INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Simulations of high-frequency vocoder on Mandarin speech recognition for acoustic hearing preserved cochlear implant

Tsung-Chen Wu[1], *Tai-Shih Chi*[1], *and Chia-Fone Lee*[2]

[1]Department of Electrical and Computer Engineering
National Chiao Tung University, Hsinchu, Taiwan 300, R.O.C.
[2]Department of Otolaryngology
Hualien Tzu Chi Hospital, Hualian City, Hualien County, Taiwan 970, R.O.C.
xyz7475wu@gmail.com, tschi@mail.nctu.edu.tw, e430013@yahoo.com.tw

Abstract

Vocoder simulations are generally adopted to simulate the electrical hearing induced by the cochlear implant (CI). Our research group is developing a new four-electrode CI microsystem which induces high-frequency electrical hearing while preserving low-frequency acoustic hearing. To simulate the functionality of this CI, a previously developed hearing-impaired (HI) hearing model is combined with a 4-channel vocoder in this paper to respectively mimic the perceived acoustic hearing and electrical hearing. Psychoacoustic experiments are conducted on Mandarin speech recognition for determining parameters of electrodes for this CI. Simulation results show that initial consonants of Mandarin are more difficult to recognize than final vowels of Mandarin via acoustic hearing of HI patients. After electrical hearing being induced through logarithmic-frequency distributed electrodes, speech intelligibility of HI patients is boosted for all Mandarin phonemes, especially for initial consonants. Similar results are consistently observed in clean and noisy test conditions.

Index Terms: vocoder simulation, Mandarin speech recognition, cochlear implant, hearing impaired model

1. Introduction

Cochlear implants (CI) have been successfully used to help severely hearing-impaired patients and auditory aging people by electrically stimulating the auditory nerves [1]. Nevertheless, intrusively implanting electrodes into the cochlea would increase the risk of losing the residual acoustic hearing and having bacterial meningitis [2]. To avoid these risks, our research group is developing a novel CI microsystem which places four electrodes on the bone surface of the cochlea [3]. The round window of the cochlea is not pierced during the surgery such that the acoustic hearing is preserved. This novel CI microsystem will provide high-frequency electrical hearing to the user while preserving his own acoustic hearing. It is suitable for people with high-frequency hearing loss most probably due to aging.

In this paper, we conduct psychoacoustic experiments on Mandarin speech recognition to test the feasibility of our CI microsystem. All experiments are under the paradigm of coexistence of high-frequency electrical hearing and low-frequency acoustic hearing. In our previous work, we developed a personalized hearing-impaired (HI) hearing model, which simulates acoustic hearing of a HI patient, in the filterbank framework [4]. This model has been validated with hearing test data from 4 HI patients. In other words, normal people hear speech processed through the personalized model would report similar speech intelligibility scores as that HI patient. The parameters of the personalized model are hearing thresholds, degrees of loudness recruitment and reductions of frequency resolution. Hearing thresholds and degrees of loudness recruitment are encoded by the minimum audible levels (MALs) in each subband, and reductions of frequency resolution are encoded by the broadened factors (BFs) of cochlear filters. Combining this model and a vocoder, we are able to evaluate the proposed CI system by conducting psychoacoustic experiments on normal-hearing (NH) subjects. In this paper, two sets of personalized parameters from patients in [4] are selected because the high-frequency MALs of these two patients are very high. In other words, these two patients severely suffer from high-frequency hearing loss.

Vocoder-based simulations have been extensively used to predict the performance of speech recognition for CI patients by imitating the electrical hearing received by patients [2, 5, 6]. During experiments, NH subjects are asked to recognize speech processed by a vocoder synthesizer, which is similar to the speech processor of the CI. The simulation results shall provide important pre-assessment information to doctors before directly conducting experiments on patients. Similarly, we evaluate the novel CI mentioned above by vocoder-based simulations in this paper. Based on the location setting of the 4 electrodes of this CI, we process the speech signal using a high-frequency vocoder instead of an all-frequency vocoder since the low-frequency acoustic hearing is preserved. Although the covered frequencies are different, we implement the vocoder based on the four-channel vocoder in [6]. We evaluate three spectral analysis strategies for the vocoder: linear-frequency distributed, logarithmic-frequency distributed, and low-map-high-frequency arrangement. The low-map-high-frequency arrangement means the low-frequency contents of speech are modulated to high frequency. These strategies affect the deployed locations of electrodes and speech coding procedures of the CI. The noise vocoder is often used to simulate CI-processed speech [7, 8]. Therefore, we also adopt the noise vocoder in our simulations.

Mandarin is a tonal language. It has four main tones, each of which has a unique fundamental frequency (F0) contour. These tones are helpful for distinguishing the word [9]. Although the tone recognition performance tends to be poor in most of the patients using conventional CIs [10], it should not pose a problem to our novel CI since patient's acoustic hearing is preserved. Simulations to mimic the acoustic hearing of HI patients show the tone recognition rates are higher than 90% [4]. Therefore, we only focus on intelligibility of phonemes. Since several studies discussed the intelligibility of vowel and consonant separately [11, 12], we follow their analysis meth-

http://dx.doi.org/10.21437/Interspeech.2017-858

ods to investigate the intelligibility improvement for vowel and consonant separately by using our novel CI.

The rest of this paper is organized as follows. In Section 2, we describe key elements of the psychoacoustic experiments. Experiment results are given and analyzed in Section 3. Conclusions and some discussions are given in Section 4.

2. Methods

2.1. Subjects and test materials

Nine normal-hearing native speakers (20-24 yrs old, 6 males and 3 females) were recruited for the Mandarin intelligibility tests. Each subject was asked to write down the words they heard. Each word was later divided into a vowel and a consonant. The sound of each word can be played repeatedly during tests.

In our experiments, five 25-Mandarin-word lists (A1, A2, B1, B2 and B3) in [13] were used for word recognition tests. Each Mandarin monosyllable word consists of three elements: two phonemes (initial consonant and final vowel) and a tone. Because recognition rates of tone are very high for processed speech by personalized HI hearing models [4], we only report intelligibility scores of vowel and consonant in this study. The intelligibility score was calculated as the ratio of the number of correctly identified vowels/consonants to the total number of played vowels/consonants. The sounds of all 125 Mandarin words for our listening tests were produced from the website (http://stroke-order.learningweb.moe.edu.tw/home.do?rd=72) developed by the Ministry of Education of Taiwan for learning Mandarin. All sounds were downsampled to 16 kHz sampling frequency and normalized to equal power. There were 5 test conditions: clean, with speech-spectrum shaped noise (SSN) of 0 and 4 dB SNR, with two-talker speech (TTS) maskers of 3 and 7 dB SNR.

2.2. Signal processing

All test sounds were processed through two parallel modules: the HI hearing model, which simulates the preserved acoustic hearing, and the vocoder, which simulates the provided electrical hearing. The block diagram of the process is shown in Fig. 1. Two processed signals from two parallel modules were summed together to generate final output speech for listening tests.

For the part of the HI hearing model, two sets of personalized parameters including MALs and BFs in Table 1 were used. These two patients (P1 and P3 in [4]) were selected because their high-frequency hearing is severely damaged such that they are better candidates for using our novel CI than the other patients. For the part of vocoder, input speech was passed through 4 band-pass filters (BPF 1 $\sim$ 4) distributed between 4000 and 8000 Hz. This frequency range was our hypothetical target for electrical stimulations. The envelope in each subband was extracted using a full-wave rectifier followed by a eight-order Butterworth low-pass filter (LPF) with the cut-off frequency of 400 Hz. Next, the white noise was band-pass filtered by the same BPF 1$\sim$ 4 to generate the four band-limited white noise carriers for carrying the extracted envelopes. All the amplitude-modulated noise signals from all BPFs were summed together to synthesize the high-frequency vocoded stimulus. Finally, the two processed speech signals from the HI model and the vocoder were combined to generate the output signal for listening tests.

Three kinds of spectral coverages of the vocoder BPFs, the logarithmic spacing (LOG), the linear spacing (LIN) and the low-map-high-frequency arrangement (LMH), were tested. For the LOG setting, 4 Butterworth filters were used with the center frequencies of 4367, 5199, 6182 and 7343 Hz and the bandwidth of 1/4 octave. For the LIN setting, 4 Butterworth filters were used with the center frequencies of 4500, 5500, 6500, 7500 Hz and the bandwidth of 1000 Hz. The same BPFs were used for extracting envelopes and generating carriers in both LOG and LIN settings. For the LMH setting, the BPFs for envelope extraction were centered at 1000, 3000, 5000 and 7000 Hz with the bandwidth of 2000 Hz to cover the whole speech spectrum, while the BPFs for generating carriers were the same BPFs in the LIN setting.

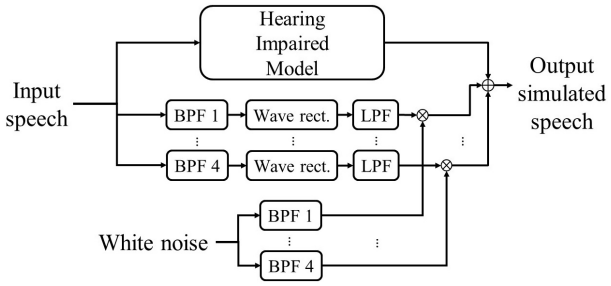

Figure 1: *Block diagram to generate test speech signals. The top branch comprises the HI hearing model, and the rest consists of a 4-channel high-frequency vocoder.*

Table 1: *MALs (in dB) and BFs of the better ear of patient 1 and patient 3 from [4].*

MALs (in dB)						
Subjects	**Age**	**250 Hz**	**500 Hz**	**1000 Hz**	**2000 Hz**	**4000 Hz**
P1	27	35	40	60	70	70
P3	36	50	50	40	65	90
BFs						
Subjects	**500 Hz**		**1000 Hz**		**2000 Hz**	
P1	2.85		3.29		3.19	
P3	3.2		2.95		4.66	

2.3. Psychoacoustic experiment procedures

Two sets of psychoacoustic experiments, simulations for determining spectral coverages of the BPFs and simulations for measuring vowel/consonant intelligibility improvement, were administered. For spectral coverage experiments, each subject participated in a total of 8 tests [= 2 personalized HI models $\times$ (no vocoder + vocoder with 3 spectral coverages)]. For intelligibility improvement experiments, each subject participated in a total of 16 tests [= 2 personalized HI models $\times$ 2 types of noise $\times$ 2 SNRs $\times$ 2 conditions (no vocoder + vocoder with the optimal spectral coverage)]. In each test, one 25-Mandarin-word list was randomly selected for generating test signals. From the results of the first set of 8 tests, we chose the optimal spectral coverage for the second set of 16 tests. That is, each subject were asked to recognize 200 and 400 words in the first and the second set of tests. During each set of tests, all processed words were randomly shuffled before presented to each NH subject via an AKG k702 headset in a semi-anechoic chamber (with a solid floor) at a comfortable volume between 65 dB and 75 dB SPL.

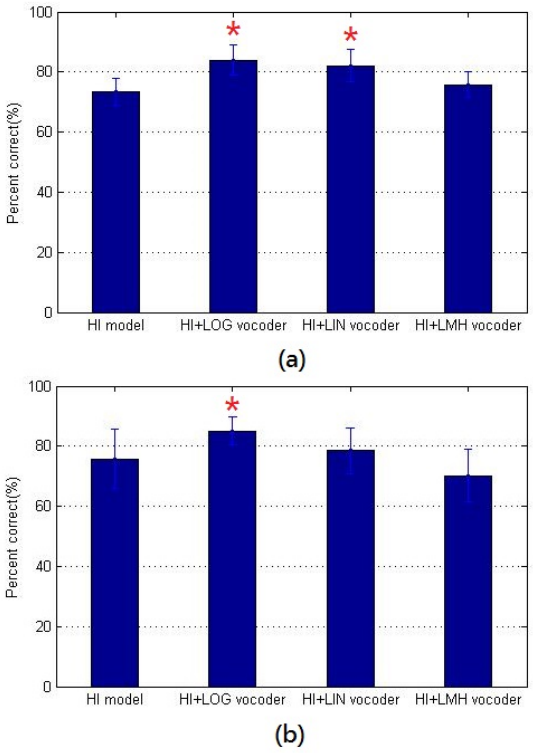

Figure 2: *Average recognition scores of phoneme (including vowel and consonant) for various types of spectral coverages using parameter sets of HI patient P1 (top panel) and P3 (bottom panel). ±1 standard deviation are superimposed on the bars. 'HI' is short for the hearing-impaired hearing model. The 'LOG', 'LIN' and 'LMH' respectively denote vocoders with logarithmic spacing, linear spacing and low-map-high arrangement. Asterisks '⋆' are placed on top of the bars whose scores are significantly ($p<0.05$) higher than scores of the condition of using HI model in paired comparisons.*

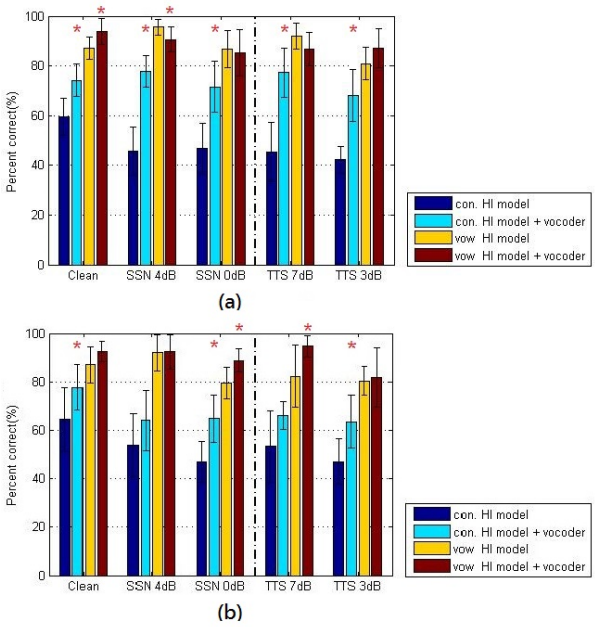

Figure 3: *Average recognition scores for vowel/consonant in each noise condition using the hearing models of patient P1 (top panel) and P3 (bottom panel). ±1 standard deviation are superimposed on the bars. The 'con.' and 'vow.' are respectively denote recognition scores for consonants and vowels. The LOG vocoder was used in this set of experiments. Asterisks '⋆' are placed on top of the bars whose scores are significantly ($p<0.05$) higher than scores of the condition of using HI model only.*

3. Experiment Results

3.1. Simulations for determining spectral coverages of the BPFs

The average recognition scores of phoneme (including vowel and consonant) for testing spectral coverages are shown in Figure 2. For statistical analysis, all recognition scores were further transformed to rational arcsine units (RAU) as the dependent variables by using rational arcsine transform [14]. The type of spectral coverage is the within-subject factor for statistical significance analysis.

Recognition scores for parameter sets of HI patient P1 are shown in Fig. 2 (a). The one-way analysis of variance (ANOVA) with repeated measures indicated significant effect from spectral coverage (F[3,32]=9.33, p=0.001). Post hoc analysis for paired comparison between each spectral coverage and HI model only showed that the scores of HI model combined with either the LOG or the LIN vocoder were significantly (both $p<0.005$) higher than those of the condition of HI model only. In contrast, the scores of HI model combined with the LMH vocoder were no significantly (p=0.2604) different from those of HI model condition.

Recognition scores for parameter sets of HI patient P3 are shown in Fig. 2 (b). Similarly, the one-way ANOVA with repeated measures indicated significant effect from spectral coverage (F[3,32]=5.87, p=0.0026). Post hoc analysis for paired comparison only showed that the scores of HI model combined with the LOG vocoder was significantly ($p<0.005$) higher than those of the condition of HI model only. The scores of HI model combined with either the LIN or LMH vocoder were no significantly (p=0.553, and p=0.1935) different from those of HI model condition.

3.2. Simulations for measuring vowel/consonant intelligibility improvement

From the spectral coverage experiments, the LOG vocoder was consistently shown beneficial to recognize phoneme. Therefore, only the LOG vocoder was adopted in this set of experiments. The average recognition scores for vowel and consonant in all noise conditions are shown in Figure 3. For statistical analysis, all recognition scores were further transformed to rational arcsine units (RAU). The SNR, phoneme type (vowel or consonant) and the vocoder presence are the three within-subject factors for statistical significance analysis. The scores of two groups of noise, SSN group (Clean, SSN 4dB and SSN 0dB) and TTS group (TTS 7dB and TTS 3dB), were discussed separately.

Recognition scores for parameter sets of HI patient P1 are plotted in Fig. 3 (a) against noise conditions. In the SSN group, the three-way ANOVA with repeated measures indicated significant effects from SNR (F[2,98]=6.07, p<0.005), phoneme type (F[1,98]=36.16, p<0.0005), and vocoder presence (F[1,98]=264.99, p<0.0005). The analysis also showed a non-significant interaction between SNR and phoneme type (F[2,98]=33.5, p=0.7343), a non-significant interaction between SNR and vocoder presence (F[2,98]=2.23, p=0.1134) and a significant interaction between phoneme type and vocoder presence (F[1,98]=33.57, p<0.005). Post hoc analysis of paired comparisons (with vocoder presence) showed that the scores by adding vocoder stimuli were significantly (p<0.005) higher than those without vocoder stimuli for consonant. Very similar to results in the SSN group, the three-way ANOVA with repeated measures also indicated significant effects from SNR (F[1,65]=8.4, p<0.05), phoneme type (F[1,65]=36.75, p<0.0005), and vocoder presence (F[1,65]=171.21, p<0.0005) in the TTS group. The analysis also showed a non-significant interaction between SNR and phoneme type (F[1,65]=1.1, p=0.2983), a non-significant interaction between SNR and vocoder presence (F[1,65]=0.04, p=0.8351) and a significant interaction between phoneme type and vocoder presence (F[1,65]=34.75, p<0.005). Post hoc analysis of paired comparisons (with vocoder presence) showed that the scores with vocoder stimuli were significantly (p<0.005) higher than those without vocoder stimuli for consonant.

Recognition scores for parameter sets of HI patient P3 are plotted in Fig. 3 (b) against noise conditions. In the SSN group, the three-way ANOVA with repeated measures indicated a significant effect from SNR (F[2,98]=10.41, p<0.005), phoneme type (F[1,98]=21.97, p<0.0005), vocoder presence (F[1,98]=209.34, p<0.0005), a non-significant interaction between SNR and phoneme type (F[2,98]=1.53, p=0.2212), a significant interaction between SNR and vocoder presence (F[2,98]=6.89, p<0.005) and a non-significant interaction between phoneme type and vocoder presence (F[1,98]=1.98, p=0.1627). Post hoc analysis of paired comparisons (with vocoder presence) showed that the scores by adding vocoder stimuli were significantly (p<0.005) higher than those without vocoder stimuli for consonant. Similar to results in the SSN group, the three-way ANOVA with repeated measures indicated a significant effect from SNR (F[1,65]=7.36, p<0.05), phoneme type (F[1,65]=20.81, p<0.0005), vocoder presence (F[1,65]=115.04, p<0.0005) in the TTS group. The analysis also showed a non-significant interaction between SNR and phoneme type (F[1,65]=1.28, p=0.2619), a non-significant interaction between SNR and vocoder presence (F[1,65]=1.72, p=0.1939) and a non-significant interaction between phoneme type and vocoder presence (F[1,65]=0.05, p=0.8317). Post hoc analysis of paired comparisons (with vocoder presence) again showed that the scores with vocoder stimuli were significantly (p<0.005) higher than those without vocoder stimuli for consonant.

4. Discussions and Conclusions

We conducted psychoacoustic experiments to predict the effects of acoustic hearing preserved CI by combining the hearing model of the HI patient and the high-frequency vocoder. The hearing model of the patient was used to simulate the preserved acoustic hearing while the vocoder was used to simulate the electrical hearing provided by the CI. We investigated three types of spectral arrangements for the vocoder and found the logarithmic arrangement of the BPFs significantly boosts the phoneme recognition rates of the two simulated HI patients. Results from the LMH arrangement demonstrated that warping the spectral distribution of envelope cues provides no benefit for recognizing phoneme.

Results of the second set of simulations, which assess vowel/consonant intelligibility improvement, showed that intelligibility scores of consonant were originally lower than scores of vowel with the HI model engaged. These results are not surprising since the selected patients have severe hearing loss in high frequencies, which affect consonants a lot more than vowels. Using the high-frequency LOG vocoder, a significant intelligibility boost for consonant was almost observed in each of the test conditions. On the other hand, intelligibility scores of vowel were originally high such that the benefit of using the LOG vocoder was not obvious for vowel. Overall speaking, intelligibility scores of phoneme were improved by utilizing the LOG vocoder.

In conclusion, this study uses vocoder-based simulations to predict phoneme intelligibility improvement by our novel CI can be expected. It validates the advantage of the acoustic hearing preserved CI. In addition, the HI hearing model plays an important role in those simulations which require the preserved acoustic hearing of the HI patient.

5. Acknowledgements

This research is supported by the Ministry of Science and Technology, Taiwan under Grant No MOST 105-2221-E-009-152-MY2 and the Biomedical Electronics Translational Research Center, NCTU.

6. References

[1] P. C. Loizou, "Introduction to cochlear implants," *IEEE Engineering in Medicine and Biology Magazine*, vol. 18, no. 1, pp. 32–42, 1999.

[2] J. Reefhuis, M. A. Honein, C. G. Whitney, S. Chamany, E. A. Mann, K. R. Biernath, K. Broder, S. Manning, S. Avashia, M. Victor *et al.*, "Risk of bacterial meningitis in children with cochlear implants," *New England Journal of Medicine*, vol. 349, no. 5, pp. 435–445, 2003.

[3] X.-H. Qian et al., "A bone-guided cochlear implant CMOS microsystem preserving acoustic hearing," in *Proc. Int. Symposium on VLSI Circuits*. IEEE, 2017, pp. C46–C47.

[4] P.-C. Tsai, S.-T. Lin, W.-C. Lee, C.-C. Hsu, T.-S. Chi, and C.-F. Lee, "A hearing model to estimate mandarin speech intelligibility for the hearing impaired patients," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5848–5852.

[5] R. V. Shannon, F.-G. Zeng, and J. Wygonski, "Speech recognition with altered spectral distribution of envelope cues," *The Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 2467–2476, 1998.

[6] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, p. 303, 1995.

[7] T. Green, A. Faulkner, and S. Rosen, "Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2155–2164, 2002.

[8] B. Roberts, R. J. Summers, and P. J. Bailey, "The intelligibility of noise-vocoded speech: Spectral information available from across-channel comparison of amplitude envelopes," *Proc. of the Royal Society of London B: Biological Sciences*, vol. 278, no. 1711, pp. 1595–1600, 2011.

[9] D. Klein, R. J. Zatorre, B. Milner, and V. Zhao, "A cross-linguistic pet study of tone perception in mandarin chinese and english speakers," *Neuroimage*, vol. 13, no. 4, pp. 646–653, 2001.

[10] D. Han, B. Liu, N. Zhou, X. Chen, Y. Kong, H. Liu, Y. Zheng, and L. Xu, "Lexical tone perception with hiresolution and hiresolution 120 sound-processing strategies in pediatric mandarin-speaking cochlear implant users," *Ear and hearing*, vol. 30, no. 2, p. 169, 2009.

[11] M. F. Dorman, P. C. Loizou, and D. Rainey, "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2403–2411, 1997.

[12] F. Chen, L. L. Wong, and E. Y. Wong, "Assessing the perceptual contributions of vowels and consonants to mandarin sentence intelligibility," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. EL178–EL184, 2013.

[13] K.-S. Tsai, L.-H. Tseng, C.-J. Wu, and S.-T. Young, "Development of a mandarin monosyllable recognition test," *Ear and hearing*, vol. 30, no. 1, pp. 90–99, 2009.

[14] G. A. Studebaker, "A rationalized arcsine transform," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 3, pp. 455–462, 1985.