Effects of Vocoder Processing on Speech Perception in Reverberant Classrooms

Qinglin Meng*, Guangzheng Yu*, Yangyang Wan†, Fanhui Kong#, Xianren Wang#, Nengheng Zheng†
* Acoustics Lab of School of Physics and Optoelectronics and State Key Laboratory of Subtropical Building Science, South China University of Technology, China
E-mail: mengqinglin@scut.edu.cn, scgzyu@scut.edu.cn Tel: +86-20-87113191
†College of Information Engineering, Shenzhen University, Shenzhen, China
E-mail: nhzheng@szu.edu.cn Tel: +86-755-22676617
Department of Otorhinolaryngology, the First Affiliated Hospital, Sun Yat-Sen University and Institute of Otorhinolaryngology, Sun Yat-Sen University, Guangzhou, China

***Abstract*— Channel vocoders have been widely used as acoustic models for current vocoder-centric cochlear implant (CI) signal processing strategies. Previous studies found that 4- to 8-channel vocoded speech in normal hearing (NH) subjects can derive comparable recognition scores to CI subjects who may have 8 to 22 frequency channels. The reasons for this overestimation might include 1) classic vocoders preserve natural intensity dynamic range within the envelopes rather than the much narrower range in CIs and 2) classic vocoders cannot simulate the effect of electric pulse rate. This study presents a novel vocoder based on a direct electric-pulse to acoustic-pulse mapping (i.e., electrodogram to spectrogram) to deal with the overestimation problem. The effects of the proposed vocoders with 22 and 16 channels on speech intelligibility in two real-measured classrooms were tested in NH listeners. Results showed that the proposed vocoders are more sensitive to changes of reverberant environment, like the previously reported actual CI results, than the classic ones, which implies that the new vocoding method could be a better alternative for acoustic modelling of current CI processing.**

I. INTRODUCTION

Homer Dudley from Bell Telephone Laboratory invented *Voder* the first apparatus to synthesize speech from a buzzer-like tone and a hiss-like noise in the 1930s. It imitated the source-tract effects of human speech production and was controlled by a single female expert to manually modify the pitch of source and spectrum distribution among ten frequency channels [1]. This idea was then widely used in the analysis-synthesis system for telecommunications, known as channel vocoder [2]. The channel vocoder takes advantage of the relative weak effect, in human hearing, of phase information (i.e., the temporal fine structure, TFS) on intelligibility and transmits a coarse spectrum envelope by temporal envelopes from multiple channels [3]. Historically, channel vocoder was replaced by subsequent vocoders which provide finer representation about the spectrum envelope (including the formant structure) and better sound quality., first in analog form and then in digital form [4].

Even though 1930s' channel vocoder was quickly replaced by advanced methods in telecommunication, 60 years later a channel vocoder like signal processing strategy, the continuous interleaved sampling (CIS; *Nature*, 1991) [5], came out as a breakthrough for multi-channel cochlear implants (CIs) and CIS-like strategies have been used by hundreds of thousands of patients. At the beginning decade of clinical application of multi-channel CIs, i.e. late 1970s to late 1980s, precisely encoding the formants of speech was thought to be important for intelligibility by some researchers [6, 7], with possible reasons from knowledge of linguistics or other speech engineering systems. Instead of explicit extraction of speech features, CIS only extract temporal envelopes from multiple bands (e.g., 6 bands in [5]). The envelopes are first sampled by carriers and then transmitted to corresponding electrodes. This processing is similar to the analysis part of channel vocoder and most current CIS-like strategies are vocoder-centric [8]. Different from the pulse-train or noise carrier of channel vocoder, nonoverlapped fixed rate electric pulse trains were used as carriers in CIS. This implicit temporal-envelope-based speech coding strategy of CIS, without any explicit feature extraction, has helped many CI users obtain good speech perception ability in quiet environment.

In a general sense, both *Voder* and CIS have demonstrated highly redundancy of speech signals and only temporal envelopes from multiple bands ($\geq$ 6), carried in amplitude by appropriate simple carries, can provide good speech intelligibility for normal hearing listeners or deaf patients. To further assess the role of spectral coarse degree in speech recognition and simulate CIS, Shannon *et al.* (1995) proposed a synthesis model, i.e., the noise-carrier vocoder, in which the extracted envelopes from all channel bands are multiplied by noises of the same bandwidths and then sum up to generate the synthesized speech [9]. They found that temporal envelopes from three bands is sufficient for speech recognition for normal hearing (NH) subjects. From then on, this noise-carrier vocoder and its younger sister sine-carrier vocoder [10], between which the main difference is on carrier selection (i.e., band-limited noise or sinusoidal signal), started to be widely used with NH subjects to simulate CI performance in many CI researches.

The basic information about these classic vocoders for CI simulation, *Voder*, channel vocoder, CIS strategy could be compared in Table 1.

However, previous studies showed that these classic

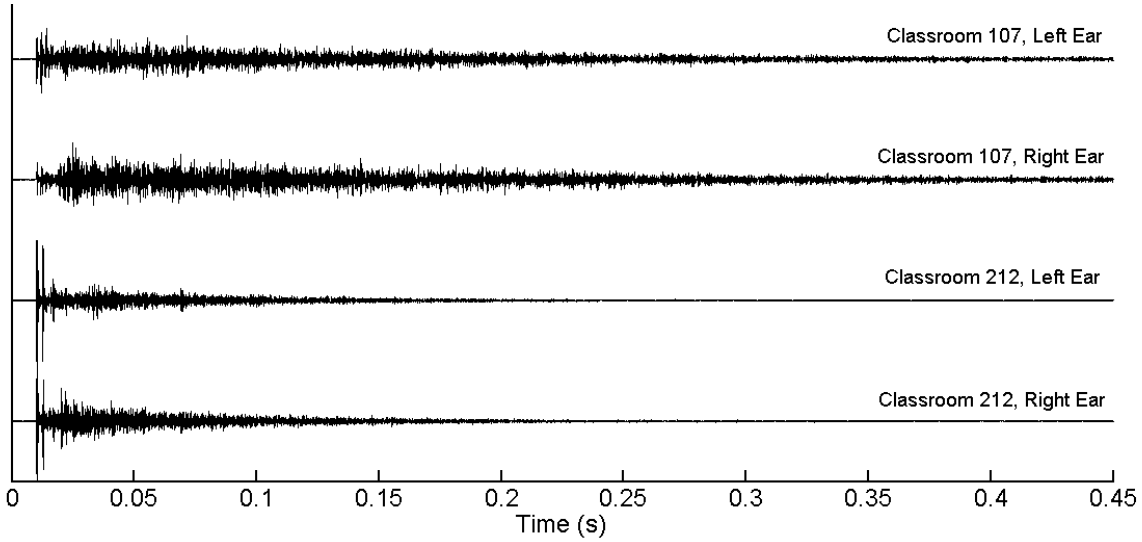

Fig.1 Binaural Room Impulse Responses measured in two classrooms

vocoders generally overestimate the performance of actual CI users [11, 12]. One possible reason was argued that CI subjects cannot make use of the provided spectral information [11]. There are some other reasons, which were easily ignored. For example, these classic vocoders have no control on the intensity dynamic range of the envelopes and cannot simulate the effect of different electric pulse rate setting. However, the relative narrower (compared with NH) acoustic input and electric output dynamic range and the pulse rate setting all have important effects on speech recognition with actual CIs [13].

Lu et al. (2007) proposed a Gaussian-enveloped tones (GET) based vocoder to simulate the pulsatile stimulation of CIs, which cannot be simulated by classic noise- or tone-exited vocoders [14, 15]. The GET vocoder have the potential to be a better alternative to classic ones. However, it was only used in some basic psychophysical experiments for binaural CI simulation [14, 16, 17].

In this study, we propose a new vocoder, which can directly transfer individual electric pulses to individual Gaussian-enveloped noise carrier or tone carrier acoustic pulses. That means any electrodogram (i.e., the time*electrode*intensity graph) from any CI system can be directly transfer to a sound consisting of many spectral-temporal pulses. The sound could be used to simulate that system with NH subjects. In this paper, we use this method to simulate the Advanced Combination Encoder (ACE) strategy [18], which is a default strategy of Cochlear company's product. It is a fast Fourier transform (FFT) based CIS-like strategy. One key difference between ACE and CIS is that ACE used a so-called *n*-of-*m* structure, i.e., for a *m*-channel CI system within each time frame only *n* channels with largest energy will generate effective pulses. Reverberant speech recognition was tested with the new vocoder and classic vocoder using noise carrier and 22 and 16 bands in NH subjects. The smearing effect was compared between conditions.

Table I. Several Famous Vocoder-centric Devices or Algorithms

Years	Name	Input	Output	Application
1930s	*Voder*	Human expert	Speech synthesis	First speaking machine
1940s	Channel Vocoder	Speech envelope and pitch analysis	Speech synthesis	Early telephone
1991-now	CIS Strategy	Speech envelope Analysis	Electric waveform	Multi-channel CIs
1995-now	Classic Vocoder	Speech envelope analysis	Speech synthesis	Acoustic model of CIS

II. METHODS

A. Binaural Room Impulse Responses (BRIRs)

We measured BRIRs in two classrooms (No.107: 10.5 m× 4.8 m×3.3 m; No. 212: 10.4 m×5.7 m × 3.3 m) of South China University of Technology. The excitation signal is a 16-stage maximal-length sequence (MLS), played through a D/A converter of a sound card (RME Fireface UC), a power amplifier (B&K 2716), and an omnidirectional source (B&K 4292-L, 12 loudspeakers) [19]. The source was placed on the platform, 2 meters from the wall, and 1.2 meters from the floor. Room impulse responses (RIRs) were firstly measured in the center of each room using a B&K 4191 microphone. Then, a KEMAR manikin with two B&K 4192 microphones was placed also at the center of the room and used to measure the BRIRs.

The reverberation time (T_{30}) at the range of 0.5-4 kHz estimated from the RIRs are 1.5 s for Classroom 107 and 0.5 s for Classroom 212. The BRIRs are illustrated in Fig. 1. Both T_{30} and the waveform of BRIRs indicate more severe reverberation in Classroom 107 than in Classroom 212. The convolution output of a sentence signal from Mandarin hearing in noise test (MHINT) [20] and the measured BRIRs are shown

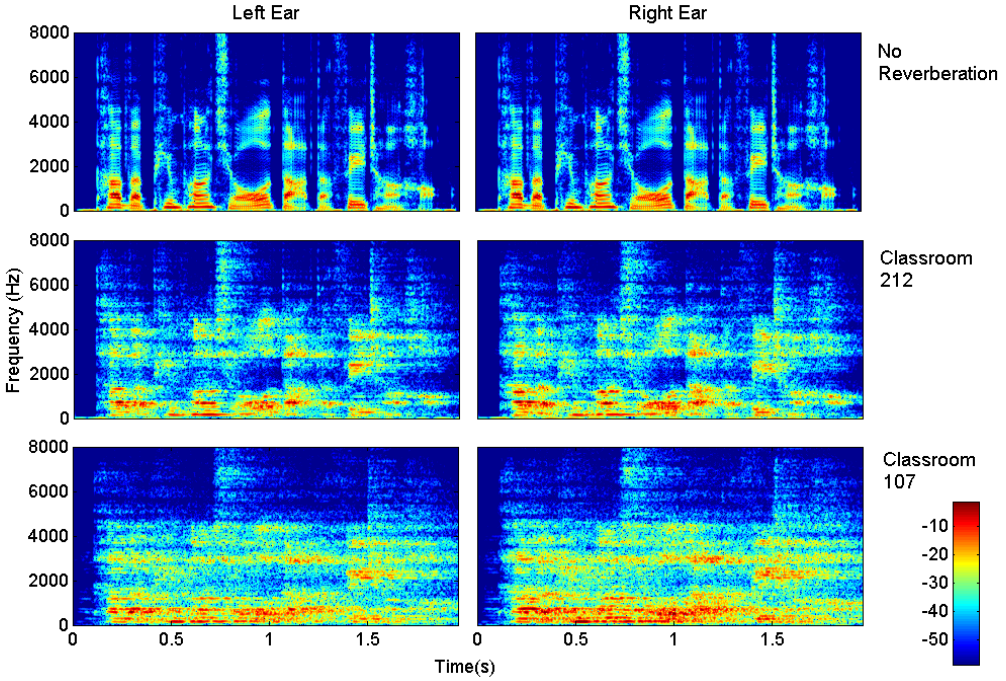

Fig.2 Convolution output of a sentence signal (top) and the measured BRIRs of Classroom 212 (middle) and 107 (bottom)

in Fig. 2. We can see that the acoustic smearing exists in both classrooms and is more severe in Classroom 107.

B. Vocoder Processing

Two types of vocoders were compared: a new pulsatile noise-excited vocoder and a classic noise-excited vocoder. Their specific implementations in this experiment are introduced here.

The new vocoder was used to simulate ACE strategy, a temporal envelope-based strategy for current Cochlear product users. In this experiment, the incoming sound, sampled at 16 kHz, was processed by a 128-point Hann window and FFT with a frame shift of 16 points (i.e., a 1 kHz pulse rate). No. 3 to 64 bins of the FFT output were divided into 22 or 16 bands. Specifically, number of bins allocated to individual bands, from low to high frequency, are (1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 5, 6, 7, 8) for the 22-band condition and (1, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5, 6, 7, 9, 11) for the 16-band condition. The power of each band is calculated and 8 bands with highest power were selected (the power of the rest bands are set as 0). These frame-wise power values represented within each band constitute the temporal envelope we need. Then the envelope within a 30-dB dynamic range was preserved by clip the possible peaks and mute the valleys and then compressed by a logarithmic function. The compressed values were linearly mapped to an electric current unit range from 100 to 255[1]. Then the synthesis part began. Firstly, an exponential function (i.e., the inverse function of the above logarithmic function) was used to transfer the electric current values to envelope power values. Then each power value P (corresponding to a single pulse at a specific time and a specific electrode) was used to generate a Gaussian-shaped envelope pulse (*GEP*), as in

$$GEP = Pe^{\frac{-\pi t^2}{D^2}} \quad (1)$$

where D is the effective duration of Gaussian envelope and was set as $D = 3/f_c$ with f_c the center frequency of corresponding band, t is set to be sampled from $-\frac{3D}{2}$ to $\frac{3D}{2}$ at the default sampling rate 16 kHz. Then all *GEPs* were assumed to occurred with a center time at the time of corresponding P. All *GEPs* along time within each band was connected including the many zero value points to generate an envelope curve. If there were coincidence between *GEPs* at a same sampling point, only the highest envelope value was preserved. Then the envelope curve was used to amplitude-modulate a band-limited noise, which was generated by summation of a set of sinusoidal signals distributed randomly (in frequency and initial phase) between the cutoff frequencies of corresponding band. In average, we used one sinusoidal signal per ten hertz. The average power of each band was kept unchanged. Finally, the modulated signals were summed up to get the vocoded stimulus.

For comparison, 22-band and 16-band classic noise-excited vocoders were used. The reverberant speech signals were splitted by 22 or 16 sixth-order Butterworth band-pass filters in frequency range of 80 to 7999 Hz into 22 or 16 band-

[1] This analysis part was implemented by using the code embedded in the CCiMobile platform, which is a CI signal processing research platform developed at CRSS-CILab, University of Texas at Dallas.

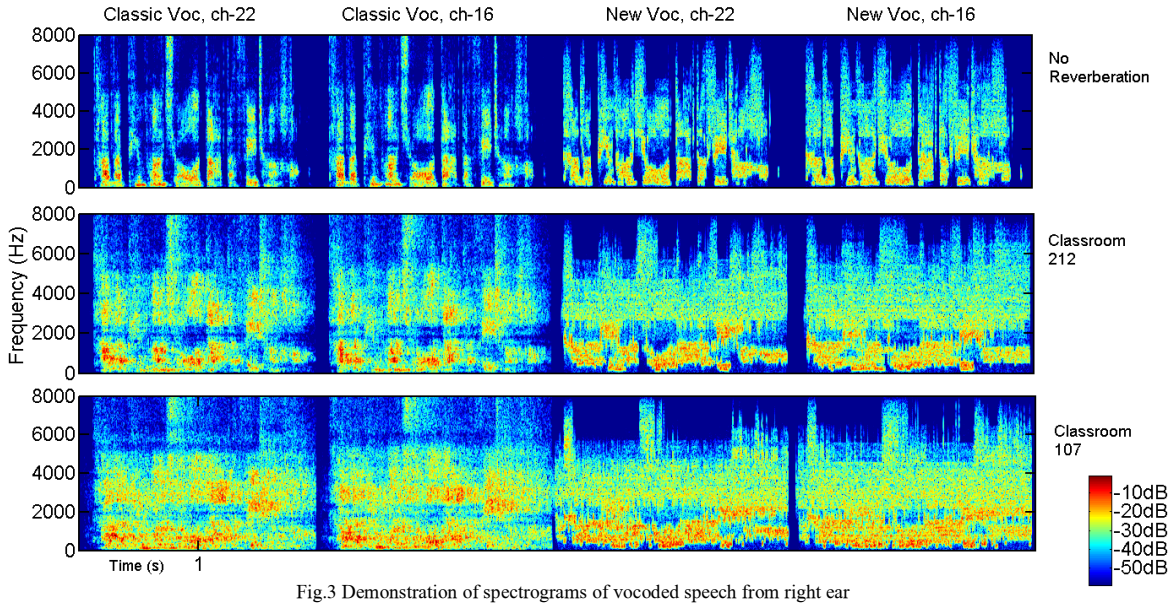

Fig.3 Demonstration of spectrograms of vocoded speech from right ear

limited signals. The cut-off frequencies of these filters were defined by equally dividing the basilar membrane according to the Greenwood function [21]. The temporal envelopes of these bandpass signals were extracted by a full-wave rectification and an eighth-order Butterworth low-pass filter. Then each envelope was multiplied by a band-limited noise, which is generated by passing a white noise to the corresponding bandpass filter used in the analysis part. Finally, the modulated signals from all bands were combined to get a vocoded stimulus.

Figure 3 shows some spectrograms of vocoded speech from right ear signals of Fig. 2. There are 12 conditions (2 vocoders × 2 channel numbers × 3 reverberation conditions). The acoustic smearing is more severe for Classroom 107 than Classroom 212. The intensity resolution is better with classic vocoders than new vocoders. What's more, because of the *n*-of-*m* feature of ACE, the new vocoders mute some time-frequency areas. All conditions were compared in the speech recognition experiment.

C. Speech Perception Experiment

Ten NH students from Shenzhen University participated in this experiment. Participation was compensated and all subjects provided informed consent in accordance with the local institution's review board. The speech material we used is the MHINT, which consists of 12 lists each with 20 sentences recoded by single male speaker. Each MHINT sentence includes 10 monosyllabic words.

The 12 lists were processed by the 12 conditions in a random order. The list order and the sentence order within each list were randomized as well. Each sentence was presented to subject up to three times and the subject was instructed to repeat as many words as possible. The stimuli were presented through an audio interface (Scarlett 2i4) and headphones (Sennheiser HD650).

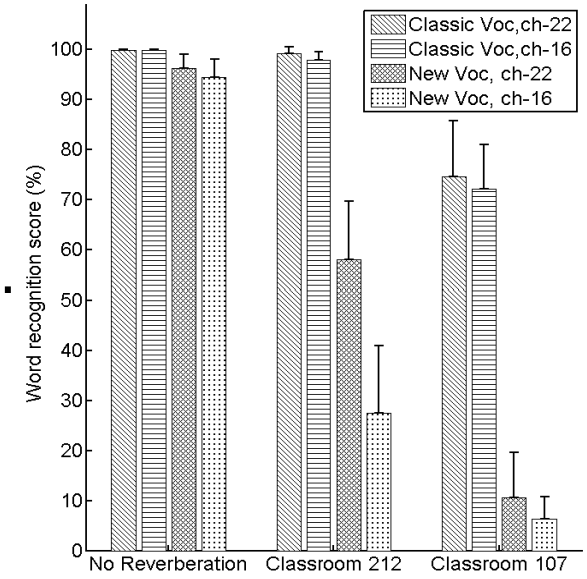

Fig.4 Word recognition scores for ten listeners as a function of reverberation and vocoder conditions

D. Results and Discussions

Speech recognition performance decreased with the new vocoder compared with the classic vocoder in all reverberation conditions (Fig.4). Two-way repeated measure analyses of variance (RM-ANOVA) showed more-reverberant room condition derived worse performance for both classic and new vocoders [$F(2,18) > 127.8$, $p < 0.001$]. However, for classic vocoder in Classroom 212, the scores are score to near saturation (100%). Pairwise comparison showed that for 22-channel classic vocoder, no significant difference was found

between results with no reverberation and with reverberation from Classroom 212 (p = 0.49). Pairwise t-test showed that, in Classroom 212, 16-channel vocoder derived significantly lower scores than 22-channel vocoder did for both vocoders. Especially, about 30% difference exists between 16- and 22-channel new vocoders, which implies that CI users need more frequency channels under reverberant conditions. Under both of the other two reverberation conditions, 16-channel condition derived statistically insignificant (p > 0/05) lower mean scores. In the most reverberant room (i.e., Classroom 107 with T_{30} = 1.5s), mean scores for classic vocoder is higher than 70%, which is much higher than both previous research [22, 23] and our observation of actual CI users. In that room, the mean scores for new vocoder is lower than 10.7%, which is nearly unusable for speech communication.

III. Conclusions

1. A new electric-pulse to Gaussian-shaped noise-excited acoustic pulse transformer based vocoder was proposed to imitate not only the spectral and temporal coarseness but also the low dynamic range and gaps between pulses of the CI electric stimuli.

2. Speech perception experiment showed that the new vocoder, compared with the classical noise-excited vocoder, is more sensitive to the variation of the reverberant condition. This effect is similar to that for actual CI users, which implies the new vocoder may be a better alternative to the classic one for future CI simulation research.

Acknowledgment

This work is jointly supported by NSF of China (Grant No. 11704129 and 61771320), the Fundamental Research Funds for the Central Universities (SCUT), State Key Laboratory of Subtropical Building Science (SCUT, Grant No. 2018ZB23), and Shenzhen Science and Innovation Funds (JCYJ 20170302145906843). Nengheng Zheng and Guangzheng Yu are corresponding authors.

References

[1] H. Dudley, "Remaking speech," *The Journal of the Acoustical Society of America,* vol. 11, pp. 169-177, 1939.

[2] B. Gold and C. Rader, "The channel vocoder," *IEEE Transactions on Audio and Electroacoustics,* vol. 15, pp. 148-161, 1967.

[3] M. R. Schroeder, "Vocoders: Analysis and synthesis of speech," *Proceedings of the IEEE,* vol. 54, pp. 720-734, 1966.

[4] A. S. Spanias, "Speech coding: A tutorial review," *Proceedings of the IEEE,* vol. 82, pp. 1541-1582, 1994.

[5] B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford, D. K. Eddington, and W. M. Rabinowitz, "Better speech recognition with cochlear implants," *Nature,* vol. 352, pp. 236-238, 1991.

[6] R. C. Dowell, P. M. Seligman, P. J. Blamey, and G. M. Clark, "Evaluation of a two-formant speech-processing strategy for a multichannel cochlear prosthesis," *Annals of Otology, Rhinology & Laryngology,* vol. 96, pp. 132-134, 1987.

[7] Y. C. Tong, G. M. Clark, P. M. Seligman, and J. F. Patrick, "Speech processing for a multiple‐electrode cochlear implant hearing prosthesis," *The Journal of the Acoustical Society of America,* vol. 68, pp. 1897-1899, 1980.

[8] P. C. Loizou, "Speech processing in vocoder-centric cochlear implants," in *Cochlear and brainstem implants*. vol. 64: Karger Publishers, 2006, pp. 109-143.

[9] R. V. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science,* vol. 270, pp. 303-304, 1995.

[10] 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, pp. 2403-2411, 1997.

[11] L. M. Friesen, R. V. Shannon, D. Baskent, and X. Wang, "Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants," *The Journal of the Acoustical Society of America,* vol. 110, pp. 1150-1163, 2001.

[12] Q. Fu and G. Nogaki, "Noise susceptibility of cochlear implant users: the role of spectral resolution and smearing," *Journal of the Association for Research in Otolaryngology,* vol. 6, pp. 19-27, 2005.

[13] F. Zeng, S. Rebscher, W. Harrison, X. Sun, and H. Feng, "Cochlear implants: system design, integration, and evaluation," *IEEE reviews in biomedical engineering,* vol. 1, pp. 115-142, 2008.

[14] T. Lu, R. Litovsky and F. Zeng, "Binaural masking level differences in actual and simulated bilateral cochlear implant listeners," *The Journal of the Acoustical Society of America,* vol. 127, pp. 1479-1490, 2010.

[15] T. Lu, J. Carroll and F. G. Zeng, "On acoustic simulations of cochlear implants," in *Conference on Implantable Auditory Prostheses (abstract)* Lake Tahoe, CA, 2007.

[16] A. Kan, C. Stoelb, R. Y. Litovsky, and M. J. Goupell, "Effect of mismatched place-of-stimulation on binaural fusion and lateralization in bilateral cochlear-implant users," *The Journal of the Acoustical Society of America,* vol. 134, pp. 2923-2936, 2013.

[17] M. J. Goupell, C. Stoelb, A. Kan, and R. Y. Litovsky, "Effect of mismatched place-of-stimulation on the salience of binaural cues in conditions that simulate bilateral cochlear-implant listening," *The Journal of the Acoustical Society of America,* vol. 133, pp. 2272-2287, 2013.

[18] A. E. Vandali, L. A. Whitford, K. L. Plant, and G. M. Clark, "Speech perception as a function of electrical stimulation rate: using the Nucleus 24 cochlear implant system," *Ear and hearing,* vol. 21, pp. 608-624, 2000.

[19] B. Xie, *Head-related transfer function and virtual auditory display*: J. Ross Publishing, 2013.

[20] L. L. Wong, S. D. Soli, S. Liu, N. Han, and M. Huang, "Development of the Mandarin hearing in noise test (MHINT)," *Ear and hearing,* vol. 28, pp. 70S-74S, 2007.

[21] D. D. Greenwood, "A cochlear frequency‐position function for several species—29 years later," *The Journal of the Acoustical Society of America,* vol. 87, pp. 2592-2605, 1990.

[22] K. Kokkinakis, C. Runge, Q. Tahmina, and Y. Hu, "Evaluation of a spectral subtraction strategy to suppress reverberant energy in cochlear implant devices," *The Journal of the Acoustical Society of America,* vol. 138, pp. 115-124, 2015.

[23] O. Hazrati, S. Omid Sadjadi, P. C. Loizou, and J. H. Hansen, "Simultaneous suppression of noise and reverberation in cochlear implants using a ratio masking strategy," *The Journal of the Acoustical Society of America,* vol. 134, pp. 3759-3765, 2013.