 32]=65.76$, $p<.001$), and a non-significant interaction ($F[2, 32]=0.95$, $p=.40$) between SNR level and vowel center duration.

It is observed in Fig. 2 (b) that, when compared with the full vowel condition [i.e., 100% vowel centers, or condition -V-center (100)], preserving 80% vowel segments at centers also carries a high level of intelligibility, i.e., 42.4% at -10 dB and 30.8% at -12 dB. This suggests that vowel centers of the IdBM-processed sentences carry rich intelligibility information. This is consistent with previous finding on the segmental contribution to unprocessed speech understanding [e.g., 14, 16]. Further reducing the duration (e.g., to 60%) of vowel centers leads to a low recognition score. This suggests that the duration of vowel centers might be a predictor of the intelligibility information contained in vowels, and this warrants further investigation in the future.

3.3. Contribution of vowel-consonant transitions to recognition score
Figure 2 (c) shows the recognition scores for the vowel-onset (i.e., -V-onset), vowel-consonant transition (i.e., -pC+pV) and consonant plus vowel onset (i.e., -C+pV) conditions. Statistical significance was determined by using the percent recognition score as the dependent variable, and SNR level, signal processing condition and proportion factor (i.e., p) as
the three within-subject factors. The scores were first converted to rational arcsine units (RAU) using the rationalized arcsine transform. Three-way ANOVA with repeated measures indicated a non-significant effect (\(F(1, 16)=3.51, p=.08\)) of SNR level, a significant effect (\(F(2, 32)=44.02, p<.001\)) of signal processing condition, proportion factor (\(F(1, 16)=289.08, p<.001\)), and a non-significant interaction (\(F(2, 32)=1.93, p=.16\)) between SNR level and signal processing condition, between SNR level and proportion factor (\(F(1, 16)=3.89, p=.07\)), and between signal processing condition and proportion factor (\(F(2, 32)=.63, p=.54\)), and a non-significant interaction (\(F(2, 32)=.48, p=.63\)) among SNR level, signal processing condition and proportion factor. The flooring effect of recognition scores at some conditions might account for the non-significant effect of SNR level at 3-way ANOVA analysis.

From Fig. 2 (e), it is seen that vowel onsets carry little intelligibility information. The recognition scores (at p=40 %) are 9.8% at -10 dB and 8.5% at -12 dB, respectively. However, when the same proportion (i.e., p=40 %) of consonant offsets are added to vowel onsets (forming vowel-consonant transition, or pC+pV condition), the recognition scores are improved to 17.8% at -10 dB and 12.5% at -12 dB, respectively. In addition, although the C-only IdBM-processed sentences carry little intelligibility information [see Fig. 2 (a)], adding a small proportion of vowel offsets significantly improves the recognition score of the C-only IdBM-processed sentences. For instance, the scores are improved to 7.9% at -10 dB and 6.1% at -12 dB, respectively, when 20% vowel onsets are combined with consonants. Note that both vowel-onset and consonant-only IdBM-processed sentences carry little intelligibility information [see Figs. 2 (a) and (c)]. However, integrating consonants and vowel onsets together leads to significant intelligibility improvement, which manifests the contribution of vowel-consonant transition for the intelligibility of the IdBM-processed sentences. This is consistent with results from many previous studies on the perceptual importance of vowel-consonant transition to speech intelligibility [e.g., 3-4].

4. Conclusions

The present work extended previous study to understand factors accounting for the intelligibility of the IdBM-processed sentences. More specifically, we assessed the segmental contributions by vowels, consonants and vowel-consonant transitions to the intelligibility of the IdBM-processed sentences. Consistent with previous work, vowels carried more intelligibility information than consonants, and vowel-consonant transitions played an important role in improving the intelligibilities of the vowel-onset and consonant-only IdBM-processed sentences. Different from previous work, the C-only IdBM-processed sentences carried little intelligibility information. This is understandable, as consonants are characterized with low energy and most consonant segments are marked by interfering noise. That is, the preserved marks labelled with ‘1’ would be dominated by those from vowel segments but not consonant segments.

However, the small amount of preserved marks (i.e., labelled with ‘1’) from consonant segments, though carrying little intelligibility contribution, are still important when understanding the IdBM-processed sentences, as the intelligibility of the V-only IdBM-processed sentences is significantly lower than that of the full IdBM-processed sentences. This suggests the presence of a high level top-down processing involved in recognizing the IdBM-processed sentences when presented segmentally. It is still unclear how those un-intelligible consonant segments contribute the intelligibility advantage of the IdBM-processed speech.

5. Acknowledgement

This study was the basis for the Bachelor’s dissertation of the second author (A.S.T.K.).

6. References