# Comparable Encoding, Comparable Perceptual Pattern: Acoustic and Electric Hearing

Fanhui Kong, Huali Zhou, *Student Member, IEEE*, Yefei Mo, Mingyue Shi, Qinglin Meng, and Nengheng Zheng, *Member, IEEE*

***Abstract*— Perception with electric neuroprostheses is sometimes expected to be simulated using properly designed physical stimuli. Here, we examined a new acoustic vocoder model for electric hearing with cochlear implants (CIs) and hypothesized that comparable speech encoding can lead to comparable perceptual patterns for CI and normal hearing (NH) listeners. Speech signals were encoded using FFT-based signal processing stages including band-pass filtering, temporal envelope extraction, maxima selection, and amplitude compression and quantization. These stages were specifically implemented in the same manner by an Advanced Combination Encoder (ACE) strategy in CI processors and Gaussian-enveloped Tones (GET) or Noise (GEN) vocoders for NH. Adaptive speech reception thresholds (SRTs) in noise were measured using four Mandarin sentence corpora. Initial consonant (11 monosyllables) and final vowel (20 monosyllables) recognition were also measured. Naïve NH listeners were tested using vocoded speech with the proposed GET/GEN vocoders as well as conventional vocoders (controls). Experienced CI listeners were tested using their daily-used processors. Results showed that: 1) there was a significant training effect on GET vocoded speech perception; 2) the GEN vocoded scores (SRTs with four corpora and consonant and vowel recognition scores) as well as the phoneme-level confusion pattern matched with the CI scores better than controls. The findings suggest that the same signal encoding implementations may lead to similar perceptual patterns simultaneously in multiple perception tasks. This study highlights the importance of faithfully replicating all signal processing stages in the modeling of perceptual patterns in sensory neuroprostheses. This approach has the potential to enhance our understanding of CI perception and accelerate the engineering of prosthetic interventions. The GET/GEN MATLAB program is freely available athttps://github.com/BetterCI/GETVocoder.**

***Index Terms*— Neural prosthetic, cochlear implant, aural rehabilitation, phoneme recognition, vocoder simulation.**

Manuscript received 19 November 2022; revised 13 February 2023 and 22 April 2023; accepted 25 April 2023. Date of publication 9 May 2023; date of current version 19 May 2023. This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2020A1515010386 and Grant 2022A1515011361, in part by the Science and Technology Program of Guangzhou under Grant 202102020944, in part by the Shenzhen Fundamental Research Program under Grant 20220809191805001, and in part by the 2020 Graduate Innovation and Development Fund Project of Shenzhen University. *(Corresponding authors: Qinglin Meng; Nengheng Zheng.)*

Fanhui Kong, Huali Zhou, Mingyue Shi, and Nengheng Zheng are with the Guangdong Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China (e-mail: nhzheng@szu.edu.cn).

Yefei Mo and Qinglin Meng are with the Acoustics Laboratory, School of Physics and Optoelectronics, South China University of Technology, Guangzhou, Guangdong 510641, China (e-mail: mengqinglin@scut.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2023.3274604, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2023.3274604

## I. Introduction

NEURAL prostheses restore or improve sensory perception for many patients by directly stimulating sensory neurons using a specifically designed series of devices. Among them, cochlear implants (CIs) are the most successful neural prostheses, which restore hearing abilities to about one million hearing impaired people [1], [2]. For people with normal sensory functions, either the impaired or artificially-restored sensation cannot be directly perceived. The challenges faced by patients after prosthesis intervention are also difficult to comprehend. This is a limitation on demonstration and education about the neuroprostheses. In addition, the engineering and development of prostheses need simulation tools to effectively predict the performance of various implementations for target patients.

To meet these needs, simulators of sensory (e.g., auditory and visual) impairments and prostheses have been proposed [3], [4], [5], [6], [7], [8]. Prostheses for different sensory domains (e.g., vision and audition) are designed to mimic healthy neural pathways, as are their simulators. Simulators for prostheses often adopt similar principles during their development, however, comparative studies of various sensory prostheses are limited in number. In 2007, Hallum et al. and colleagues performed a review comparing auditory and visual prostheses [9]. While there have been hundreds of papers using CI acoustic models since the 1980s, image models for visual prostheses have been few [6], [7], [9]. The authors concluded with an outlook on the modeling of visual prosthesis with the background of both the well-known advantages and limitations of acoustic modeling of CIs [9].

In this paper, we propose a critical principle needed for the development of simulations of protheses: the encoding stages of the simulator and the encoding stages of the protheses should be mirror, in that, the quality of the signal

characteristics should be the same or as close as possible. We argue that this critical principle could help to minimize the performance discrepancy between actual and simulated implantees. To examine it, CIs and CI simulators were studied.

## II. Related Works

In the 1990s, open speech communication with CIs became possible due to advances in signal processing. One important milestone is the continuous interleaved sampling (CIS) strategy [10]. The success of CIS was attributed to an effective combination and exact implementation of a series of signal processing stages [11]. Input microphone signal is filtered into a few frequency bands covering a wide range of frequencies which are thought to be important for speech intelligibility; Second, temporal envelopes from each band are extracted and compressed to fit into the electrical dynamic range of individual users; then, the compressed envelopes are used to modulate the amplitudes of high-rate biphasic charge-balanced electric pulses for corresponding electrode stimulation. Among bands, the electric pulses are non-simultaneously fired, i.e., in an interleaved sampling manner. The CIS strategy coarsely mimics the frequency analysis of the basilar membrane and allows for simplification and optimization of an engineered implementation.

CIS abandoned the philosophy of explicitly extracting phonetic cues, such as fundamental frequency (F0) and formants. Instead, temporal envelopes from fixed bands are extracted where phonetic cues are found to be implicitly represented. For example, F0 can be found in the periodicity on the envelope [12] and formants can be estimated by comparing the relative power of all bands [13]. More recent strategies have inherited key features of CIS, and the approach is still an option for almost all modern CI products.

Modern CIs are still far from perfect, which means CI users face challenges in many sound perception tasks. Simulators of CIs have been investigated throughout the history of CI technology development, e.g., [14], [15], [16]. They were proposed to model the electric hearing of CIs recipients using acoustic stimuli presented to normal hearing (NH) people. Electric hearing refers to the perception of sound through the stimulation of the auditory nerve with electrical currents of CI, rather than through air vibrations.

The most widely used CI simulators are temporal-envelope-based vocoders [15]. In these vocoders, temporal envelopes from multiple bands are extracted and then directly used to modulate band-limited noise or sine-wave carriers. The modulated carriers are summed up to obtain the output signal for stimulation. Because of the similarities in temporal envelope extraction, these kinds of vocoded sounds were assumed to be transmitting similar information as CI devices. This is a widely accepted assumption and consistent performance trends between these two hearing modes were often reported [8], [17]. The advantage of these conventional continuous carrier vocoders is obvious: They can be used to predict the overall trend of intelligibility with CIs.

However, there are significant simulation-to-real disparities in absolute perceptual test results [8], [18]. Some degradation methods with good physiological hypotheses can be used to decrease the disparities. One solution is to adjust the degree of current spread in the vocoder [19]. The current spread between CI electrodes refers to the distribution of electrical stimulation delivered by the electrodes to the auditory nerve fibers in the cochlea. This current spread can affect the frequency resolution, potentially causing speech degradation. Another solution is to include shallow insertion or frequency shift in the simulations [20].

Instead, here we focus on encoding or signal processing implementations of CI strategies. According to the proposed critical principle, the encoding stages of a CI simulator and the encoding stages of a target CI strategy should be the same or as close as possible. However, signal processing of conventional temporal-envelope-based vocoders and CIS strategy as described above have many obvious differences. They may share the same envelope extraction methods, but the latter stages, including envelope compression and pulsatile carrier modulation of CIS, are not included in the processing stages. This could be another reason behind the simulation-to-real disparities.

We developed an advanced simulator, i.e., the GET vocoder in [21], and proposed a variant Gaussian-enveloped Noise (GEN) in current work. The idea is to transform the pulsatile CI electric stimuli directly into an acoustic sound in a pulse-to-pulse manner. Technically, GET/GEN vocoded sound could maintain the same information as a CI strategy, as they use the same process to encode the original sound. To learn more about the connection or difference between GET and other related vocoders, please see [21] for an in-depth introduction. In the preliminary studies [21], [22], the theoretical analysis based on the time-frequency uncertainty principle was done[1] and the advantage of GET on a single SRT test has also been verified.

In this study, the GET vocoder and its newly proposed variant GEN vocoder, are used to examine the proposed critical principle for prosthesis simulators. We argue that, following the proposed critical principle, the GET/GEN vocoder could derive comparable perceptual patterns in multiple tasks across NH and CI listeners in SRTs as well as consonant and vowel confusion rates. We selected CI users who used the advanced combinational encoder (ACE) strategy in their clinical processors as the target to simulate, as this kind of processor is estimated to be used by half of global users [2]. In Sec. III, several algorithms are introduced, i.e., a standard ACE strategy, the GET vocoder and the variant GEN vocoder for ACE-CI simulation, and the control simulators (i.e, conventional vocoders with continuous sine-wave carriers and current spread manipulation). In the following sections, the speech intelligibility simulation performance of GET/GEN is systematically validated. To conclude, CI simulators and their implications on general neuroprosthesis simulations are further discussed. The GET/GEN MATLAB program is freely available at https://github.com/BetterCI/GETVocoder.

[1]According to the time-frequency uncertainty principle, it is not possible for acoustic GET or GEN pulses to be extremely short. There is a compromise between the temporal duration of a signal and its frequency content. This limits the effectiveness of GET/GEN in simulating electric pulses with microsecond widths in CI [21].

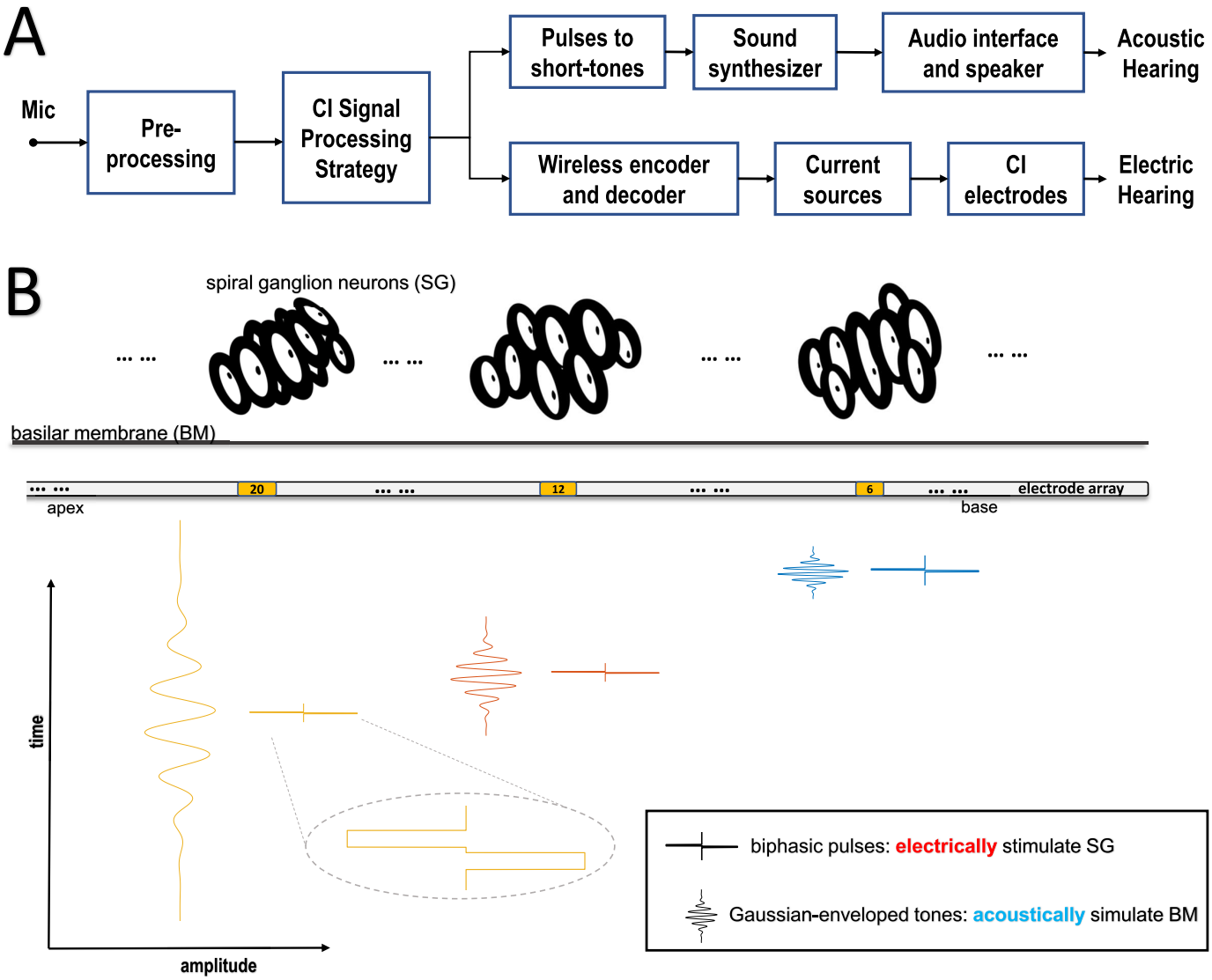


Fig. 1. Conceptualization of the engineering framework of this study. A: Signal processing flowchart for the acoustic and electric hearing; B: Cochlear stimulation interface and the stimulus atoms for the acoustic hearing and electric hearing. Acoustic hearing and electric hearing represent simulated and actual CI hearing, respectively.

## III. Algorithms and Models

The proposed framework for a comprehensive simulation of CI processing is conceptually demonstrated in Fig. 1. According to Fig. 1. A, the incoming sound captured by a microphone is pre-processed and then passes through a CI signal processing strategy. For a real CI, the information to be represented at the electrodes has been determined by the strategy, so the signal then branches and is delivered to an NH acoustic ear (by sound synthesizing and electroacoustic devices) or a CI electric ear (by CI implant device and electrodes). The GET and GEN vocoders follow the acoustic branch and their difference is only at the carrier signals for synthesizing. They simulate individual electric pulses using individual acoustic pulses with the assumption of delivering the same sound information to the same group of neurons (see Fig. 1. B). They were used in NH listeners to simulate perception of CI users using the ACE strategy and was compared with actual ACE CI users. Conventional sine-wave vocoders with current spread manipulation were also used as control.

### A. ACE Strategy[2]

ACE and CIS processing share the following features such as band-pass filtering, temporal envelope representation, and interleaved sampling.Their main difference is that ACE uses an $n$-of-$m$ maxima selection, i.e., in each time frame among the total $m$ ($\leq$ 22) bands only $n$-maxima (default: 8) with the highest energies generate electric pulses to corresponding electrodes [22], [23].

The incoming sound $x$ is sampled at a sampling rate $fs$ = 16000 Hz. It is pre-emphasized by a high-pass filter:

$$y[n] = 0.5006x[n] - 0.5006x[n-1] + 0.0012y[n-1]. \quad (1)$$

[2]This study adopted the MATLAB code for the ACE strategy from the publicly available code of the CCi-Mobile research platform, which can be found at https://Github.com/CILabUTD/CCi-MOBILE

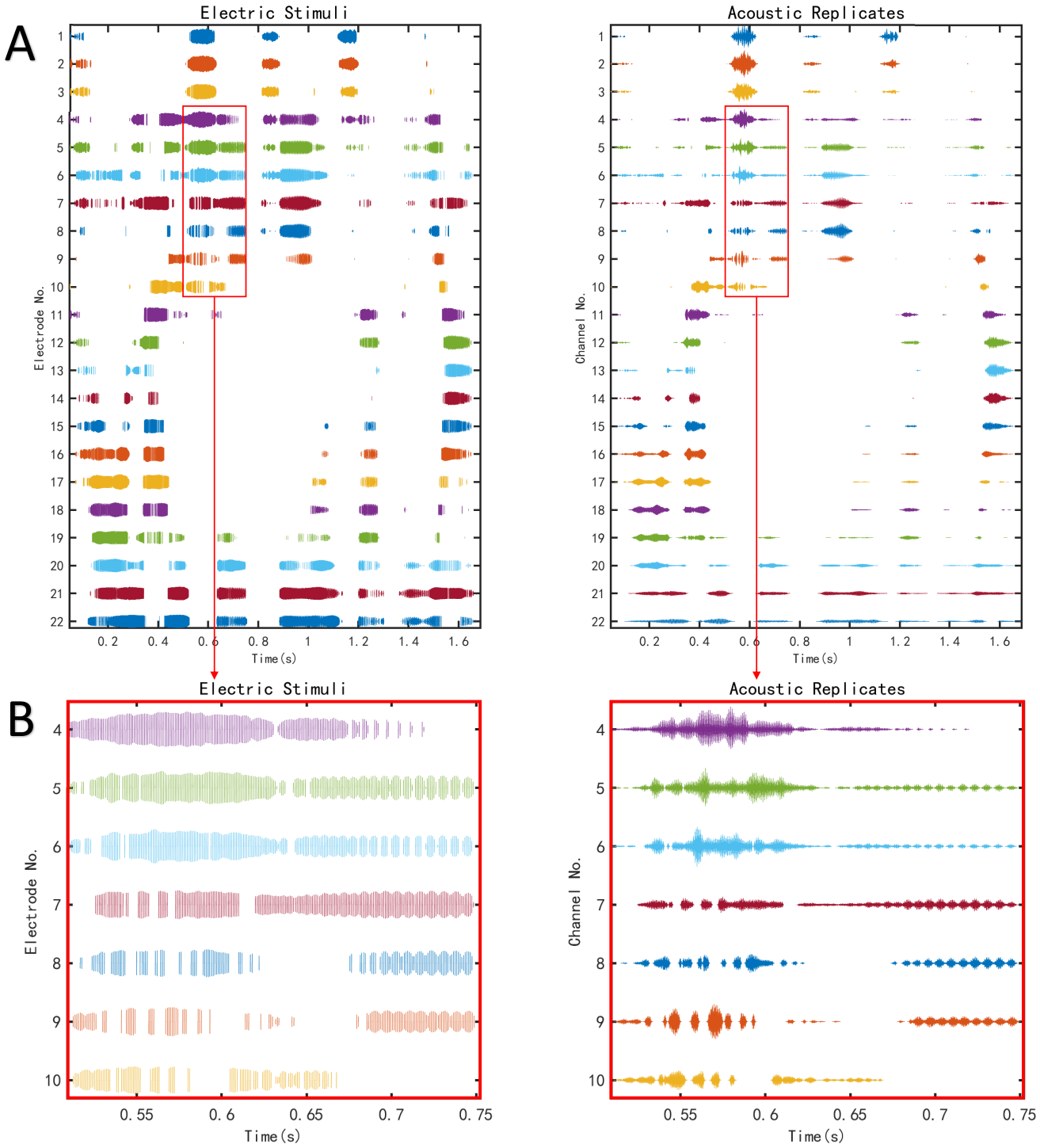


Fig. 2. Demonstration of the acoustic (GET) and electric (ACE) stimuli. The electric stimulus was generated from a Mandarin sentence speech using an advanced combinational encoder (ACE) CI strategy. The right acoustic band signals were generated by transforming the left electric signal using a Gaussian-enveloped Tone (GET) vocoder. A: all band signals; B: Inset local band signals. The "atoms" for the left electric and right acoustic stimuli are the bi-phasic current pulse and Gaussian-enveloped tone demonstrated in Fig. 1.

Then $y$ is processed in an overlapped frame manner. Each frame, denoted by $\mathbf{v}[n]$ has 128 sampling points (corresponding to 8 ms; $n \in \{0, 1, 2, \ldots, 127\}$) windowed by a Hanning window. The frameshift is determined by a pre-selected stimulation rate. In this case, the stimulation rate $r$ was 900 pps (pulse-per-second), so the frameshift was $s = \lceil fs/r \rceil = 18$ points. Discrete Fourier transform (realized by a fast Fourier transform, FFT) is used to transfer $\mathbf{v}[n]$ to $\mathbf{V}[k]$ at the frequency domain and only the left half of the symmetric bins is preserved:

$$\mathbf{V}[k] = \sum_{n=0}^{127} \mathbf{v}[n] \cdot \mathrm{e}^{-\frac{i2\pi}{127}kn}, \quad k \in \{0, 1, 2, \ldots, 64\}. \tag{2}$$

The bins between $k = 2$ and 63 were combined into 22 bands, i.e., bin numbers of low-to-high bands are [1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 5, 6, 7, 8]. The power of each band is calculated by:

$$\mathbf{u} = \sqrt{\mathbf{w}[|\mathbf{V}[0]|^2, |\mathbf{V}[1]|^2, \ldots, |\mathbf{V}[64]|^2]^T}. \quad (3)$$

The weight matrix $\mathbf{w} \in \mathbb{R}^{22\times 65}$) is determined by the frequency response of the window function. For each frame, $\mathbf{u}$ has 22 values, among which the 14 lowest values are rejected (i.e., set to zero). In this way, the so-called "$n$-of-$m$" is realized. Then the preserved maxima $\mathbf{u}'$ are compressed according to

$$\mathbf{g} = \frac{\ln(1+\alpha\mathbf{u}')}{\ln(1+\alpha)}, \quad \alpha = 416.0. \quad (4)$$

Then $\mathbf{g}$ is quantized with eight bits into the range between threshold (T) and comfortable (C) current levels. The quantized $\mathbf{g}'$ is then used to control the magnitude of the individual current-balanced bi-phasic electric pulse (phase duration: 25 $\mu s$; inter-phase gap: 8 $\mu s$, see the inset in Fig. 1). The ACE-electrodogram[3] of a sentence speech is demonstrated in the left panels of Fig. 2.

### B. Gaussian-Enveloped Tone/Noise Vocoders

We have argued that faithfully replicating the signal processing stages in a CI strategy is critical for a good simulation of CI. The GET vocoder [21] meets this requirement. In this study, we proposed a new variant of GET, i.e., the GEN vocoder. The ACE strategy is used in our study. Each electric pulse calculated using the ACE strategy is directly mapped into an envelope pulse according to:

$$p_{n_0,t_0}(t) = g_{n_0,t_0} e^{-\frac{\pi(t-t_0)^2}{2\sigma^2}}, \quad (5)$$

where $n_0$ and $t_0$ denote the band number and time center of the electric as well as the consequent acoustic pulse, $g_{n_0,t_0}$ is an element of $\mathbf{g}$ recovered from $\mathbf{g}'$, and the standard deviation $\sigma$ is used to control the duration of the acoustic pulse and equaled to $3/f_c$ in our experiments.

To synthesize a sound for simulation, an envelope pulse $p_{n_0,t_0}$ with unit amplitude (i.e., let $g_{n_0,t_0} = 1$ in Eq. 5) is used as an impulse response of the electric stimuli. Therefore, convolution is used to transform the electric band signal to an envelope band signal $P_{n_0}(t)$. In this convolution, a traditional simplified CI electrodogram was used rather than that demonstrated in Fig. 2. This means that individual electric pulses were represented by a $\delta$ function (i.e. a single vertical line) rather than the detailed bi-phasic waveform. The final acoustic band signal is

$$A_{n_0}(t) = P_{n_0}(t)\sin(2\pi f_c t + \varphi_0) \quad (6)$$

where $f_c$ is the center frequency of the current band and $\varphi_0$ is an arbitrary initial phase and equaled to zero in our experiments. The band signals for the acoustic sound are demonstrated in the right panels of Fig. 2. The band signals are then summed to generate a GET-vocoded sound.

Although there is a "tone" in the name of GET, intrinsically the carrier is not limited to sine waves. Here, we add another possibility, i.e., the noise carrier from GEN vocoders. This implementation method simply replaces the carrier signal in Eq. (6) according to

$$A_{n_0}(t) = P_{n_0}(t)\sum_{l=1}^{L_{n_0}} \frac{1}{L_{n_0}}\sin(2\pi f_{n_0,l} t + \varphi_{n_0,l}), \quad (7)$$

where $L_{n_0} = \lceil 0.1 B_{n_0} \rceil$ (i.e., 0.1 times the bandwidth of the No. $n_0$ band) and then rounded up to a whole number. In specific, $L_{n_0}$ =13, 13, 13, 13, 13, 13, 13, 13, 13, 25, 25, 25, 25, 38, 38, 50, 50, 63, 63, 75, 88, and 100 for the 22 bands from low to high in our experiments. For individual sine waves in Eq. (7), the frequency $f_{n_0,l}$ was a random value uniformly distributed in the corresponding frequency band; the initial phase $\varphi_{n_0,l}$ was a random value uniformly distributed in the range of $[0, 2\pi]$.

### C. Conventional Sine-Wave Vocoders With Current Spread Manipulation

To evaluate the potential superiority of GET or GEN, conventional vocoders [15], [16] should be used as controls. The current spread around individual CI electrodes may interact with one another, and the naive implementation in the conventional vocoders without current spread simulation often overestimates CI performance. Previous studies have shown that certain degrees of current spread simulation could be used to approximate average CI performance on speech-in-noise intelligibility [19]. The simulation of the assumed spread of current from each electrode to remote neurons can be performed by summing the weighted energies of the envelopes of each channel to the energies of other channels. Here, we implemented conventional sine-wave vocoders with current spread simulation as follows.

The incoming sound $x[n]$ is sampled at a sampling rate $fs$ = 16000 Hz. A pre-emphasized signal $y[n]$ is acquired by Eq. (1). Then a bank of 22-band $6^{th}$-order bandpass Butterworth filter is used to process $y$. The cutoff frequencies are equal to the FFT-based bandpass filters in the ACE strategy. The $k^{th}$ band output $v_k[n]$ is full-wave-rectified by $V_k[n] = |v_k[n]|$. Then it is filtered by an $8^{th}$-order low-pass Butterworth filter with a cutoff frequency of 250 Hz. The filter output is a temporal envelope of the $k^{th}$ band, denoted as $E_k[n]$. The current spread degree in dB/octave is denoted by $w$. The spectral smearing effect is realized by

$$E'_k[n] = \sqrt{\sum_{m=1}^{22} 10^{-w|\log_2(f_m/f_k)|/20} E_m^2[n]}, \quad (8)$$

in which $f_m$ represents the center frequency of the $m^{th}$ band. In reference to the literature [19], [24], in our experiments, we used $w$ = 8, 10, and 12 dB/octave. Then the vocoded

[3]In this article, an electrodogram displays the amount of electrical stimulation given to each electrode graphically, and a spectrogram visually portrays the distribution of energy across various frequencies in a signal as it changes over time.

TABLE I
CI PARTICIPANT DEMOGRAPHICS

| ID | Gender | Age at testing | Etiology | Processor | CI experience (yr) | EDR[b] (CL) |
|---|---|---|---|---|---|---|
| C1 | F | 25 | Congenital | CP810 | 21 | 57.7 |
| C2 | M | 28 | Drug induced | CP950 | 19 | 64.0 |
| C3 | F | 24 | Drug induced | CP900 | 20 | 84.9 |
| C14[a] | F | 44 | Drug induced | CP810 | 15 | 96.1 |
| C26 | M | 24 | Meningitis | Freedom | 15 | 78.1 |
| C48 | F | 23 | Enlarged Vestibular Aqueduct | CP900 | 6 | 55.7 |
| C58 | F | 31 | Unknown | CP900 | 2 | 48.6 |
| C59 | M | 51 | Unknown | CP802 | 3 | 30.4 |
| C60 | M | 18 | Unknown | N7 | 17 | Unknown |
| mean | – | 30 | – | – | 13 | 64.4 |

*a*: the only bilateral user, tested unilaterally with her preferred ear.
*b*: mean electrical dynamic range among all active electrodes.

sound signal is generated by

$$O[n] = A \sum_{m=1}^{22} E'_k[n] \sin\left(2\pi f_m \frac{n}{fs} + \phi_{0m}\right) \tag{9}$$

in which $A$ is adjusted to keep the root-mean-square of the whole signal unchanged after the spectral smearing.

### D. Spectrograms of Vocoded Sounds

One mandarin sentence from the MSP corpus was processed by six algorithms, i.e., ACE, GET50, GET150, GEN50, Sine12, Sine10, and Sine8. The electrodogram or spectrograms of these output signals are illustrated in Fig. 3. The six algorithms were used in two experiments (ACE, GET50, and GET150 in the first; ACE, GET50, GEN50, Sine12, Sine10, and Sine8 in the second).

## IV. MODEL VALIDATION

### A. Overview

Neural prostheses are expected to improve sensation performance in all aspects of the corresponding sensation rather than in a single task. Here, we examine the proposed critical principle for effective prosthesis simulators in CIs, i.e., faithfully following the signal processing details of the prosthesis as critical for obtaining the best simulation performance in multiple tasks.

The framework of the recently proposed GET vocoder adheres to this critical principle in its approach to simulation. It could derive similar means and variances in SRTs in noise with a sentence corpus as actual CIs [21], [22]. In this work, we designed a GET variant (i.e., GEN) and developed a systematic test battery to validate the GET/GEN model. Training effects (Exp. 1), multiple sentence corpora (Exp. 1 & 2), and vowel and consonant confusion (Exp. 2) were included in two experiments. The algorithms in Sec. III were all tested.

If one vocoder outperforms the others in most tasks in the battery, it would be recognized as optimal among the included vocoders. In total, nine CI users (see Table I) and 66 NH listeners who listened to vocoded speech were recruited. NH participants all self-reported no otological disease or hearing loss. The total number of person-visit-time conducted in the study was 242. In terms of time, the total number of hours spent on these experiments was about 125 hours. All participants received financial compensation or credits for their participation. Written informed consent was obtained from all participants before the experiment, and all procedures were approved by the ethical review board at Shenzhen University.

### B. Experiment 1: SRT in Noise-Training Effects and Multiple Corpora

Difficulty in speech perception in noise is one of the most common complaints of the CI listening experience. The signal-to-noise ratio (SNR) at which one listener could recognize a certain percentage (e.g., 50%) of the words in a sentence is defined as an SRT. To quantitatively estimate speech-in-noise perception ability, SRTs with target sentences in stable or babble noises are usually measured through an adaptive psychophysical procedure [25].

In two previous studies by the authors [21], [22], we have measured SRTs with a GET simulation using a single corpus. Specifically, in [22] we evaluated the effects of the electrical dynamic range (EDR) of CIs with both actual CI users and GET vocoders. EDR is defined as the difference between the comfortable (C) level and detection threshold (T) level at individual electrodes for individual implantees. Three EDRs were compared, i.e., 30, 100, and 150 CL (current level; defined by a CI manufacturer). We found that narrower EDR in the perceivable range would lead to worse speech performance (i.e., higher SRT in noise). GET simulation with EDR = 150 CL derived comparable SRTs to actual CI users with much narrower EDRs (typically 20-80 CL). We argue that this discrepancy could be attributed to the listening experience difference. For newly implanted CI users, training and rehabilitation may take months or longer to help their brains learn to use this new type of stimuli used to represent sound [26]. CI listeners had months or years of CI listening experience, while NH participants are naive listeners to the vocoded sound with limited training experience during the experiment itself.

In this experiment, GET vocoders with EDR = 50 and 150 CL were used in two groups of NH participants. On a daily basis, SRTs were measured using four different corpora. The NH participants visited the laboratory for five consecutive days and the CI participants visited only once. It was hypothesized that using similar strategy parameters encoded speech could derive similar intelligibility with any corpus for both electric CI and acoustically simulated CI listeners, given enough training.

*1) Methods:* Nine CI users (see Table I) and 16 NH listeners (college students; age range: 20 to 26; 9 males) were recruited in Experiment 1.

CI users were all tested with their daily used processor and ACE strategy (Sec. III-A). NH listeners were tested with GET (Sec. III-B) vocoded sound delivered through a sound card and headphones. ACE and GET algorithms were implemented as described in Sec. III. The clinical EDRs of the CI users were in the range of 30-100 CLs. For NH with GET vocoded sound, EDR = 50 CL and 150 CL were tested in two groups (each with eight listeners).

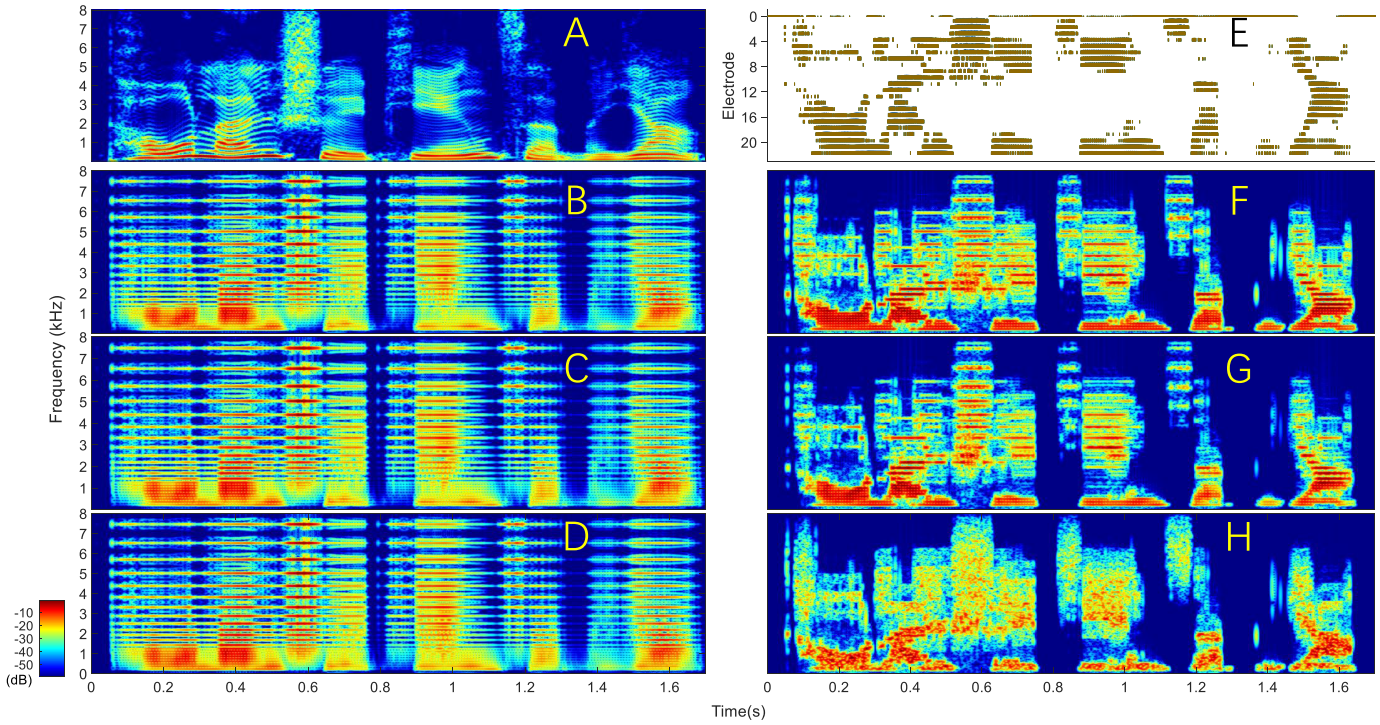


Fig. 3. Demonstrations of the spectrograms and electrodogram of the stimuli used in this study. The original speech is a sentence from the MSP corpus. The sentence is 投篮水平怎么样 (Pinyin: tóu lán shuǐ píng zěn me yà ng; meaning: How is the shoot level?) The spectrogram of the original speech is shown in A.The ACE strategy was used to generate the electrodogram (in E). The six vocoders, i.e, F. GET50, G. GET150, H. GEN50, B. Sine12, C. Sine10, and D. Sine8 were used to generate the vocoded sounds. (GET50: Gaussian-Enveloped Tones with EDR = 50 CL; GET150: Gaussian-Enveloped Tones with EDR = 150 CL; GEN50: Gaussian-Enveloped Noise with EDR = 50 CL; SineX: conventional sine-wave vocoders with current spread degree of X = 12, 10, and 8 dB/octave.)

The four Mandarin sentence corpora were Mandarin speech perception (MSP) [27], Mandarin hearing in noise test (MHINT) [28], an in-house corpus developed by our group in South China University of Technology (SCUT), and Mandarin Chinese matrix (CMNmatrix) [29]. More details about the SCUT corpus are provided in the supplemental materials. There are seven, ten, and ten monosyllabic words in the sentence of MSP, MHINT, and SCUT respectively. There are five disyllabic words with a fixed structure "name-verb-number-adjective-object" in CMNmatrix. All corpora were provided by corresponding developers. MHINT materials were recorded by a male speaker. The other corpora were recorded by a female speaker.

During each visit, listeners were tested using four 20-sentence lists respectively from the four corpora in the order of MSP, SCUT, MHINT, and CMNmatrix. For each participant, no list was used more than once. An adaptive procedure with one list generated an SRT50, i.e., the SNR at which the subject has a 50% (by a 1-down-1-up procedure) possibility to recognize a certain percentage of the words in a sentence. The percentage criteria for MSP, SCUT, MHINT, and CMNmatrix were designated as 70%, 75%, 75%, and 50% respectively. The noise was a 20-talker (10 male-10 female) babble noise, which was generated using the same method as described in [22]. The adaptive procedure of CMNmatrix also followed the method in [22], while the others followed the method in [30]. For training purposes, at the end of each trial, the sentence text was provided and the corresponding audio was also replayed to the listener.

*2) Results:* SRT results are shown in Fig. 4. To understand the effect of training and EDR on simulation SRTs in NH listeners, a two-way mixed-design analysis of variance (ANOVA) was administered separately for each corpus, with EDR as the between-subjects factor and the day number as the within-subjects factor. Data normality was confirmed by the Shapiro-Wilk test ($p > 0.05$). Significant main effects of EDR and training were found in all four corpora. There was no significant interaction between EDR and day number except for the CMNMatrix corpus. Detailed statistical results are provided in Table II. The interaction between EDR and day number for the CMNMatrix corpus was mainly caused by the comparable performance with EDR = 150 CL in the first two days. When the first day was excluded, the interaction became insignificant ($F(3, 42) = 0.648$, $p = 0.589$).

Overall, for the simulation results in NH listeners, the wider EDR (150 CL) had a better performance than the narrower EDR (50 CL); both EDRs demonstrated a training effect, indicating a significant improvement as the number of training days increased.

It is of interest to see how many days of training are needed before the NH listeners with vocoder simulations reach a comparable performance to the actual CI users. Independent T-tests were used to examine the significance of the mean SRT difference between each simulated condition and the corresponding CI condition (see Fig. 4 and detailed statistical results in supplementary materials). On the third to fifth day of training, for all four corpora, GET with EDR = 50 CL derived comparable mean SRTs as the CI group (range of mean EDRs:

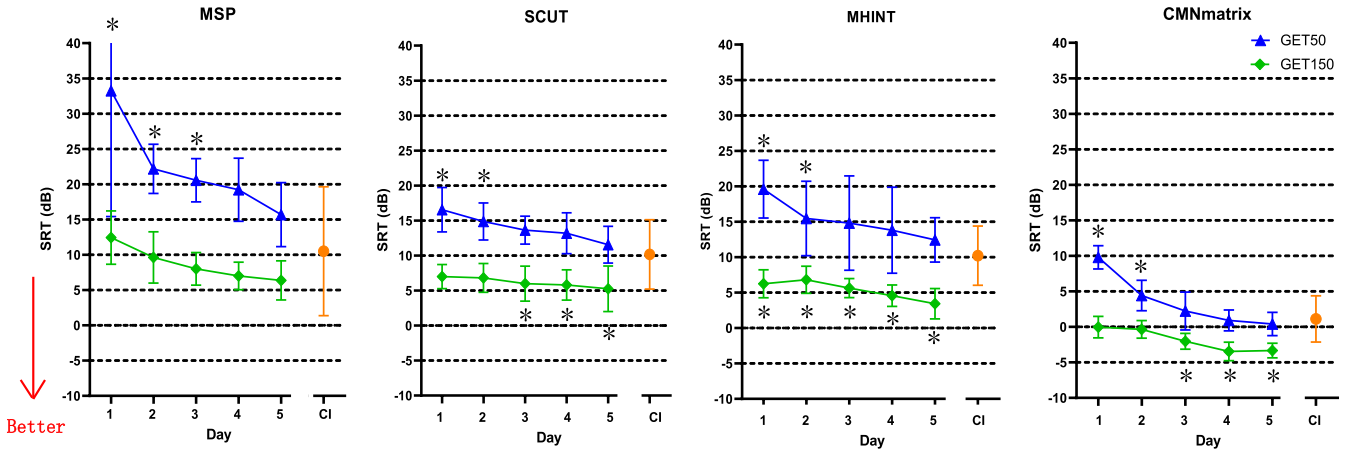


Fig. 4. Experiment 1 results: training effects on speech reception thresholds (SRTs) in noise with GET vocoded speech (with EDR = 50 and 150 CL, denoted by GET50 and GET150 respectively) in NH listeners in comparison with actual CI listeners using four Mandarin sentence corpora (i.e., MSP, SCUT, MHINT, and CMNmatrix). The range of our CI mean EDRs was 30.4-96.1 CL. For the mean SRT with each corpus and each EDR on each day, if it is significantly different ($p < 0.05$) from that of the CI user, a single asterisk is marked.

TABLE II
STATISTICAL RESULTS OF THE GET SIMULATION SRT DATA IN EXPERIMENT 1

| | MSP | | SCUT | | MHINT | | CMNMatrix | |
|---|---|---|---|---|---|---|---|---|
| | F | p | F | p | F | p | F | p |
| EDR | 49.891 | <0.001(*) | 89.717 | <0.001(*) | 45.353 | <0.001(*) | 65.106 | <0.001(*) |
| Day | 9.327 | 0.003(*) | 5.541 | <0.001(*) | 7.062 | <0.001(*) | 95.513 | <0.001(*) |
| EDR × Day | 2.136 | 0.156 | 1.160 | 0.338 | 1.856 | 0.131 | 21.346 | <0.001(*) |

30.4-96.1 CL). The SRT variances were also comparable, particularly for SCUT and MHINT. On the first day, GET with EDR = 150 CL may derive similar mean SRTs as the CI group, but several training days significantly improved performance (i.e., decreased SRTs; outperforming actual CIs). Detailed statistical results are provided in supplementary materials.

Regarding the absolute mean SRTs, the mean SRTs of CI users and the NH group with EDR = 50 on the fifth day were higher than 10 dB for MSP, SCUT, and MHINT, but they were close to zero for CMNMatrix. The reasons include 1) learning effects, i.e., participants had more practice because CMNMatrix was tested after the other three during each visit test; 2) CMNmatrix was tested in a much different psychophysical procedure from the others, i.e., lower percentage criteria for trial correctness and close-set testing rather than open-set.

### C. Experiment 2: Consonant/Vowel Identification and Confusion

The first experiment demonstrated that after about two days of training, the GET vocoder with the same strategy and parameters as actual CI users could derive similar mean SRTs with multiple corpora as experienced CI users. However, this result is still not significant enough to support GET as an optimal CI simulator. Oxenham and colleagues have shown that conventional sine-wave vocoders can also derive SRTs close to CI results if the current spread was manipulated to degrade the sound signal to a proper degree [19], [24]. In order to further investigate the potential benefits of the GET vocoder and its variants, as well as the proposed framework for prosthetics emulation, it is necessary to add these sine-wave vocoders with current spread manipulations as controls.

We know that clear speech signals are highly redundant and robust to many kinds of distortions like clipping, band-limiting, compression, and vocoder processing. However, in noise, the intelligibility may be degraded (i.e., with increased SRTs) by all kinds of distortions. Toward our study aim, two more monosyllabic tests were included, i.e., vowel identification and consonant identification. We assumed that the SRT in noise and the clear phoneme discrimination may rely on different acoustic cues, which may be distorted in different patterns for different vocoders, e.g., the GET vocoder and sine-wave vocoder. The identification scores and the confusion patterns were compared between CI users and five simulation groups of NH listeners. If a vocoder could provide a better simulation in all tests (i.e., SRTs in noise, vowel, and consonant identification), it would be recognized as a better simulator.

*1) Methods:* Nine CI users (see Table I) and 50 NH listeners (college students; age range: 19 to 24; 22 males) were recruited in Experiment 2. The NH listeners were assigned into five groups, each of which included ten people, and were tested using one of the five vocoders. GET50 and GEN50 were implemented according to Sec. III-B with EDR = 50 CL. Sine12, Sine10, and Sine8 were implemented using the conventional 22-channel sine-wave vocoder as described in Sec. III-C respectively using three different degrees of current spread simulation, i.e., 12, 10, and 8 dB/octave.

SRTs in noise were measured in all six groups (one CI + five NH) with MSP, SCUT, and MHINT corpora using the

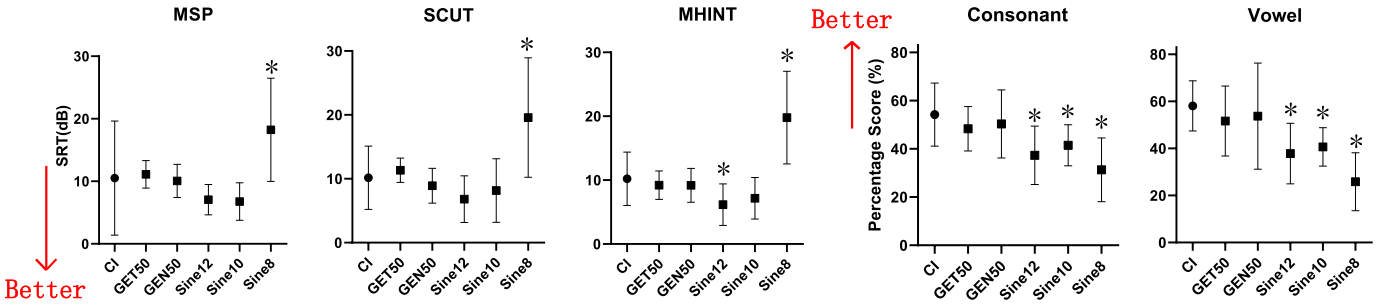


Fig. 5. Experiment 2 results: Speech reception threshold (SRT) in noise and consonant and vowel identification with actual CIs and five vocoders (GET50: Gaussian-Enveloped Tones with EDR = 50 CL; GEN50: Gaussian-Enveloped Noise with EDR = 50 CL; SineX: conventional sine-wave vocoder with current spread degree of X = 12, 10, and 8 dB/octave). For the mean results for each simulated condition, if it is significantly different ($p < 0.05$) from that of the CI user, a single asterisk is marked.

TABLE III
PAIRWISE STATISTICAL RESULTS OF THE SIX ALGORITHM DATA IN EXPERIMENT 2

| | MSP | | SCUT | | MHINT | | Consonant | | Vowel | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean Diff. (dB) | P | Mean Diff. (dB) | P | Mean Diff. (dB) | P | Mean Diff. (%) | P | Mean Diff. (%) | P |
| CI vs. GET50 | -0.6 | 0.813 | -1.2 | 0.639 | 1.0 | 0.616 | 5.9 | 0.304 | 6.5 | 0.348 |
| CI vs. GEN50 | 0.5 | 0.858 | 1.3 | 0.601 | 1.0 | 0.599 | 3.9 | 0.480 | 4.4 | 0.515 |
| CI vs. Sine12 | 3.5 | 0.174 | 3.4 | 0.168 | 4.1 | 0.040* | 16.9 | 0.003** | 20.31 | 0.004** |
| CI vs. Sine10 | 3.7 | 0.150 | 2.0 | 0.422 | 3.1 | 0.129 | 12.8 | 0.028* | 17.5 | 0.013* |
| CI vs. Sine8 | -7.7 | 0.003** | -9.4 | <0.001*** | -9.6 | <0.0001**** | 22.9 | 0.0001*** | 32.3 | <0.0001**** |
| GET50 vs. GEN50 | 1.1 | 0.673 | 2.4 | 0.317 | 0 | 0.991 | -2.0 | 0.724 | -2.1 | 0.753 |
| GET50 vs. Sine12 | 4.1 | 0.110 | 4.5 | 0.065 | 3.1 | 0.117 | 11.1 | 0.050 | 13.9 | 0.042* |
| GET50 vs. Sine10 | 4.4 | 0.096 | 3.2 | 0.205 | 2.1 | 0.304 | 6.9 | 0.228 | 11.0 | 0.112 |
| GET50 vs. Sine8 | -7.1 | 0.006** | -8.3 | 0.001** | -10.6 | <0.0001**** | 17.1 | 0.003** | 25.8 | 0.0003*** |
| GEN50 vs. Sine12 | 3.0 | 0.223 | 2.1 | 0.374 | 3.1 | 0.111 | 13.0 | 0.019* | 16.0 | 0.017* |
| GEN50 vs. Sine10 | 3.3 | 0.193 | 0.7 | 0.762 | 2.0 | 0.297 | 8.9 | 0.114 | 13.1 | 0.054 |
| GEN50 vs. Sine8 | -8.2 | 0.002** | -10.7 | <0.0001**** | -10.6 | <0.0001**** | 19.0 | <0.001*** | 27.9 | <0.0001**** |
| Sine12 vs. Sine10 | 0.294 | 0.907 | -1.4 | 0.573 | -1.0 | 0.601 | -4.2 | 0.454 | -2.8 | 0.672 |
| Sine12 vs. Sine8 | -11.2 | <0.0001**** | -12.8 | <0.0001**** | -13.6 | <0.0001**** | 6.0 | 0.268 | 12.0 | 0.070 |
| Sine10 vs. Sine8 | -11.5 | <0.0001**** | -11.4 | <0.0001**** | -12.6 | <0.0001**** | 10.2 | 0.071 | 14.8 | 0.031* |

same procedure as Experiment 1 in Sec. IV-B. Consonant and vowel tests were measured using a close-set alternative forced choice manner. There were 11 final consonants embedded in a /Ca/ syllable in Tone 1, which refers to the first of the four tones used in the Mandarin language [31]. There were 20 initial vowels embedded in a /dV/ syllable in Tone 1 [32], [33]. We used recorded speech from three female and three male speakers, resulting in 66 (= 11 × 6) and 120 (= 20 × 6) tokens in total for consonants and vowels respectively. For the CI group, SRT results taken from Exp. 1 were used to compare with the GEN and GET processed speech in Exp. 2.

The first experiment told us that NH subjects need at least two days to train their brains to become familiar with the vocoded speech. In this experiment, NH participants did the test remotely on three consecutive days. CI participants were tested in the laboratory only once. Remote tests were carried out through a customized website that could play audio stored on a server and also collect the subjects' feedback. On each day, the task order is from SRTs (from MSP through SCUT to MHINT) through the consonant to vowel test. In SRT tasks, answer correctness was provided in the same way as in Experiment 1. In consonant and vowel tasks, no feedback about the answer's correctness was provided.

*2) Results:* After an initial examination of the data collected in Exp. 1 and 2, we noticed that the remotely collected data from Exp. 2 had sometimes odd results and exhibited larger variances than the data from Exp. 1 collected in the lab. Two of the 50 participants, i.e., S14 (for GET50) and S54 (for Sine10), were excluded from the analysis of the results. S14 was excluded due to missing data, which could have affected the validity of the results. S54 was excluded because the subject consistently selected a fixed answer in the vowel and consonant tasks (see the supplementary materials), indicating that the subject may not have fully understood the tasks or may not have been fully engaged in the study, making their data unreliable for analysis. This exclusion was done to ensure the integrity and accuracy of the results. For each of the five tests with each participant, the better result (i.e., lower SRT or higher percentage scores) between the second and third days was preserved for further analysis.

The mean results are illustrated in Fig. 5. Consistent with the first experiment, all actual and simulated CI SRTs were higher than theoretically much lower SRTs with non-vocoded original signals in NH listeners (not tested in the experiment). The mean scores for both consonants and vowels were below 60%, putting them far below the typical performance of NH individuals when listening to non-vocoded speech.

The results of one-way ANOVA revealed significant differences among the outcomes produced by the six algorithms. Post-hoc pairwise comparison with no correction showed that in all of the five tasks, mean results with GET50 and GEN50 showed no significant difference from those of

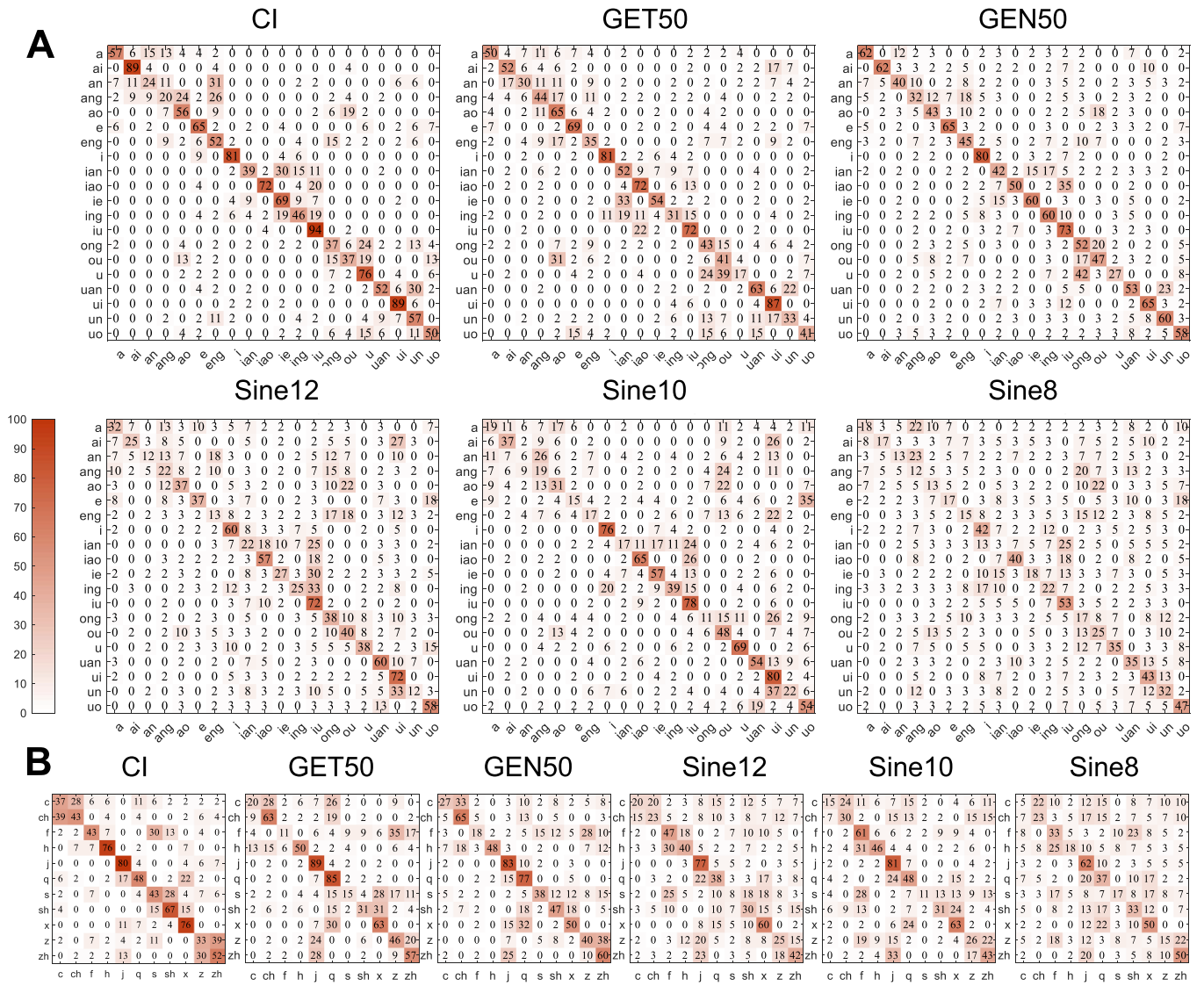


Fig. 6. Average phoneme confusion matrices for each of the six groups of listeners, i.e., CI, GET50 (Gaussian-Enveloped Tones with EDR = 50 CL), GET50 (Gaussian-Enveloped Noise with EDR = 50), SineX (conventional sine-wave vocoder with current spread degree of X = 12, 10, and 8 dB/octave). A: Final vowel; B: Initial consonant. Unit: %.

the actual CI group. Sine8 derived the poorest performance among the six groups. The mean results from Sine12 and Sine10 showed no significant difference.[4] The fundamental cause remained elusive. They outperformed the CI group in the SRT tasks (by more than 2.0 dB; statistically significant for Sine12 with MHINT, but not for the others) but worse than CI in monosyllabic phoneme recognition tasks (by more than 12% significantly). This indicates that conventional sine-wave vocoders degrade speech information in a different manner from ACE and GET/GEN. Detailed descriptive statistical analysis results are in Table III.

To further observe the detailed confusion patterns, the mean vowel and consonant confusion matrices are shown in Fig. 6. The distances between the five simulated matrices (GET50, GEN50, Sine12, Sine10, and Sine8) and the CI matrices were respectively 6.6, 5.7, 7.9, 7.5, and 9.6 for the vowel results and 11.2, 9.1, 9.5, 9.6, 10.9 for the consonant results. The distance was calculated by the root-mean-square (RMS) of the differences between corresponding elements in the matrices. The GEN-processing resulted in the least RMS difference and the closest in similarity to actual CI processing. The results suggest that GEN50 is the optimal simulator of the tested ones. Based on the above distance results, GET50 performed better (i.e., lower distance) compared to the SineX vocoders in vowel recognition but worse (i.e., higher distance) in consonant recognition.

## V. Discussion

### A. CI Simulator

In this work, to experimentally validate the GET/GEN vocoders and the proposed prosthesis simulation principle, two

[4]The results suggest a nonlinear association between the current spread degree and the tested performance, particularly for the SRT tests (refer to Fig. 6)

experiments were carried out. The most widely used CI signal processing strategy, ACE, and corresponding CI users were included. After about two days of training, NH participants listening to the sounds processed by the GET or GEN vocoders showed consistent mean results in all tests as CI users. However, the conventional sine-wave vocoders with 12 and 10 dB/oct current spread simulation outperformed the CI group in SRTs, but did worse than the CI group in consonant and vowel recognition. This interaction indicates that conventional sine-wave vocoders do not transmit speech information in the same way as CIs. For example, the dynamic range was not compressed in sine-wave vocoders. The wide dynamic range is critical to speech-in-noise intelligibility [22] and would partly offset the degradation introduced by the current spread. Identification of clear phonemes may be more sensitive to current spread or not substantially supported by the wide dynamic range. Regarding the consonant and vowel confusion patterns, the newly proposed GET variant, i.e., the GEN vocoder, also performed better. The timbre of the "tone" carriers might highly influence consonant discrimination with GET50, which was compensated by the noise carrier in GEN50.

This study is limited in several ways. The sample size of the study is small ($N \leq 10$). Experiment 1 revealed that after three to five days of half-hour training sessions per day, NH listeners were able to comprehend the GET/GEN-processed speech at a level comparable to that of CI listeners. While performance appeared to level off by the fifth day, it cannot be ruled out that further training might lead to lower speech recognition thresholds (SRTs). Further research is needed to simulate the rehabilitation process for CI users following activation, including their performance at 1, 3, and 12 months after activation. Additionally, the current spread [34] or frequency-place shift [20] were not quantified or manipulated in GET/GEN vocoders. The GET/GEN vocoders are highly effective in simulating the group performance of those CI users. In future studies, more individual parameters will be considered to determine their impact on predicting individual performance.

Instead of human listeners, computation algorithms have also been suggested to be used for CI performance prediction. Brochier and colleagues combined biophysical models of the electric field in the cochlea, neural model of signal processing in the auditory nerve, and automatic speech recognition to predict the perception and misperception of phonemes and SRT with CIs [35]. Both the method and our GET/GEN vocoders have a similar philosophy in terms of utilizing as much information from the CI processing to encode sounds. Both human simulation and machine simulation have their own advantages. Human simulation could assume that the cognitive and language abilities in the brain are equivalent at least for some post-lingual CI users. The time course of learning and rehabilitation could be simulated in human subjects. Human simulation could also be used as a tool not only for CI performance prediction but for demonstration and education. Machine simulation is much less time-consuming, a benefit to the industry. In this work, our focus is on human simulation.

### B. Neuroprothesis Simulators

The lessons from CIs to other neuroprostheses have been discussed in previous literature [7], [9], [36]. These limited works were mainly comparing bionic hearing and vision. Among them, the most systematic comparison with balanced detailed contents on both sides was provided in [9] published in 2007. This status might be due to the limited overlaps between researchers, knowledge, and tools of these two fields. In recent years, many simulation works added new features like gaze contingency [37], temporal features [38], infrared image [39], end-to-end optimization [40], semantic face image translation [41], and electrode-retina distance [6] to supplement the previously un-simulated features or to add new features to current prosthetic vision techniques. This kind of work also happened in other prosthetic fields, e.g., prosthetic arm and hand [42]. In these experiments, there were usually only normal subjects without real implanted ones. This is mainly limited by the small population of implanted patients.

Previously, we know most CI simulators can predict actual performance trends with various simulation parameters. For example, in [43] the effects of channel number (from 1 to 9) on sentence recognition in quite was examined. However, the absolute quantitative scores with CIs could not be easily predicted by these simulators [8], [9]. Some studies have introduced certain degrees of current spread to derive similar intelligibility in noise as CI user [19]. However, in this work, Experiment 2 showed that this method cannot derive satisfactory simulation results simultaneously in SRT and phoneme recognition tasks. We argue that only simulating key features (e.g., the temporal envelope) is not enough. The current work verified that faithfully replicating the signal processing details of the real implant could derive comparable perceptual results and patterns in multiple psychophysical tasks.

## VI. Conclusion

The GEN vocoder, a variant of our recently proposed GET vocoder, after enough training, could be used to simulate the mostly widely used ACE strategy. The perceptual patterns in SRTs with multiple corpora and vowel/consonant phoneme recognition were quantitatively consistent between the simulated and actual CI users. The successful simulation proved that faithfully replicating the signal processing (or degradation) details of the real prosthesis is a necessary step in the simulations of prostheses. This is a lesson from this audition study and may be helpful for other sensory neuroprostheses, in considering their great similarities in electrode-to-brain interface and brain sensation and cognition.

## Acknowledgment

The authors would like to thank all volunteers who have participated in their experiments. They thank XU Rui and CAO Teng for their assistance in the data collection in Experiment one. They thank Li Xu for providing the speech materials for vowel and consonant recognition tests. Language polishing assisted by OpenAI's ChatGPT in several parts of the article during revision.

## REFERENCES


[1] F.-G. Zeng, "Challenges in improving cochlear implant performance and accessibility," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1662–1664, Aug. 2017.

[2] F.-G. Zeng, "Celebrating the one millionth cochlear implant," *JASA Exp. Lett.*, vol. 2, no. 7, Jul. 2022, Art. no. 077201.

[3] J. Lewis, D. Brown, W. Cranton, and R. Mason, "Simulating visual impairments using the unreal engine 3 game engine," in *Proc. IEEE 1st Int. Conf. Serious Games Appl. Health (SeGAH)*, Nov. 2011, pp. 1–8.

[4] Z. Tu, N. Ma, and J. Barker, "Optimising hearing aid fittings for speech in noise with a differentiable hearing loss model," in *Proc. Interspeech*, Jun. 2021, pp. 691–695.

[5] T. Baer and B. C. J. Moore, "Effects of spectral smearing on the intelligibility of sentences in noise," *J. Acoust. Soc. Amer.*, vol. 94, no. 3, pp. 1229–1241, Sep. 1993.

[6] D. Avraham and Y. Yitzhaky, "Simulating the perceptual effects of electrode–retina distance in prosthetic vision," *J. Neural Eng.*, vol. 19, no. 3, 2022, Art. no. 035001.

[7] H. T. Nguyen et al., "Thalamic visual prosthesis," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 8, pp. 1573–1580, Aug. 2016.

[8] F. Kong, Y. Mo, H. Zhou, Q. Meng, and N. Zheng, "Channel-vocoder-centric modelling of cochlear implants: Strengths and limitations," in *Proc. 9th Conf. Sound Music Technol.*, X. Shao, K. Qian, X. Wang, and K. Zhang, Eds. Singapore: Springer, 2023, pp. 137–149.

[9] L. E. Hallum, G. Dagnelie, G. J. Suaning, and N. H. Lovell, "Simulating auditory and visual sensorineural prostheses: A comparative review," *J. Neural Eng.*, vol. 4, no. 1, pp. S58–S71, Mar. 2007.

[10] 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, no. 6332, pp. 236–238, Jul. 1991.

[11] B. S. Wilson, "Getting a decent (but sparse) signal to the brain for users of cochlear implants," *Hearing Res.*, vol. 322, pp. 24–38, Apr. 2015.

[12] H. Zhou, A. Kan, G. Yu, Z. Guo, N. Zheng, and Q. Meng, "Pitch perception with the temporal limits encoder for cochlear implants," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2528–2539, 2022.

[13] 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. Roy. Soc. B, Biol. Sci.*, vol. 278, no. 1711, pp. 1595–1600, May 2011.

[14] P. J. Blamey, R. C. Dowell, Y. C. Tong, A. M. Brown, S. M. Luscombe, and G. M. Clark, "Speech processing studies using an acoustic model of a multiple-channel cochlear implant," *J. Acoust. Soc. Amer.*, vol. 76, no. 1, pp. 104–110, Jul. 1984.

[15] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, Oct. 1995.

[16] 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," *J. Acoust. Soc. Amer.*, vol. 102, no. 4, pp. 2403–2411, Oct. 1997.

[17] Q.-J. Fu, R. V. Shannon, and X. Wang, "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," *J. Acoust. Soc. Amer.*, vol. 104, no. 6, pp. 3586–3596, Dec. 1998.

[18] Q. Meng et al., "Time-compression thresholds for Mandarin sentences in normal-hearing and cochlear implant listeners," *Hearing Res.*, vol. 374, pp. 58–68, Mar. 2019.

[19] A. J. Oxenham and H. A. Kreft, "Speech perception in tones and noise via cochlear implants reveals influence of spectral resolution on temporal processing," *Trend. Hear.*, vol. 18, Sep. 2014, Art. no. 2331216514553783.

[20] N. Zhou, L. Xu, and C.-Y. Lee, "The effects of frequency-place shift on consonant confusion in cochlear implant simulations," *J. Acoust. Soc. Amer.*, vol. 128, no. 1, pp. 401–409, Jul. 2010.

[21] Q. Meng, H. Zhou, T. Lu, and F.-G. Zeng, "Pulsatile Gaussian-enveloped tones (GET) for cochlear-implant simulation," *Appl. Acoust.*, vol. 208, Jun. 2023, Art. no. 109386.

[22] Y. Mo et al., "Effects of number of maxima and electrical dynamic range on speech-in-noise perception with an 'n-of-m' cochlear-implant strategy," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104169.

[23] J. N. Saba, H. Ali, and J. H. L. Hansen, "Formant priority channel selection for an 'n-of-m' sound processing strategy for cochlear implants," *J. Acoust. Soc. Amer.*, vol. 144, no. 6, pp. 3371–3380, Dec. 2018.

[24] E. R. O'Neill, M. N. Parke, H. A. Kreft, and A. J. Oxenham, "Role of semantic context and talker variability in speech perception of cochlear-implant users and normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 149, no. 2, pp. 1224–1239, Feb. 2021.

[25] H. Zhou, N. Wang, N. Zheng, G. Yu, and Q. Meng, "A new approach for noise suppression in cochlear implants: A single-channel noise reduction algorithm," *Frontiers Neurosci.*, vol. 14, p. 301, Apr. 2020.

[26] C. Cusumano, D. R. Friedmann, Y. Fang, B. Wang, J. T. Roland, and S. B. Waltzman, "Performance Plateau in prelingually and postlingually deafened adult cochlear implant recipients," *Otol. Neurotol.*, vol. 38, no. 3, pp. 334–338, 2017.

[27] Q.-J. Fu, M. Zhu, and X. Wang, "Development and validation of the Mandarin speech perception test," *J. Acoust. Soc. Amer.*, vol. 129, no. 6, pp. EL267–EL273, Jun. 2011.

[28] L. L. N. Wong, S. D. Soli, S. Liu, N. Han, and M.-W. Huang, "Development of the Mandarin hearing in noise test (MHINT)," *Ear Hearing*, vol. 28, no. 2, pp. 70S–74S, 2007.

[29] H. Hu, X. Xi, L. L. N. Wong, S. Hochmuth, A. Warzybok, and B. Kollmeier, "Construction and evaluation of the Mandarin Chinese matrix (CMNmatrix) sentence test for the assessment of speech recognition in noise," *Int. J. Audiol.*, vol. 57, no. 11, pp. 838–850, Nov. 2018.

[30] Q. Meng, N. Zheng, and X. Li, "Mandarin speech-in-noise and tone recognition using vocoder simulations of the temporal limits encoder for cochlear implants," *J. Acoust. Soc. Amer.*, vol. 139, no. 1, pp. 301–310, Jan. 2016.

[31] S. Qi et al., "Effects of adaptive non-linear frequency compression in hearing aids on Mandarin speech and sound-quality perception," *Frontiers Neurosci.*, vol. 15, p. 1038, Aug. 2021.

[32] X. Chen et al., "Effects of nonlinear frequency compression on Mandarin speech and sound-quality perception in hearing-aid users," *Int. J. Audiol.*, vol. 59, no. 7, pp. 524–533, Jul. 2020.

[33] J. Yang et al., "Effects of nonlinear frequency compression on the acoustic properties and recognition of speech sounds in Mandarin Chinese," *J. Acoust. Soc. Amer.*, vol. 143, no. 3, pp. 1578–1590, Mar. 2018.

[34] A. J. Oxenham, "How we hear: The perception and neural coding of sound," *Annu. Rev. Psychol.*, vol. 69, no. 1, pp. 27–50, Jan. 2018.

[35] T. Brochier et al., "From microphone to phoneme: An end-to-end computational neural model for predicting speech perception with cochlear implants," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 11, pp. 3300–3312, Nov. 2022.

[36] B. S. Wilson and M. F. Dorman, "Interfacing sensors with the nervous system: Lessons from the development and success of the cochlear implant," *IEEE Sensors J.*, vol. 8, no. 1, pp. 131–147, Jan. 2008.

[37] N. Paraskevoudi and J. S. Pezaris, "Full gaze contingency provides better reading performance than head steering alone in a simulation of prosthetic vision," *Sci. Rep.*, vol. 11, no. 1, pp. 1–17, May 2021.

[38] D. Avraham, J.-H. Jung, Y. Yitzhaky, and E. Peli, "Retinal prosthetic vision simulation: Temporal aspects," *J. Neural Eng.*, vol. 18, no. 4, Aug. 2021, Art. no. 0460d9.

[39] J. Liang et al., "An infrared image-enhancement algorithm in simulated prosthetic vision: Enlarging working environment of future retinal prostheses," *Artif. Organs*, vol. 46, no. 11, pp. 2147–2158, Nov. 2022.

[40] J. de Ruyter van Steveninck, U. Güçlü, R. van Wezel, and M. van Gerven, "End-to-end optimization of prosthetic vision," *J. Vis.*, vol. 22, no. 2, p. 20, Feb. 2022.

[41] X. Xia et al., "Semantic translation of face image with limited pixels for simulated prosthetic vision," *Inf. Sci.*, vol. 609, pp. 507–532, Sep. 2022.

[42] S. Mick et al., "Shoulder kinematics plus contextual target information enable control of multiple distal joints of a simulated prosthetic arm and hand," *J. NeuroEng. Rehabil.*, vol. 18, no. 1, pp. 1–17, Dec. 2021.

[43] L. Xu, X. Xi, A. Patton, X. Wang, B. Qi, and L. Johnson, "A cross-language comparison of sentence recognition using American English and Mandarin Chinese HINT and AzBio sentences," *Ear Hearing*, vol. 42, no. 2, pp. 405–413, 2021.