INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Effect of Carrier Bandwidth on Understanding Mandarin Sentences in Simulated Electric-acoustic Hearing

Feng Wang [1], *Jing Chen* [2, 3], *Fei Chen* [1]

[1] Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China
[2] Department of Machine Intelligence, Speech and Hearing Research Center, and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China
[3] Peng Cheng Laboratory, Shenzhen, China
fchen@sustech.edu.cn

Abstract

For patients suffering with high-frequency hearing loss and preserving low-frequency hearing, combined electric-acoustic stimulation (EAS) may significantly improve their speech perception compared with cochlear implants (CIs). In combined EAS, a hearing aid provides low-frequency information via acoustic (A) stimulation and a CI evokes high-frequency sound sensation via electrical (E) stimulation. The present work investigated the EAS advantage when only a small number (i.e., 1 or 2) of channels were provided for electrical stimulation in a CI, and the effect of carrier bandwidth on understanding Mandarin sentences in a simulation of combined EAS experiment. The A-portion was extracted via low-pass filtering processing and the E-portion was generated with a vocoder model preserving multi-channel temporal envelope waveforms, whereas a noise-vocoder and a tone-vocoder were used to simulate the effect of carrier bandwidth. The synthesized stimuli were presented to normal-hearing listeners to recognize. Experimental results showed that while low-pass filtered Mandarin speech was not very intelligible, adding one or two E channels could significantly improve the intelligibility score to above 86.0%. Under the condition with one E channel, using a large carrier bandwidth in noise-vocoder processing provided a better intelligibility performance than using a narrow carrier bandwidth in tone-vocoder processing.

Index Terms: combined electric-acoustic stimulation, combined stimulation advantage, cochlear implants

1. Introduction

A cochlear implant (CI) is presently the only medical treatment for partially restoring hearing to patients with profound-to-severe hearing loss [e.g., 1-2]. The incoming speech signal is coded into multi-channel envelope waveforms, which are subsequently used to stimulate auditory nerves via an electrode array and elicit speech perception [1]. A number of factors may affect the performance of CI users in speech understanding. To understand the perceptual contributions of multi-channel envelope waveforms, vocoder simulations have been long used because they can avoid the patient-specific factors affecting speech understanding by patients fitted with CIs [3]. In a vocoder simulation, the speech signal is processed by an acoustic model simulating CI speech processing, and the vocoder-processed stimuli are presented to normal-hearing (NH) listeners to recognize. Many factors have been found to affect the understanding of vocoded stimuli, including the number of spectral channels [3], the cutoff frequency used to extract the envelope waveform [3], the type of carrier signal in vocoder-based speech synthesis [4-5], etc. The effects of these factors provided implications on how CI settings in its speech processing and coding affected CI based speech perception, e.g., the number of channels for electrical stimulation, the modulation rate of extracted envelope waveform, the excitation spread in electrical stimulation [6], etc.

Many hearing-impaired patients preserve some degree of low-frequency (LF) residual hearing and largely suffer from hearing loss at high frequency. The 'combined-stimulation advantage' refers to an improvement in speech recognition when electrical stimulation in CI is supplemented by LF acoustic information [e.g., 7-11]. Recent work has consistently shown that adding LF acoustic information may significantly improve CI speech perception, such as in noise [e.g., 11] and reverberation [e.g., 12]. Several studies have investigated factors affecting understanding of vocoded speech plus LF acoustic information. Chen and Loizou found that acoustic landmarks carry critical information for understanding combined electric-acoustic stimulation (EAS)-processed stimuli [13]. Carroll et al. showed that the fundamental frequency (F0) cue significantly contributes to the benefit received with combined EAS hearing [10]. They attributed this benefit to the modulation of the frequency rather than the amplitude component. Chen and Chen recently assessed the perceptual impact of vowels and consonant-vowel transitions in simulated EAS hearing [14]. They found that adding consonant-vowel transitions in combined EAS yields sentence recognition performance equivalent to that observed with E stimulation and full speech segments. Wu et al. studied Mandarin sentence understanding when the electric (E) and acoustic (A) portions were not temporally aligned in simulated combined EAS [15]. They showed a significant decrease of the intelligibility score caused by the temporal misalignment in the two portions of EAS processing, suggesting the need to avoid temporal misalignment in EAS.

Early work has shown the benefits of using a large number of spectral channels and focused electrical stimulation for CI speech understanding. More non-overlapping spectral channels may lead to a better spectral resolution for speech perception, and focused electrical stimulation can largely avoid the interaction of envelope information between adjacent stimulation channels. In terms of vocoder simulation, white noise and pure tone have been commonly used as carrier

http://dx.doi.org/10.21437/Interspeech.2021-24

Table 1. *Cutoff frequencies (in Hz) for the channel allocation of lowpass and bandpass filters for all test conditions in this experiment.*

A-only stimulation	E-only stimulation		Combined EAS	
	N=1	N=2	N=1	N=2
250	80, 6000	80, 1158, 6000	250, 6000	250, 1498, 6000
500			500, 6000	500, 1902, 6000

signals to synthesize vocoded stimuli. They have different bandwidth, and may simulate the effect of current spread and/or spread of excitation in CI. Chen et al. simulated the effects of decay rates of excitation spread in CIs on the intelligibility of Mandarin speech in noise [16]. They showed that significant benefit for Mandarin sentence recognition in noise was observed with narrower type of excitation. Many studies have suggested that tone-vocoded (TV) stimuli (using pure tone as the carrier signal) offer an intelligibility advantage over noise-vocoded (NV) stimuli (using white noise as the carrier signal) [4-5] in vocoder simulations. One explanation for this intelligibility advantage concerns the spectral sidebands that are contained in TV stimuli when a pure tone is multiplied by the envelope waveform. The amplitude-modulated tone carrier has two spectral sidebands that impose a periodic temporal structure in voiced speech segments on the tone-vocoder's output, with the talker's pitch being preserved over most voiced segments. Hence, the spectral sidebands contain an additional cue that is beneficial for speech intelligibility.

Although more E channels may increase the spectral resolution in electrical stimulation, the need for more E channels poses challenges for electrode design. This work investigated whether the EAS advantage still held when only a limited number (e.g., 1 or 2) of E channels were used for electrical stimulation, which was the first aim of the present work. Second, while early work suggested the advantage of focused electrical stimulation, those findings were reported with a large number of E channels in electrical stimulation. It is unclear, under the EAS condition with a limited number (e.g., 1 or 2) of E channels, whether focused electrical stimulation (i.e., simulated with pure-tone carrier in a tone-vocoder) still shows its intelligibility advantage, which was the second aim of this work. Two cases of carrier bandwidth were implemented in this work, with white noise carrier simulating a spread of electrical excitation and pure-tone carrier for a focused electrical excitation.

2. Methods

2.1. Subjects and materials

This experiment involved 15 NH listeners (7 males and 8 females, and pure-tone thresholds better than 20 dB HL at octave frequencies from 125 to 8000 Hz in both ears). All subjects were native speakers of Mandarin Chinese and were paid for their participation. The experimental procedure involving human subjects was approved by the Institution's Ethical Review Board of Southern University of Science and Technology.

Speech material comprised sentences extracted from the Mandarin Hearing in Noise Test (MHINT) database [17]. The MHINT corpus includes 24 lists, each with 10 sentences and 10 keywords per sentence. All sentences were spoken by a male native Mandarin Chinese speaker having a fundamental frequency of 75–180 Hz, which was recorded at a sampling rate of 16 kHz.

2.2. Signal processing

This experiment included three signal-processing conditions. The first condition simulated acoustic-only (A-only) stimulation. Speech signal was processed by low-pass (LP) filtering to generate the LP-processed stimuli. LP filtering was implemented by using a linear-phase finite impulse response filter with filter order of $10 \times f_s/f_{cut}$, where f_s is the sampling rate (16 kHz) and f_{cut} is the LP cutoff frequency (f_{cut} = 250 or 500 Hz).

The second processing condition simulated N-channel electric-only (E-only) stimulation (N = 1 or 2). All MHINT sentences were processed by a tone- or noise-vocoder [5]. To implement the tone vocoder, speech signals were first processed through a pre-emphasis filter (first-order high-pass filter with 1200-Hz cutoff frequency). Then, signals were bandpass-filtered into N frequency bands between 80 and 6000 Hz with sixth-order Butterworth filters. Cutoff frequencies for the channel allocation of bandpass filters are given in Table 1 [18]. From each band, the envelope was extracted by full-wave rectification and LP filtering with a 200-Hz cutoff frequency by way of a fourth-order Butterworth filter. Sine waves at the center frequencies of the bandpass filters were generated with amplitudes modulated by the extracted envelopes. All amplitude-modulated sine waves from the resultant set of bands were summed to generate a TV stimulus, whose amplitude was adjusted to have the same root-mean-square (RMS) energy as the original input speech signal. Implementation of the noise vocoder was similar to that of the tone vocoder, except that white noise instead of a sine wave was used as the carrier signal, and amplitude-modulated by the extracted envelope. Output from each band was further band-limited with the same bandpass filter at that band. All amplitude-modulated noises (with band-limiting processing) were summed to generate the NV stimulus, with its amplitude adjusted to have the same RMS power as the original input speech signal.

The third processing condition simulated the combined EAS. To simulate the effect of EAS with LF residual hearing up to the LP cutoff frequency, we first generated the LP-processed acoustic stimulus with the specified LP cutoff frequency. Next, we synthesized the vocoded stimulus spanning the frequency range from the LP cutoff frequency to 6000 Hz. Cutoff frequencies for the channel allocation of bandpass filters are given in Table 1. Finally, we combined the LP-processed acoustic stimulus with the vocoded stimulus to generate the EAS-processed stimulus. Figure 1 shows the examples of spectrograms of two EAS-processed stimuli. As seen in Fig. 1, using pure-tone carrier (Fig. 1 (b)) yields

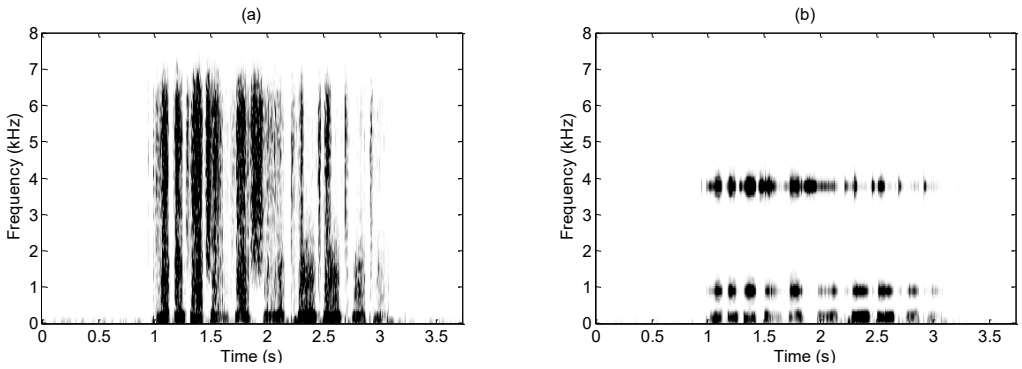

Figure 1: *Spectrograms of two EAS-processed stimuli whereas the acoustic portion was extracted with a 250-Hz-cutoff low-pass filter and the electric portion included N=2 channels. Panels (a) and (b) use white noise and pure-tone as carrier signals, respectively, in the electric portion.*

narrower bandwidth (simulating focused electrical stimulation in a CI) than using white-noise carrier (Fig. 1 (a)).

2.3. Procedure

The experiment was performed in a sound booth, and stimuli were played to listeners monaurally through an HD 650 circumaural headphone (Sennheiser, Germany) set at a comfortable listening level. Before the actual testing session, each subject participated in a 10-min training session and was given four lists of 10 MHINT sentences (different with those used in testing session). The training session familiarized the subjects with the testing procedure and conditions. During the training session, the subjects were allowed to read transcriptions of the training sentences while they were listening to the sentences. In the testing session, the order of the conditions was randomized across subjects, and the subjects were asked to orally repeat all of the words they heard. In addition, the lists were randomized across listeners. The sentences used during testing were not the same as any of the training sentences. Each subject participated in a total of 14 conditions [= 2 conditions of acoustic-only stimulation (fcut=250 and 500 Hz) + 4 conditions of electric-only stimulation (i.e., 2 carrier signal types × 2 channel numbers) + 8 conditions of combined EAS (2 LP cutoff frequencies in acoustic stimulation × 2 channel numbers in electric stimulation × 2 carrier signal types)]. One list of 10 Mandarin sentences was used per test condition, and none of the sentences was repeated across conditions. Subjects were allowed to listen to each stimulus a maximum of three times, and were asked to repeat as many words as they could recognize. A simple custom software interface was designed for the listening experiment, which each participant used to control the auditory delivery of the processed stimuli. During the testing session, a tester accompanied the participant and scored his/her response in the computer. A 5-minute break was given every 30 minutes to avoid listening fatigue. The intelligibility score for each condition was computed as the ratio between the number of correctly recognized words and the total number of words contained in each MHINT list. The total testing time was around one hour.

3. Results

Mean sentence recognition scores for all conditions are shown in Fig. 2. For all data analysis, recognition scores were first converted to rational arcsine units by using the rationalized arcsine transform [19]. Figure 2 (a) gives the results of the acoustic-only stimulation. One-way repeated-measures analysis of variance (ANOVA) indicated a significant effect of LP cutoff frequency ($F_{1, 14} = 108.67$, $p < 0.005$).

Figures 2 (b) and (c) show the results of the combined EAS stimulation with N=1 and 2 E channels, respectively. For comparison purposes, results of the E-only stimulation are also shown. For the E-only stimulation, the recognition scores of the TV condition are significantly ($p < 0.05$) larger than those of the NV condition.

For the combined stimulation, statistical significance was determined by using the recognition score as the dependent variable and using the LP cutoff frequency and carrier signal type as within-subject factors. With one E channel (Fig. 2 (b)), two-way repeated-measures ANOVA indicated a significant effect of LP cutoff frequency ($F_{1, 14} = 183.15$, $p < 0.005$), a significant effect of carrier signal type ($F_{1, 14} = 20.92$, $p < 0.005$), and a non-significant interaction between these two within-subject variables ($F_{1, 14} = .96$, p=0.345). Paired t-tests revealed significant performance differences ($p < 0.05$) between paired TV and NV stimuli at all LP cutoff frequencies. When the LP cutoff frequency was 250 or 500 Hz, the recognition score of the NV condition was significantly ($p < 0.05$) larger than that under the TV condition. Compared to the results under the electric-only condition, the combined-stimulation advantage was seen under LP cutoff frequencies of 250 and 500 Hz.

For the results with two channels (Fig. 2 (c)), two-way repeated-measures ANOVA indicated a significant effect of LP cutoff frequency ($F_{1, 14} = 124.80$, $p < 0.005$), a non-significant effect of carrier signal type ($F_{1, 14} = 2.84$, $p = 0.11$), and a non-significant interaction between these two variables ($F_{1, 14} = .026$, $p = 0.873$). Paired t-tests revealed non-significant performance differences ($p > 0.05$) between paired TV and NV speech under LP cutoff frequencies of 250 and 500 Hz. In addition, when compared to the results under the E-only condition, the combined-stimulation advantage was seen under LP cutoff frequencies of 250 and 500 Hz.

4. Discussion and conclusions

This study used a small number (N = 1 or 2) of E channels in vocoding processing to study the combined stimulation advantage on understanding Mandarin sentences. Early CI (i.e., electric-only stimulation) studies showed that Mandarin-speaking listeners could have an almost-perfect understanding of the electric-only stimuli with up to 4 E channels in a

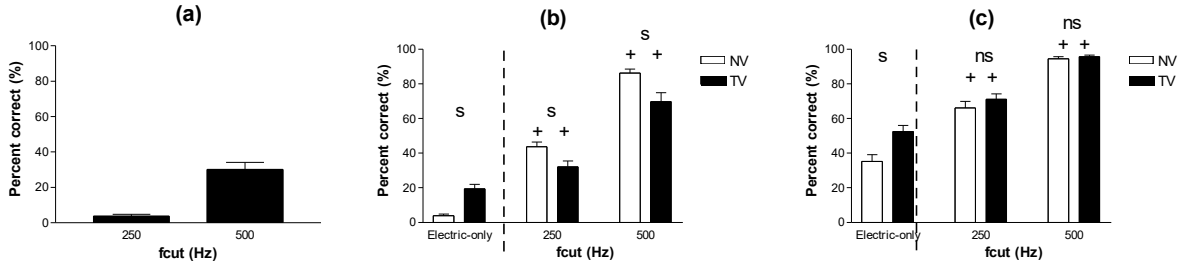

Figure 2: *Mean sentence recognition scores for all conditions. Panel (a) shows the processing conditions of acoustic-only stimulation. Panels (b) and (c) show the processing conditions of electric-only and combined stimulations with N=1 and N=2 channels, respectively. The error bars denote ±1 standard error of the mean. 's' and 'ns' denote significant and non-significant difference, respectively, between paired NV and TV conditions at the same LP cutoff frequency. '+' denotes that the recognition score is significant larger than that of the paired electric-only condition with the same type of carrier signal.*

vocoder [e.g., 20]. The present work further showed that, in the context of combined electric-acoustic stimulation, intelligibility score with N=1 or 2 E channels in a vocoder was significantly improved by adding LP acoustic information with a LP cutoff frequency of 250 or 500 Hz.

Most early studies on the intelligibility of electric-only stimuli showed the perceptual advantage of tone versus noise carriers in vocoder simulations [4-5]. This TV vs. NV advantage is also seen from the results of the electric-only stimuli in Figs. 2 (b) and (c). However, in the scenario of combined stimulation, the present work found that this advantage was affected by the number of E channels in electrical stimulation. With N=1 E channel (Fig. 2 (b)), NV electric-only stimuli were more intelligible than TV electric-only stimuli at LP cutoff frequencies of 250 and 500 Hz. In contrast, with N=2 E channels (Fig. 2 (c)), differences of the intelligibility scores between the NV and TV conditions were not significant ($p > 0.05$) at LP cutoff frequencies of 250 and 500 Hz. These results suggested that adding LF acoustic information in combined stimulation might affect the perceptual advantage of TV over NV.

The intelligibility advantage of noise versus tone carriers in the combined stimulation in Fig. 2 (b) might be attributed to the effect of small channel numbers (i.e., N = 1) in electrical stimulation, which was also reported in [21]. For small numbers of E channels, TV stimuli using low envelope cutoff frequencies were less intelligible than NV stimuli [21]. Recently, Fu et al. assessed the perceptual importance of carrier bandwidth in CI simulations [6]. Carrier bandwidth was varied across three carriers: broad-band noise, narrow-band noise, and sine waves. Reducing the bandwidth in CI simulations significantly affected the intelligibility of electric-only stimuli, but not EAS performance. However, they used eight-channel electrical stimulation, in contrast to the small numbers (i.e., N = 1 and 2) of electric channels used in this work.

In conclusion, the present work used vocoder simulation to examine how LF acoustic information in acoustic stimulation and carrier signal type in electrical stimulation affected the combined-stimulation advantage. Specially, this work investigated the EAS advantage under the condition with a small number of E channels in EAS. Results showed that, under the condition with a small number of E channels (e.g., N = 1 or 2) in electrical stimulation, adding LF (≥ 250 Hz) acoustic information might increase the intelligibility of EAS stimuli (with either tone or noise carrier) relative to the electric-only stimuli; however, EAS stimuli with the noise carrier were more intelligible than those with the tone carrier when only one E channel was utilized in electrical stimulation.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61771023). Part of this study was the basis for the Bachelor's thesis of the first author (F.W.).

6. References

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

[2] Chen, F., Ni, W. L., Li, W. Y., and Li, H. W., "Cochlear Implantation and Rehabilitation," in Hearing Loss: Mechanisms, Prevention and Cure, Eds: Huawei Li, and Renjie Chai, 129–144, 2019.

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

[4] Whitmal, N. A., Poissant, S. F., Freyman, R. L., and Helfer, K. S., "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," J. Acoust. Soc. Am., 122: 2376–2388, 2007.

[5] Chen, F., Zheng, D. C., and Tsao, Y., "Effects of noise suppression and envelope dynamic range compression to the intelligibility of vocoded sentences for a tonal language," J. Acoust. Soc. Am., 142: 1157–1166, 2017.

[6] Fu, Q. J., Galvin III, J. J., and Wang, X. S., "Effect of carrier bandwidth on integration of simulations of acoustic and electric hearing within or across ears," J. Acoust. Soc. Am., 142: EL561–EL566, 2017.

[7] Qin, M., and Oxenham, A., "Effects of introducing unprocessed low-frequency information on the reception of the envelope-vocoder processed speech," J. Acoust. Soc. Am., 119: 2417–2426, 2006.

[8] Luo, X., and Fu, Q. J., "Contribution of low-frequency acoustic information to Chinese speech recognition in cochlear implant simulations," J. Acoust. Soc. Am., 120: 2260–2266, 2006.

[9] Micheyl, C., and Oxenham, A. J., "Comparing models of the combined-stimulation advantage for speech recognition," J. Acoust. Soc. Am., 131: 3970–3980, 2012.

[10] Carroll, J., Tiaden, S., and Zeng, F. G., "Fundamental frequency is critical to speech perception in noise in combined acoustic and electric hearing," J. Acoust. Soc. Am., 130: 2054–2062, 2011.

[11] Chang, J. E., Bai, J. Y., and Zeng, F. G., "Unintelligible low-frequency sound enhances simulated cochlear-implant speech

recognition in noise," IEEE Trans. Biomed. Eng., 53: 2598–2601, 2006.
[12] Tillery, K. H., Brown, C. A., and Bacon S. P., "Comparing the effects of reverberation and of noise on speech recognition in simulated electric-acoustic listening," J. Acoust. Soc. Am., 131: 416–423, 2012.
[13] Chen, F., and Loizou, P., "Contribution of consonant landmarks to speech recognition in simulated acoustic-electric hearing," Ear Hear., 31: 259–267, 2010.
[14] Chen, F., and Chen, J., "Perceptual contributions of vowels and consonant-vowel transitions in simulated electric-acoustic hearing", J. Acoust. Soc. Am., 145: EL197–EL202, 2019.
[15] Wu, H. D., Lin, W. H., Chen, F., and Zheng, D. C., "Effect of temporal misalignment on understanding Mandarin sentences in simulated combined electric-and-acoustic stimulation," J. Acoust. Soc. Am. 148, EL433–EL439, 2020.
[16] Chen, F., Guan, T., and Wong, L. L. N., "Effects of excitation spread on the intelligibility of Mandarin speech in cochlear implant simulations," in Proc. of 8th International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong, December 5-8, 2012, pp. 35–39.
[17] Wong, L. L. N., Soli, S. D., Liu, S., Han, N., and Huang, M. W., "Development of the Mandarin Hearing in Noise Test (MHINT)," Ear Hear., 28: 70S–74S, 2007.
[18] Greenwood, D. D., "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am., 87: 2592–2605, 2009.
[19] Studebaker, G. A., "A 'rationalized' arcsine transform," J. Speech Hearing Research 28: 455–462, 1985.
[20] Fu, Q. J., Zeng, F. G., Shannon, R. V., and Soli, S. D., "Importance of tonal envelope cues in Chinese speech recognition," J. Acoust. Soc. Am., 104: 505–510, 1998.
[21] Rosen, S., Zhang, Y., and Speers, K., "Spectral density affects the intelligibility of tone-vocoded speech: Implications for cochlear implant simulations," J. Acoust. Soc. Am., 138: EL318–EL323, 2015.