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

Comparing the Contributions of Amplitude and Phase to Speech Intelligibility in a Vocoder-based Speech Synthesis Model

Fei Chen [1], *Benson C. L. Chiao* [2]

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

[2] Division of Speech and Hearing Sciences, The University of Hong Kong, Hong Kong, China

fchen@sustc.edu.cn

Abstract

Vocoder-based speech synthesis model has been long used to assess the contribution of acoustic cue for speech recognition. This study compared the perceptual contributions of amplitude and phase by using two types of stimuli, i.e., amplitude- and phase-based vocoded stimuli. The amplitude-based vocoded stimuli were synthesized by preserving amplitude fluctuation cue but discarding phase cue (i.e., setting phase to zero), while the phase-based vocoded stimuli were synthesized by preserving phase cue and discarding amplitude cue (i.e., setting amplitude to unit). Listening experiments with normal-hearing participants showed consistent findings with earlier studies that the intelligibility scores of both amplitude- and phase-based vocoded stimuli increased when using a large number of channels in vocoder-based speech synthesis. In addition, at all tested conditions, the intelligibility scores of amplitude-based vocoded stimuli were significantly larger than those of phase-based vocoded stimuli, suggesting that amplitude might carry more perceptual contribution than phase. This intelligibility advantage of amplitude over phase may be attributed to the difference in the amount of envelope information contained in the two types of vocoded stimuli.

Index Terms: Speech intelligibility, amplitude and phase, vocoder simulation.

1. Introduction

Amplitude and phase are two pieces of important acoustic information for speech perception. A number of work has been carried out to understand their importance for speech understanding in different listening environments [e.g., 1-6]. Particularly, over the past years, studies were conducted in investigating the effect of preserving temporal envelope (i.e., amplitude fluctuation) cue in a speech signal to speech perception [e.g., 1-3]. Among many, one motivating reason for these studies is the success of using temporal envelope cue in the present cochlear implant (CI, an electronic device designed to restore sound perception to deaf patients by directly stimulating their intact auditory nerves [7]) speech processors. Vocoder-based speech synthesis model has been long used to investigate the importance of amplitude cue to speech perception, and indeed it is the basis for the envelope-based speech processing of the present CI speech processors [e.g., 1-2, 7]. In vocoder-based speech synthesis, the speech signal is first decomposed into many frequency channels. The envelope waveform in each channel is extracted by wave rectification and low-pass (LP) filtering, and modulates a carrier signal (e.g., sinusoid or white noise). The outputs from all channels are finally summed together to generate the envelope-based stimuli, i.e., containing little phase information in the process of speech synthesis [7]. Results showed that using a large number of channels and a high LP cut-off frequency to extract the envelope waveform was favorable to a better perception of speech synthesized primarily with amplitude cue [1-3, 7].

While a lot of studies have suggested the importance of amplitude or envelope cue for speech recognition, much work also revealed that the recognition of amplitude-based stimuli would be affected greatly in different conditions, such as in noise and in music appreciation. For example, studies found that with a limited number of channels (or spectral resolution) used in vocoder simulation, the amplitude cue was able to support the recognition of phoneme and sentence with a high success rate in a quiet environment, but the performance in speech recognition was severely affected when there was a competing voice in an environment [8]. Besides the effect on speech perception brought by noise, Kong *et al*. found that CI listeners had a poorer performance in rhythmic pattern identification and melody recognition than normal-hearing (NH) listeners [9-10]. As the operation of most existing CI speech processors involved the extraction of amplitude cue and the elimination of phase cue contained in the original signal, it could be deduced that amplitude cue alone might not be effective enough in perceiving changes in rhythm and pitch. Furthermore, it should be noted that only a little information about fundamental frequency (F0) could be found in an amplitude-based speech (mainly dependent on the cut-off frequency used to extract envelope waveform), and the difference in the fundamental frequency of a syllable is a determining factor to distinguish the meanings of different words with identical phonemes in a tonal language [13]. Using Mandarin as an example, there are four lexical tones in total in this language system. Words with same phonemic structure but with different tones contain different meanings. A failure in recognizing the lexical tone would lead to a difficulty in identifying the Mandarin word [14].

Although phase information was eliminated in the present envelope-based speech synthesis, many recent studies found that phase (or temporal fine-structure waveform, which contains phase fluctuation information) cue carries important information for tone identification, speech perception in noise, and music appreciation [e.g., 8-12]. Unfortunately, our knowledge on the effect of phase information to speech perception is still limited. Chen and Guan investigated the effect of temporal modulation rate (equivalent to the low-pass cut-off frequency in extracting envelope waveform) on the intelligibility of phase-based speech, which was generated by

http://dx.doi.org/10.21437/Interspeech.2016-66

using a phase-based vocoder to eliminate envelope cue (i.e., setting amplitude to unit) and preserve phase cue in each channel during vocoder processing (see more in section 2.2) [4]. Their result showed that the intelligibility of phase-based speech was significantly improved when using a high temporal modulation rate and/or a large number of channels in vocoder-based speech synthesis.

While many studies have assessed the importance of amplitude and phase cues for speech recognition, so far little has been done to compare the contributions of amplitude and phase to speech intelligibility. This is partially because earlier work was designed with different speech synthesis methods, e.g., vocoder or short-time Fourier transform based model. The purpose of present work was to use a vocoder-based simulation to compare the relative perceptual contributions of amplitude and phase cues for speech understanding, and to examine the effect of vocoding parameter (i.e., the number of channels in this study) on the intelligibility of amplitude- and phase-based vocoded speech.

2. Methods

2.1. Subjects and materials

Eighteen NH listeners who were native-speakers of Mandarin Chinese participated in this experiment. Test sentences were taken from Mandarin Hearing in Noise Test (MHINT) database [4]. Twenty-four lists of words were present in the MHINT test items. In each list there were 10 sentences, with each sentence containing 10 target words. All the sentences were spoken by a native male Mandarin speaker. The F0 of the speaker ranged from 75 to 180 Hz, and the voice was recorded at a sampling rate of 16 kHz. A speech-spectrum shaped noise (SSN) was used to corrupt the test sentences at 0 and 5 dB signal-to-noise ratio (SNR), and the SNR levels were selected from a pilot experiment to avoid the ceiling/floor effect in speech perception.

2.2. Signal processing

Speech signals were first masked by the SSN masker at 0 or 5 dB SNR. To synthesize the amplitude- and phase- based vocoded stimuli, a pre-emphasis (high-pass) filter (2000 Hz cut-off) with a 3 dB/octave roll-off was used to process the speech signals. After that, sixth-order Butterworth analysis filters were used to band-pass the signals into N (N=4, 8, or 12 in this study) frequency channels between 80 and 6000 Hz. The cut-off frequencies of the N band-pass analysis filters were computed according to the cochlear frequency-position mapping function [15]. For the amplitude-based vocoded stimuli, sinusoids were generated with amplitudes equal to the root-mean-square (RMS) energies of the band-passed signals (computed every 2.5 ms), initial phases equal to 0, and frequencies equal to the center frequencies of the bandpass filters. After summing up the sinusoids of all channels, the RMS value of the synthesized stimulus segment was adjusted to the same value as the original signal. For the phase-based vocoded stimuli, sinusoids were generated with amplitudes equal to one (i.e., a constant number), frequencies equal to the center frequencies of the bandpass filters, and initial phases estimated from the fast Fourier transform of every 2.5 ms of non-overlapping speech frames. After summing up the sinusoids of all channels, the RMS value of the synthesized stimulus was adjusted to the same value as the original signal. The signal processing method to generate the phase-based vocoded stimuli followed that used in [4].

2.3. Procedure

The experiment was performed in a sound-proof room, and the participants were required to listen to the stimuli which were played monaurally at a comfortable listening level. The reason for playing stimuli monaurally was to simulate the hearing of most CI users, who were implanted with CI devices unilaterally. Before the testing session, each participant attended a training session (about 10 minutes) to listen to 40 amplitude-/phase-based vocoded sentences (different from those used in testing session) in order to familiarize himself/herself with the testing procedure. During the testing session, participants were instructed to listen to the sentences, and then were required to orally repeat the sentences they heard. Each target word in the sentence was scored as correct or incorrect. There were 18 [=three numbers of channels (i.e., N=4, 8, and 12) × two SNR levels (i.e., 0, and 5 dB) × two signal processing conditions (i.e., amplitude- and phase-based vocoded stimuli)] testing conditions in total. Each condition was tested with one list of MHINT sentences (i.e., 10 sentences), and there was no repetition in the use of the same list across testing conditions. Randomization of the order of the testing conditions was performed across subjects. A 5-minute break was given to subjects every 30 minutes during the test. The percentage intelligibility score was calculated by dividing the number of correctly identified target words over the total number of target words in a testing condition.

3. Results

Figure 1 shows the mean sentence recognition scores for all conditions. Statistical significance was determined by using the percent correct score as the dependent variable, and SNR level (0 and 5 dB), the number of channels (i.e., 4, 8 and 12) and signal processing condition (i.e., amplitude- and phase-based vocoded speech) as the three within-subjects factors. Because of the floor effect, the scores were first converted to rational arcsine units (RAU) by using the rationalized arcsine transform [16]. Three-way repeated measures analysis of variance indicated a significant effect of SNR level (F[1, 17]=84.20, p<0.001), the number of channels (F[2, 34]=141.71, p<0.001), and signal processing condition (F[1, 17]=141.71, p<0.001), a significant interaction between SNR level and the number of channels (F[2, 34]=4.13, p<0.05), between SNR level and signal processing condition (F[1, 17]=10.40, p<0.05), and between the number of channels and signal processing condition (F[2, 34]=12.98, p<0.05), and a non-significant interaction among SNR level, the number of channels and signal processing condition (F[2, 34]=0.52, p>0.05). The significant interaction appears to be due to the floor effect of intelligibility scores of phase-based vocoded speech synthesized at 4 or 8 channels. Post hoc pairwise comparisons showed significant (p<0.05) difference between paired scores at the same SNR level and the same number of channels.

4. Discussion and conclusions

The present work showed that at all tested conditions, the intelligibility score of amplitude-based vocoded stimuli (containing little phase information) was significantly larger than that of phase-based vocoded stimuli (eliminating amplitude fluctuation information). This result suggested that envelope (or amplitude) cue may carry more perceptual contribution than phase cue in speech perception. This intelligibility advantage of amplitude over phase is clearly

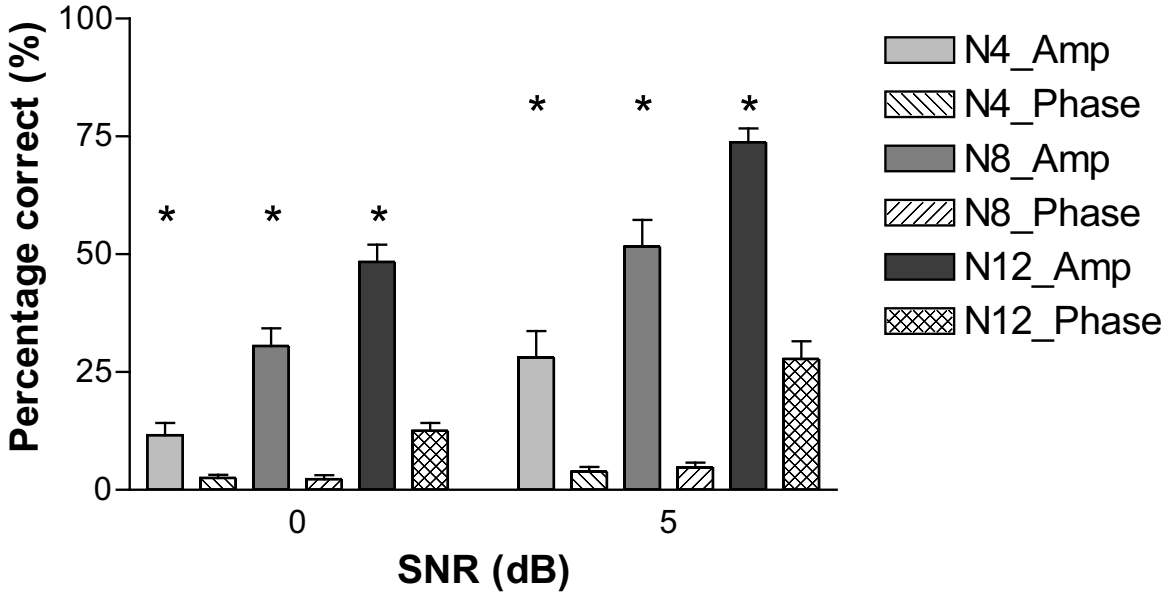

Figure 1. *Mean sentence recognition scores for all conditions. The error bars denote ±1 standard error of the mean. 'N' denotes the number of channels used in vocoder simulation, and 'Amp' and 'Phase' are for amplitude- and phase-based vocoded conditions, respectively. Asterisks denote that the score is significantly ($p<0.05$) larger than its paired score from phase-based condition.*

shown at conditions with a small number of channels (e.g., N=4 or 8) in Fig. 1. Earlier work showed that speech synthesized with envelope waveforms from 4 to 8 channels may give almost perfect speech perception in quiet [1-2]. When the SSN masker was combined with clean speech at 5 dB SNR level, the envelope-based vocoded stimuli were still intelligible, i.e., recognition scores of 28.1% and 51.6% for N=4 and 8, respectively. However, the phase-based vocoded speech contained little intelligibility information, i.e., 3.9% and 4.8%, respectively. This clearly demonstrates the advantage of amplitude against phase for speech understanding. In addition, Fig. 1 shows that the intelligibility score of amplitude-based vocoded speech continuously increases when using a larger number of channels from 4 to 12; however, only when the number of channels is increased to 12, a noticeable intelligibility improvement is seen for phase-based vocoded speech.

Although the exact mechanism of phase-based speech recognition is not clear now, many earlier studies suggested that it may be largely attributed to the recovered envelope from the phase-based speech [e.g., 5]. The process to extract recovered envelope waveform can by implemented by the functional modules of bandpass filtering, waveform rectification and low-pass filtering in the signal transmission pathway in the periphery auditory system. Hence, the perception of both envelope- and phase-based speech may be rooted in the usage of envelope cue (e.g., recovered envelope in phase-based speech) for speech understanding. The difference in the amount of envelope information contained in the two types (i.e., amplitude- and phase-based) of vocoded stimuli may account for the amplitude advantage over phase for speech intelligibility. This is so because the amplitude-based vocoded stimuli directly make use of the amplitude modulation in their speech synthesis; on the other hand the phase-based vocoded stimuli use phase modulation information during the synthesis, but their perception relies on the recovered envelope cue. We suppose that more envelope cue is present in the amplitude-based vocoded stimuli than in the phase-based vocoded stimuli, yielding a higher intelligibility of amplitude-based vocoded stimuli than phase-based vocoded stimuli revealed in this study.

This study used vocoder model to compare the perceptual impacts of amplitude and phase to speech perception. Note that other model might also be utilized for this comparison purpose, e.g., using the STFT-based model [6]. It is unclear whether the advantage of amplitude over phase can also be achieved when using the STFT-based model for speech synthesis. Different model may use its own parameters to control the process of speech synthesis. Vocoder-based speech synthesis model uses LP cut-off frequency and the number of channels to control the temporal and spectral resolutions, respectively, in its speech synthesis. On the other hand, window length is an important factor determining the intelligibility of speech synthesized with the STFT-based model. Kazama *et al*. assessed the roles of spectral resolution and temporal resolution on the significance of phase information in the STFT spectrum for speech intelligibility, and their speech intelligibility data showed the significance of phase spectrum for long (> 256 ms) and for very short (< 4 ms) windows [6]. It is possible that new findings may be obtained with using the STFT-based model to replace the vocoder model in this study, which warrants further investigation

In conclusion, the present study used the vocoder-based speech synthesis model to compare the relative perceptual contributions of amplitude and phase cues. Results found that the amplitude-based vocoded speech consistently yielded a larger intelligibility score than the phase-based vocoded speech. This finding suggested that amplitude might carry

more perceptual contribution than phase, which might be attributed to the difference in the amount of envelope information contained in the two types of vocoded stimuli.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61571213). This study was the basis for the Bachelor's thesis of the second author (B.C.L.C.).

6. References

[1] Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M., "Speech recognition with primarily temporal cues," Science, 270 (5234): 303–304, 1995.

[2] Dorman, M. F., Loizou, P. C., 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 (4): 2043–2411, 1997.

[3] Xu, L., Thompson, C. S., and Pfingst, B. E., "Relative contributions of spectral and temporal cues for phoneme recognition," J. Acoust. Soc. Am., 117 (5): 3255–3267, 2005.

[4] Chen, F. and Guan, T., "Effect of temporal modulation rate on the intelligibility of phase-based speech," J. Acoust. Soc. Am., 134 (6): EL520–EL526, 2013.

[5] Zeng, F. G., Nie, K. B., Liu, S., Stickney, G., Del Rio, E., Kong, Y. Y., and Chen, H. B., "On the dichotomy in auditory perception between temporal envelope and fine structure cues (L)," J. Acoust. Soc. Am., 116 (3): 1351–1354, 2004.

[6] 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 (3): 1432–1439, 2010.

[7] Loizou, P. C., "Introduction to cochlear implants," IEEE Eng. Med. Biol., 18 (1): 32–42, 1999.

[8] Nie, K., Barco, A., and Zeng, F. G., "Spectral and temporal cues in cochlear implant speech perception," Ear Hear., 27(2): 208–217, 2006.

[9] Kong, Y. Y., Cruz, R., Jones, J. A., and Zeng, F. G., "Music perception with temporal cues in acoustic and electric hearing," Ear Hear., 25 (2): 173–185, 2004.

[10] Kong, Y. Y. and Zeng, F. G., "Temporal spectral cues in Mandarin tone recognition," J. Acoust. Soc. Am., 120 (5): 2830–2840, 2006.

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

[12] 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.

[13] Klein, D., Zatorre, R. J., Milner, B., and Zhao, V., "A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers," Neuroimage, 13 (4): 646–653, 2001.

[14] Chen, F. and Loizou, P. C., "Predicting the intelligibility of vocoded and wideband Mandarin Chinese," J. Acoust. Soc. Am., 129 (5): 3281–3290, 2011.

[15] Greenwood, D. D., "A cochlear frequency-position function for several species-29 years later," J. Acoust. Soc. Am., 87 (6): 2592–2605, 1990.

[16] Studebaker, G. A., "A 'rationalized' arcsine transform," J. Speech Lang. Hear. Res., 28: 455–462, 1985.