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

Relative Contributions of Amplitude and Phase to the Intelligibility Advantage of Ideal Binary Masked Sentences

Lei Wang, Shufeng Zhu, Diliang Chen, Yong Feng, Fei Chen

Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China

fchen@sustc.edu.cn

Abstract

Many studies have shown the advantage of using ideal binary masking (IdBM) to improve the intelligibility of speech corrupted by interfering maskers. Given the fact that amplitude and phase are two important acoustic cues for speech perception, the present work further investigated the relative contributions of these two cues to the intelligibility advantage of IdBM-processed sentences. Three types of Mandarin IdBM-processed stimuli (i.e., amplitude-only, phase-only, and amplitude-and-phase) were generated, and played to normal-hearing listeners to recognize. Experiment results showed that amplitude- or phase-only cue could lead to significantly improved intelligibility of IdBM-processed sentences in relative to noise-masked sentences. A masker-dependent amplitude over phase advantage was observed when accounting for their relative contributions to the intelligibility advantage of IdBM-processed sentences. Under steady-state speech-spectrum shaped noise, both amplitude- and phase-only IdBM-processed sentences contained intelligibility information close to that contained in amplitude-and-phase IdBM-processed sentences. In contrast, under competing babble masker, amplitude-only IdBM-processed sentences were more intelligible than phase-only IdBM-processed sentences, and neither could account for the intelligibility advantage of amplitude-and-phase IdBM-processed sentences.

Index Terms: Speech intelligibility, ideal binary masking, amplitude and phase.

1. Introduction

It is well known that amplitude and phase are two acoustic properties carrying important information for speech perception [1-4]. The relative importance of amplitude to speech recognition was extensively investigated through, for instance, envelope-based vocoder simulation studies, which preserve envelope cue and discard temporal fine-structure (TFS) or phase cue by replacing it with a sinusoidal or band-limited noise carrier [e.g., 5-6]. It was shown that using a relatively large number of channels and a high low-pass (LP) cut-off frequency to extract the envelope waveform favors better recognition of the envelope-based vocoded speech [e.g., 5-7]. Earlier studies also showed that phase was also important for speech perception. Gilbert and Lorenzi studied the intelligibility of temporal fine-structure stimuli, which is believed to contain primarily phase information. Their results showed that the TFS stimuli could also be highly intelligible, and the underlying mechanism for understanding TFS stimuli was attributed to the recovered envelope (i.e., amplitude fluctuation) in the auditory pathway [8]. Chen and Guan further studied the factors influencing the intelligibility of phase-based stimuli, and suggested that the temporal modulation rate of phase information affected the amount of intelligibility information contained in the phase-based stimuli [9]. Recent speech perception studies also suggested that phase information played an important role for tonal language understanding [e.g., 10-11] and speech recognition in noise [e.g., 12].

Many studies indicated that, as a state-of-the-art speech enhancement technique, ideal binary mask (IdBM) processing could significantly improve the intelligibility of speech under interfering noises [13-15]. IdBM uses the time-frequency (T-F) representation of a mixture signal (of clean speech and noise masker), and then compares the local ideal (presumably available from clean speech and noise masker) signal-to-noise ratio (SNR) of each T-F unit with a present threshold (e.g., 0 dB SNR). Afterwards, T-F units with SNR levels lower than the threshold are eliminated, while the others with SNR levels larger than the threshold are preserved [13]. Studies are still actively ongoing to understand the underlying mechanism accounting for the intelligibility advantage of IdBM-processed sentences [13-20]. For instance, Li and Loizou examined the effect of spectral resolution on the intelligibility of ideal binary masked speech, and showed that using ideal binary masks could bring substantial gains in speech intelligibility, particularly at low SNR levels (e.g., −5 and 0 dB), even when the spectral resolution was relatively low (16–24 channels) [18]. Chen and Kwok recently assessed the segmental contributions to the intelligibility of IdBM-processed speech, and found that the recognition score of vowel-only IdBM-processed sentences was significantly higher than that of consonant-only IdBM-processed sentences, suggesting a greater contribution of vowels than consonants to the intelligibility of IdBM-processed sentences [19].

However, so far little has been done to examine how amplitude and phase cues account for the intelligibility advantage of IdBM-processed speech. Many questions remain unclear, including 1) which cue (i.e., amplitude and/or phase) contained in those SNR-larger-than-threshold (or masker-‘1’) T-F units accounts for the intelligibility advantage of IdBM-processed sentences; 2) what is the relative perceptual importance of amplitude and phase cues if they both contribute the intelligibility advantage of IdBM-processed sentences; and 3) how masker signal interacts with amplitude and phase cues to affect the intelligibility of amplitude- and phase-only IdBM-processed sentences. The purpose of this study was to investigate the perceptual contributions of amplitude and phase cues to the intelligibility advantage of

http://dx.doi.org/10.21437/Interspeech.2016-18

IdBM-processed sentences. In addition to the traditional IdBM condition which uses both amplitude and phase information in masker-‘1’ T-F units, two additional IdBM conditions were generated to contain amplitude-only or phase-only information in masker-‘1’ T-F units. It was hypothesized that both amplitude and phase cues contributed to the intelligibility advantage of IdBM-processed sentences. However, since environmental noise has different influence to amplitude and phase cues, there might exist a difference between the intelligibility advantages of amplitude- and phase-only IdBM-processed sentences under different noise background.

2. Methods

2.1. Subjects and materials

Eight (18-29 yrs, and 4 female) normal-hearing (NH) native Mandarin speakers were recruited to attend this experiment. The sentence materials were taken from the Mandarin version of the Hearing in Noise Test (MHINT) [9]. The MHINT database consisted of 24 lists of sentences and each list had 10 ten-syllable Mandarin sentences. All the sentences were produced by a male speaker, with fundamental frequency ranging from 75 to 180 Hz. Steady-state speech-spectrum shaped noise (SSN) and two-talker babble (2-talker) were used to corrupt test sentences at two SNR levels of −5 and −10 dB, which were chosen by a pilot study to avoid the ceiling/floor effect in speech recognition.

2.2. Signal processing

In this study, the synthesis of IdBM-processed stimuli followed the procedure used in [18]. A 24-channel sinewave-excited vocoder was used to process the speech materials with the utilizing of IdBM processing. When implementing the sinewave-excited vocoder, signals were first processed through a pre-emphasis filter (with 2000 Hz cutoff and 3 dB/octave rolloff), and then bandpassed into 24 frequency bands using sixth-order band-pass Butterworth filters. Mel filter spacing was used to determine the cutoff frequencies of bandpass filters. Full-wave rectification and second-order Butterworth low-pass filtering (with a 400 Hz cutoff frequency) were used to extract the envelope waveform of the band-pass-filtered signal. Afterwards, the masker signal was scaled to obtain the desired SNR level. The target (clean) and masker signals were bandpass filtered independently into the 24 channels, and the envelope waveforms were extracted by low-pass filtering the full-wave rectified waveforms. The filtered target and masker signals were used to estimate the (true) instantaneous local 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 larger than a pre-set threshold 0 dB, the mixture envelope of that channel was reserved (i.e., noted as masker-‘1’); otherwise, the mixture envelope of that channel was eliminated (i.e., noted as masker-‘0’). 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, frequencies set to the center frequencies equal to those of the corresponding bandpass filters, and phase estimated from the fast Fourier transform of every 4 ms of non-overlapping speech frames [21].

The above-mentioned signal processing generated traditional IdBM-processed stimuli containing both amplitude and phase information in masker-‘1’ T-F units [i.e., condition IdBM(A+P)]. The present work also synthesized stimuli only containing amplitude or phase information in masker-‘1’ TF units, noted as condition IdBM(A) or IdBM(P), respectively. To synthesize the IdBM(A)-processed stimuli, the phase value was set to zero for the sinewaves in masker-‘1’ TF units; similarly, the amplitude value was set to one in masker-‘1’ TF units to generate the IdBM(P)-processed stimuli. Note that when synthesizing the IdBM(A+P)-, IdBM(A)- and IdBM(P)-processed stimuli, the sinusoids of each band were finally summed up, and the level of the synthesized speech was adjusted to have the same RMS level as the original speech.

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 IdBM(A+P)-processed sentences were played to the participants as a practice session (i.e., with feedback) before the experiment to familiarize them with the experiment procedure. Participants attended all the 16 tested conditions [= 2 maskers (i.e., SSN and 2-talker) × 2 SNR levels (i.e., −5 and −10 dB) × 4 signal processing conditions of Noisy, IdBM(A+P), IdBM(A) and IdBM(P)]. The order of the 16 tested 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. The intelligibility score for each tested condition was computed as the ratio between the number of the correctly recognized words and the total number of words contained in each list of 10 MHINT sentences.

3. Results

Figure 1 (a) shows the mean sentence recognition scores of all conditions with SSN masker. 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. Two-way analysis of variance (ANOVA) with repeated measures indicated a significant effect (*F*[1, 7]=101.7, *p*<0.001) of SNR level, signal processing condition (*F*[3,21]=53.0, *p*<0.001), and a significant interaction (*F*[3, 21]=14.5, *p*<0.001] between SNR level and signal processing condition.

Post hoc pairwise comparisons at each SNR level showed that recognition scores of conditions IdBM(A+P), IdBM(A) and IdBM(P) were all significantly (*p*s<0.01) larger than of condition Noisy, and there were no significant difference between paired conditions of IdBM(A+P) and IdBM(A) (*p*=0.21), IdBM(A+P) and IdBM(P) (*p*=1), and IdBM(A) and IdBM(P) (*p*=0.02).

Figure 1 (b) shows the mean sentence recognition scores of all conditions with 2-talker masker. 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. Two-way ANOVA with repeated measures indicated a significant effect (*F*[1, 7]=158.6, *p*<0.001) of SNR level, signal processing condition (*F*[3,21]=188.0, *p*<0.001), and a non-significant interaction (*F*[3, 21]=2.4, *p*=0.1] between SNR level and signal processing condition.

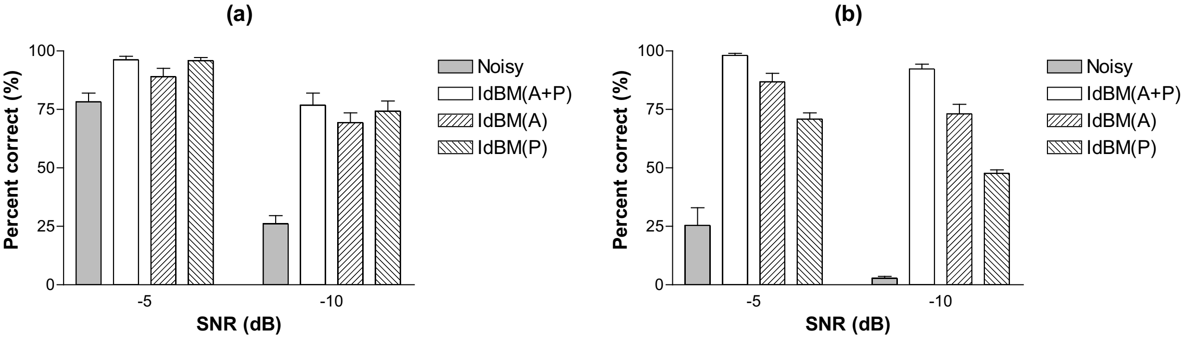

Figure 1. *Mean sentence recognition scores for all conditions with (a) SSN masker and (b) 2-talker masker. The error bars denote ±1 standard error of the mean.*

Post hoc pairwise comparisons at each SNR level showed that the recognition scores of conditions IdBM(A+P), IdBM(A) and IdBM(P) were all significantly (*p*s<0.01) larger that of condition Noisy, and there were significant difference among the four conditions (*p*s<0.01).

4. Discussion and conclusions

Following earlier studies to account for the mechanism of intelligibility advantage by IdBM processing, the present work assessed the relative contributions of amplitude and phase cues to the intelligibility advantage of IdBM-processed sentences. More specifically, this study synthesized, in addition to the traditional IdBM condition, two types of IdBM conditions containing only amplitude or phase information in those maker-‘1’ T-F units. The following observations were noted from this study. First, while the traditional IdBM-processing involved both amplitude and phase information in those preserved T-F units with local SNR larger than threshold (i.e., 0 dB), the present work found that the amplitude- or phase-only information in those masker-‘1’ T-F units could lead to significantly improved intelligibility scores in relative to unprocessed noise-masked sentences. For instance, at SSN masker and −10 dB SNR, the intelligibility score of noise-masked speech (i.e., 26.1%) was significantly improved to 69.4% at IdBM(A) condition and 74.3%% at IdBM(P) condition. Hence, the improved intelligibility by IdBM processing could be attributed to either its amplitude or phase information in selected masker-‘1’ T-F units. This finding was seen in conditions from both maskers in Figs. 1 (a) and (b), which implies that the future design of IdBM-centric algorithm may consider preserving either amplitude or phase information in those masker-‘1’ T-F units for improving speech intelligibility.

Second, the relative contributions of amplitude and phase cues to the intelligibility advantage of IdBM-processed sentences depended on the masker under investigation. Under SSN masker in Fig. 1 (a), it is seen that both IdBM(A) and IdBM(P) conditions could lead to improved speech intelligibility close to IdBM(A+P) condition, which involves both amplitude and phase information in masker-‘1’ T-F units. For instance, the intelligibility scores are 69.4%, 74.3% and 76.9%, respectively, for IdBM(A), IdBM(P) and IdBM(A+P) conditions at SSN masker and −10 dB SNR level. In other words, when masked by steady-state noise, either amplitude- or phase-only information could lead to intelligibility improvement that was traditionally observed from utilizing both amplitude and phase information. However, when the masker is interfering noise (e.g., 2-talker babble in this study), IdBM(A) or IdBM(P) condition could not lead to the same amount of intelligibility improvement as IdBM(A+P) condition did. The intelligibility score of IdBM(A+P)-processed sentences is 92.4% at 2-talker masker and −10 dB SNR, while those of IdBM(A)- and IdBM(P)-processed sentences are 73.1% and 47.6%, respectively. In addition, it is also seen that the IdBM(A)-processed sentences have a larger intelligibility score than IdBM(P)-processed sentences, e.g., 86.8% vs. 70.9% at −5 dB SNR and 73.1% vs. 47.6% at −10 dB SNR. This implies that amplitude cue may carry more perceptual information than phase cue when IdBM-processed sentences are corrupted by competing maskers. The intelligibility difference between IdBM(A) and IdBM(P) conditions in Fig. 1 (b) increases at lower SNR level, i.e., from 15.9% percentage point at −5 dB to 25.5% percentage point at −10 dB, suggesting that the intelligibility information contained in phase-only masker-‘1’ T-F units are more susceptible to competing-masker’s influence than that in amplitude-only masker-‘1’ T-F units. However, it is unclear whether the relative perceptual contributions of amplitude and phase cues would vary with the settings of IdBM processing (e.g., the number of channels and frame duration), which warrants further investigation.

The following limitations are noted from this work. Firstly, the experiment in this work was performed with a tonal language (i.e., Mandarin). Many studies have shown that phase carries important information for tonal language perception [10-12]. It is unclear whether the same findings would be achieved with a non-tonal language (e.g., English). Secondly, the test materials in this study were produced by a male speaker. Speech intelligibility score may also depend on the talker gender; hence it is necessary to extend this work with test materials pronounced by a female speaker. Thirdly, future work may assess the sensibility of the contribution on amplitude and phase with errors in the binary mask as in a real system where the ideal binary mask is not available.

In conclusion, the present study suggested that both amplitude and phase cues accounted for the intelligibility improvement observed from IdBM-processed sentences. However, their relative contributions depended on the masker under investigation. Under steady-state masker, the intelligibility improvement of the traditional IdBM-processed sentences [i.e., condition IdBM(A+P)] could be achieved by either amplitude or phase cue contained in those maker-‘1’ units. On the other hand, when corrupted by competing masker, it is the integrated contributions from both amplitude and phase cues that accounted for the intelligibility advantage of IdBM-processed sentences, and amplitude-only IdBM-processed sentences were more intelligible than phase-only IdBM-processed sentences.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61571213). This study was also supported by a grant from Neural and Cognitive Sciences Research Center, Southern University of Science and Technology, Shenzhen, China.

6. References

[1] Smith, Z. M., Delgutte, B., and Oxenham, A. J., “Chimaeric sounds reveal dichotomies in auditory perception,” Nature, 416: 87–90, 2002.

[2] Kazama, M., Gotoh, S., Tohyama, M., and Houtgast, T. “On the significance of phase in the short term Fourier spectrum for speech intelligibility,” J. Acoust. Soc. Am., 127: 1432–1439, 2010.

[3] Mowlaee, P., Saeidi, R., and Stylianou, Y., “Recent advances in phase-aware signal processing,” Speech Com., 81, 1–29, 2016.

[4] Mowlaee, P., Kulmer, J., Stahl, J., and Mayer, F., “Single-channel phase-aware signal processing in speech communication: Theory and practice,” John Wiley & Sons, 2016.

[5] Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., and Ekelid, M., “Speech recognition with primarily temporal cues,” Science, 270: 303–304, 1995.

[6] Dorman, M., Loizou, P., and Rainey, D., “Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs,” J. Acoust. Soc. Am., 102: 2403–2411, 1997.

[7] Xu, L., Thompson, C. S., and Pfingst, B. E., “Relative contributions of spectral and temporal cues for phoneme recognition,” J. Acoust. Soc. Am., 117: 3255–3267, 2005.

[8] Gilbert, G., Lorenzi, C., “The ability of listeners to use recovered envelope cues from speech fine structure,” J. Acoust. Soc. Am., 119: 2438–2444, 2006.

[9] Chen, F. and Guan, T., “Effect of temporal modulation rate on the intelligibility of phase-based speech,” J. Acoust. Soc. Am., 134: EL520–EL526, 2013.

[10] Wang, D. M., Kates, J. M., and Hansen, J. H. L., “Investigation of the relative perceptual importance of temporal envelope and temporal fine structure between tonal and non-tonal languages,” in Proc. of 15th Annual Conference of the International Speech Communication Association (InterSpeech), Singapore, 2014, pp. 495–498.

[11] Chen, F. and Zhang, Y. T., “Zerocrossing-based nonuniform sampling to deliver low-frequency fine structure cue for cochlear implant,” Digital Signal Processing, 21: 427–432, 2011.

[12] Nie, K. B., Stickney, G., and Zeng, F. G., “Encoding frequency modulation to improve cochlear implant performance in noise,” IEEE Trans. Biomed. Eng., 52: 64–73, 2005.

[13] Wang, D., “On ideal binary mask as the computational goal of auditory scene analysis,” in Speech Separation by Humans and Machines, edited by P. Divenyi (Kluwer Academic, Dordrecht), 181–197, 2005.

[14] Brungart, D., Chang, P., Simpson, B., and Wang, D., “Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation,” J. Acoust. Soc. Am., 120(6): 4007–4018, 2006.

[15] Cooke, M., “A glimpsing model of speech perception in noise,” J. Acoust. Soc. Am., 119(3): 1562–1573, 2006.

[16] Cao, S., Li, L., and Wu, X.H., “Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise,” J. Acoust. Soc. Am., 129(4): 2227–2236, 2011.

[17] Li, N. and Loizou, P. C., “Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction,” J. Acoust. Soc. Am., 123(3): 1673–1682, 2008.

[18] Li, N., and Loizou, P. C., “Effect of spectral resolution on the intelligibility of ideal binary masked speech,” J. Acoust. Soc. Am., 123: EL59–EL64, 2008.

[19] Chen, F. and Kwok, A. S. T., “Segmental contribution to the intelligibility of ideal binary-masked sentences,” in Proc. of 16th Annual Conference of the International Speech Communication Association (InterSpeech), Dresden, 2015, 3404–3407.

[20] Mayer, F. and Mowlaee, P., “Improved phase reconstruction in single-channel speech separation,” in Proc. of 16th Annual Conference of the International Speech Communication Association (InterSpeech), Dresden, 2015, 1795–1799.

[21] McAulay, R., and Quatieri, T., “Sinusoidal coding,” in Speech Coding and Synthesis, edited by W. Kleijn and K. Paliwal (Elsevier Science, New York), 1995.