ElectrodeNet – A Deep Learning Based Sound Coding Strategy for Cochlear Implants

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

18-02-2022 / 23-02-2022

CITATION

Huang, Enoch Hsin-Ho; Chao, Rong; Tsao, Yu; Wu, Chao-Min (2022): ElectrodeNet – A Deep Learning Based Sound Coding Strategy for Cochlear Implants. TechRxiv. Preprint.
<https://doi.org/10.36227/techrxiv.19195013.v1>

DOI

[10.36227/techrxiv.19195013.v1](https://doi.org/10.36227/techrxiv.19195013.v1)

ElectrodeNet – A Deep Learning Based Sound Coding Strategy for Cochlear Implants

Enoch Hsin-Ho Huang, Rong Chao, Yu Tsao, and Chao-Min Wu

Abstract— Objective: ElectrodeNet, a deep-learning based sound coding strategy for the cochlear implant (CI), is proposed in this study. The performance between ElectrodeNet and the advanced combination encoder (ACE) coding strategy in speech intelligibility is compared. **Methods:** ElectrodeNet emulates the ACE strategy and replaces the conventional envelope detection using various forms of artificial neural networks. Network models of deep neural network (DNN), convolutional neural network (CNN), and long short-term memory (LSTM) were trained using the fast Fourier transformed clean speech and the corresponding electrode stimulation patterns. Objective speech intelligibility was estimated for ElectrodeNets for the factors of loss function, network architecture, language, and noise type. Subjective listening tests for vocoded Mandarin speech were conducted with 40 normal-hearing listeners. **Results:** DNN, CNN, and LSTM based ElectrodeNets exhibited strong correlations with the ACE strategy in short-time objective intelligibility (STOI) and normalized covariance metric (NCM) scores. For objective evaluations, small mean squared error (MSE) scores between ACE and ElectrodeNets were less than 0.01 under all experimental conditions, whereas linear correlation coefficient (LCC) and Spearman's rank correlation coefficient (SRCC) were obtained in large values greater than 0.97 and 0.96, respectively. According to the listening test results, substantial positive relationships were also observed between ACE and both DNN and CNN based ElectrodeNets with MSEs smaller than 0.02, and LCCs and SRCCs greater than 0.9. **Significance:** This study demonstrates the feasibility of using deep learning to encode sound into meaningful patterns for CI listening.

Index Terms— cochlear implant, deep learning, sound coding strategy, vocoder simulation

I. INTRODUCTION

In biomedical engineering, the cochlear implant (CI) is a revolutionary device that enables people with severe-to-

profound hearing loss to restore partial sense of hearing [1]–[3]. With the design of an external CI sound processor and an internal device to stimulate the auditory nerves in meaningful pulse patterns, the lives of hundreds of thousands of CI users have been changed. With contributions from multidisciplinary researchers in electrical engineering, biomaterials, otolaryngology, audiology and so on, the CI system today helps a recipient not only improve lip-reading ability but also recognizes speech in quiet conditions and distinguishes different environmental sounds [1]–[5]. Apart from the present achievements, challenges of electric hearing are present in recognizing speech under interferences, and clean sounds such as pitch, tonal languages, and music [1]–[4]. Furthermore, the difficulties for CI and hearing aid users in listening to speech through face masks and to conversations in distance learning and video conferences are getting noticed during the recent COVID-19 pandemic [6]. Therefore, innovations in sound processing technologies and algorithms are in demand to improve the CI user experiences.

Deep learning [7], [8], a subset of artificial intelligence (AI) and machine learning, has an increasing influence and potential on the innovation of CI. In the recent decade, AI based approaches have demonstrated their supreme capability over traditional signal processing algorithms in computer vision, natural language processing, gaming, and biomedical engineering, including hearing healthcare [9]. Machine learning and deep learning have changed various aspects of the 'AI + CI' research [10], including prognosis estimation [11], electrode placement [12], robotic surgery [13], mapping [14], [15], and sound signal processing [16], [17]. In addition, the successful stories of deep learning in general audio processing [18], [19] and hearing aids [20], [21] may also significantly change the related field of CI sound processing.

Deep learning based on various types of neural networks has been used in speech and music processing for the CI. The network models used for CI applications include the feed-forward neural network [22], [23], deep neural network (DNN) [24], [25], convolutional neural network (CNN) [26], fully convolutional network (FCN) [27], [28], recurrent neural network (RNN) based on the long short-term memory (LSTM) architecture [29], [30], and deep recurrent neural network (DRNN) [31]. Among the aforementioned studies, deep learning based speech enhancement approaches outperform traditional approaches in objective and subjective CI simulation as well as in clinical tests with CI recipients [22]–[26]. Furthermore, the multimodal approaches can also

This work was supported by the Ministry of Science and Technology of Taiwan under Grant MOST 108-2221-E-008-066. (Corresponding author: Chao-Min Wu.)

E. H.-H. Huang is with the Department of Electrical Engineering, National Central University, Taoyuan 320317, Taiwan. He is also with the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taipei 115201, Taiwan.

R. Chao is with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701401, Taiwan. He is also with the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taipei 115201, Taiwan.

Y. Tsao is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115201, Taiwan.

C.-M. Wu is with the Department of Electrical Engineering, National Central University, Taoyuan 320317, Taiwan. (e-mail: wucm@ee.ncu.edu.tw).

improve the speech intelligibility in forms of electric and acoustic stimulation (EAS) [27] and audiovisual integration [28]. In music processing, source separation and remixing techniques adjust the balance between vocal and instrumental sound sources of songs for improving CI recipients' enjoyment with music [32]–[34]. These speech and music processing approaches can improve the CI perception and overcome limitations existing in traditional signal processing techniques. However, these approaches typically modify the front-end processing only. For the remaining signal processing path, particularly for the core sound coding strategy, the efficacy of utilizing deep learning approaches have to be thoroughly studied.

The CI sound coding strategy plays a key role in representing acoustic signals as recognizable stimulation patterns using traditional signal processing techniques [35], which may possibly be replaced using deep learning approaches. The two prominent sound coding strategies, continuous interleaved sampling (CIS) [36] and advanced combination encoder (ACE) [37], exhibit a reasonable good performance for speech understanding in quiet, but they have already existed for 20–30 years. More recent commercial sound coding strategies, including MP3000 [38], [39], fine structure processing (FSP) [40], [41], HiRes 120, HiRes Optima [42], [43], and CRYSTALIS [44], have been proposed with some improvements. However, further innovated coding strategies are still needed [35], [45]. In the era of AI, it is important to investigate the feasibility of applying deep learning in replacing traditional approaches for sound coding strategies [46]. Furthermore, the effectiveness of using different network architectures to replace a portion of the linear processing in the ACE strategy can be studied in preparation of more powerful nonlinear computations in the future. Consequently, ElectroNet, a deep learning based sound coding strategy, was proposed to imitate the conventional ACE strategy [46]. A comprehensive objective evaluation using CI simulations was conducted using loss function, neural network architectures, languages of speech materials, and noise type as factors. Correlation between the ACE and ElectroNets have been measured for the objective speech intelligibility. Sentence recognition tests for vocoded Mandarin speech were conducted with normal-hearing listeners. The approaches and outcomes of this study illuminates several aspects of transforming conventional sound coding strategies into deep learning based strategies.

The rest of this article is organized as follows. Section II describes the proposed ElectroNet and experimental methods. Section III presents the objective and subjective evaluation results. Section IV provides the discussions. Section V draws the conclusions.

II. METHODS

This section describes the ElectroNet, which models the ACE strategy, and the experimental methods for objective and subjective evaluations.

A. The ACE Strategy

ACE, the most commonly used sound coding strategy for Nucleus CI systems, is illustrated in Fig. 1 (a). Sound signals

are collected by microphones and enhanced in the preprocessing stage. K samples are Hann windowed and converted into L bins using Fast Fourier Transform (FFT), where generally $K = 128$ and $L = (K/2 + 1) = 65$. Magnitudes of the complex FFT outputs represent the envelopes in L channels with a typical bandwidth of 125 Hz linearly assigned between 0 and 8 kHz. The envelope detection stage reduces the number of frequency bands from L bins into M channels. Each bin at approximately 1 kHz and below is directly output to a channel, and two to eight higher frequency bins are combined together to form a single channel. The L -to- M envelope detection uses the power-sum combination method to derive gains for each channel [47]. Each channel corresponds to one physical electrode for the Nucleus-24 implant device, which includes $M = 22$ intracochlear electrodes and two extra-cochlear ground electrodes. An N -of- M selection stage is applied to M channels by enabling typically $N = 8$ channels with the maximum amplitudes. A mapping stage adjusts stimulation levels with respect to the loudness growth function (LGF) and personalized electric current levels. Pulse patterns are transmitted from a coil as radio frequencies (RFs) to an implant device. In addition, the electrode stimulation patterns can also be synthesized by a tone vocoder [48] or noise vocoder [49] as audible sound for normal-hearing listeners. The ACE strategy defined in Cochlear's Nucleus MATLAB Toolbox (NMT) v4.31 [50] was used in this study.

B. Artificial Neural Networks

In deep learning, the artificial neural network simulates the human neural system capable of learning. An artificial neural network consists of a number of neuron layers, and each artificial neuron is a perceptron, which combines multiple inputs as a nonlinear output through an activation function. Under supervised learning, a neural network learns the interconnection weights from a large amount of labelled training data using the back-propagation algorithm to efficiently compute the derivatives or gradients. By measuring the loss function or objective function during the training process, the network is shaped towards an optimized model. Apart from the number of neurons and layers, a neural network is also defined using hyperparameters, such as the number of neurons in each layer, learning rate, and number of epochs. The trained network model is evaluated for its performance using test data.

DNN, CNN, and LSTM are three networks with representative architectures in deep learning. A typical DNN [51] or the multilayer perceptron (MLP) [52], is a feedforward network containing multiple hidden layers, where neurons between adjacent layers are fully connected to each other. A neural network layer with fully connected neurons is also referred to as a dense layer. For a CNN network, it typically consists of the convolutional layers, pooling layers, and dense layers [53]. The convolution process extracts features using the definitions of hyperparameters such as kernel size, padding, and stride. The pooling operation, which reduces the sizes of feature maps, is absent in some audio processing models and the FCN networks to avoid its downsampling functionality. As for the recurrent type of networks, LSTM [54] is a refined

Fig. 1. Signal processing for (a) the ACE strategy, and (b) the ElectrodeNet strategy. The DNN, CNN, or LSTM model within ElectrodeNet is trained using the corresponding L-bin input and M-channel output data at the envelope detection stage in the ACE coding strategy.

type of RNN [55]. The LSTM network is designed to process sequences using cells and gates to operate as memory with different lengths over time.

C. ElectrodeNet

ElectrodeNet utilizes an artificial neural network to imitate the CI electrode output behaviors of the ACE strategy. The signal processing path for ElectrodeNet is illustrated in Fig. 1 (b), which includes processing stages comparable to the ACE strategy in Fig. 1 (a). The preprocessing stage was bypassed and the FFT stage of the ACE strategy was adopted. Using supervised learning, an artificial neural network learned to mimic the conventional envelope detection stage of ACE. The input layer of the neural network was set to $L = 65$ neurons, which corresponded to the FFT magnitudes. The neural network was implemented in different architectures such as DNN, CNN, and LSTM. Pairs of spectral-temporal matrices in L and M spectral bins for clean speech generated by ACE were used to train the neural networks within ElectrodeNet. The N -of- M channel selection stage was preserved to process the neural network output of $M = 22$ neurons. Electrode stimulation patterns were synthesized as audible sounds using the tone vocoder of NMT and normalized in sound levels. Training and testing for deep learning were implemented using PyTorch 1.1.0, and the other processing was performed using MATLAB R2020a.

Fig. 1 (b) illustrates the ElectrodeNet coding strategy using a DNN, CNN, or LSTM model. The DNN model investigated for ElectrodeNet had four dense layers containing 1024, 512, 256, and 22 neurons, respectively. The CNN model had two layers of 1-D convolution, including 1024 channels and 512 channels with a kernel size of 3 and padding of 2. The convolutional layers were followed by two dense layers, each consisting of 256 neurons and 22 neurons, and no pooling layers. The LSTM network used one 1-D convolution with 1024 channels, one directional LSTM layer with 512 nodes, and two fully connected layers with 256 and 22 neurons, respectively.

Common settings were applied to the DNN, CNN, and LSTM based ElectrodeNets. The learning rate was 0.0001,

TABLE I

THE TMHINT SPEECH MATERIALS USED FOR TRAINING, OBJECTIVE EVALUATION, AND SUBJECTIVE LISTENING TEST.

Speech Materials	TMHINT-AS	TMHINT-NCU		
Recorded By	Academia Sinica	National Central University		
Usage	ElectrodeNet Training	Objective Evaluation	Subjective Listening Test	
List No.	9 - 16	1-8	1-8	16
Speaker	4 male 4 female	1 male 1 female	1 female	1 female
Noise Type	-	SSN white noise street noise cafeteria	SSN	SSN
SNR Levels	Quiet	-10, -5, 0, 5, 10, 15 dB Quiet	-5, 0, 5 dB Quiet	-5, 0, 5 dB Quiet

and the number of epochs was 100. The rectified linear unit (ReLU) was used as the common activation function and causal implementations were used to fit the scenarios of real-time signal processing.

D. Datasets

Speech materials of a tonal language, Mandarin Chinese, and a non-tonal language, English, were investigated in this study. Distinct sentences spoken by different people were used in training and testing the neural networks for ElectrodeNet. The sampling frequency for all datasets was 16 kHz.

1) *Mandarin dataset*: The Taiwan Mandarin Hearing in Noise Test (TMHINT) sentences [56] was used for both objective and subjective evaluations of CI speech intelligibility. The TMHINT speech material contains 16 lists and each list has 20 sentences. All sentences have an identical length of 10 syllables. Each syllable corresponds to a Chinese character and has one of the four primary lexical tones, tone 1 (high and flat), tone 2 (rising), tone 3 (falling then rising), and tone 4 (falling), or an additional neutral tone. TMHINT is designed to have phonemically balanced sentence lists [56]. To validate the efficacy of ElectrodeNet, sound files of different

speakers recorded by two institutes, Academia Sinica (AS) and the National Central University (NCU), were separately used for training and testing. The sound levels were normalized between the TMHINT(AS) and TMHINT(NCU) sentences. The TMHINT(AS) sentences were spoken by four male and four female speakers, whereas the TMHINT(NCU) sentences were spoken by one male and one female.

Table I shows the arrangement of TMHINT sentences for training and testing in this study. In training ElectroNet, 320 clean sentences from Lists 9-16 of the TMHINT(AS) material were used. More specifically, two distinct lists uttered by each of the four male or female speakers were selected for training, so the total number of sentences were 160 (4 speakers $\times$ 2 lists/speaker $\times$ 20 sentences/list) for each gender. In testing, the TMHINT(NCU) sentences under various noise conditions were processed. Lists 1-8 spoken by one male speaker and one female speaker were mixed with speech shaped noise (SSN), white noise, street noise and cafeteria babble at seven SNR levels (-10, -5, 0, 5, 10, 15 dB, and quiet) for estimating the objective speech intelligibility. As for the listening test, female sentences spoken by a professional pronouncer were used at four SNR levels (-5, 0, 5 dB SNRs and quiet) with SSN. List 16 was used in the practice session.

2) English dataset: In addition to the Mandarin sentences, an English corpus was used in the objective evaluation. Developed by research teams at Texas Instruments (TI) and the Massachusetts Institute of Technology (MIT), TIMIT is one of the most famous speech materials used for automatic speech recognition [57]. TIMIT has been previously used in many CI studies [26], [30], [58]–[60]. The complete TIMIT dataset consists of 6,300 sentences spoken by 630 different speakers from 8 major dialect regions in the United States. In this study, 300 and 80 non-repetitive sentences recorded by different speakers were used as the training and test datasets, respectively. Equal number of male and female sentences were used in the training and test datasets.

3) Bilingual dataset: In the objective evaluation, a third dataset combining Mandarin and English sentences was used. The TMHINT + TIMIT ‘bilingual’ dataset, consisting of 320 TMHINT(AS) sentences and 300 TIMIT sentences, was used for training. The objective scores were separately calculated for TMHINT(NCU) and TIMIT sentences.

E. Electrode Stimulation Patterns

Speech signals processed by ElectroNet may be understood using the electrode stimulation patterns shown in Fig. 2. The speech waveform in Fig. 2 (a) demonstrates the trisyllabic Mandarin phrase “Gōng Chéng Shī” (“Engineer” in Chinese with tones 1, 2 and 1) [61] with syllabic transitions at approximately 0.3, and 0.7 seconds. The spectrogram in Fig. 2 (b) shows that the consonants of the second and third syllables predominantly exist in the frequency range above 2 kHz and the strongest envelope amplitudes for three vowels are below 1 kHz. Similar trends can be observed in the electrodograms for the reference ACE strategy in Fig. 2 (c) and ElectroNets based on DNN, CNN, and LSTM networks in Fig. 2 (d), (e), and (f), respectively. For the first syllable

“Gōng,” the ‘g’ consonant (unaspirated velar voiceless stop /k/), which takes less than 30 ms in a previous study for the utterances by 30 native Mandarin speakers [62], is quite short and not easy to be identified. On the contrary, the second and third consonants with lengths of approximately 100 ms (0.3-0.4 s in the time axis) and 200 ms (0.7-0.9 s in the time axis) are clearly depicted in Fig. 2 (b)-(f), compared to the previous lengths for the 131-184 ms ‘ch’ consonant (aspirated retroflex voiceless affricate /tʂʰ/) and 158-256 ms ‘sh’ consonant (retroflex voiceless fricative /ʃ/) in [62]. At the lower frequencies from Channel 1 (250 Hz) to Channel 8 (1,125 Hz), the pulses are with higher amplitudes represent vowels carrying strong energy in comparison to relatively weak consonants. Some differences between ACE and ElectroNets, particularly the DNN based ElectroNet in Fig. 2 (d), can be observed between channel 9 (1,250 Hz) and Channel 17 (3,812Hz) for the three vowels. However, these pulses in the middle to high spectrum only have small effects on the speech intelligibility for vowels, which are primarily contributed by low-frequency components.

F. Experimental Factors

ElectroNets under various experimental factors were evaluated in comparison with the ACE strategy. The effects of loss function, neural network architecture, noise type, and language were examined using objective evaluation, while the DNN and CNN based ElectroNets were further investigated in the subjective listening test.

1) Loss function: Two loss functions were investigated for ElectroNet. The L_1 norm (mean absolute error, MAE) and L_2 norm (mean squared error, MSE) were measured between the electrode stimulation patterns generated by envelope detection for ACE and corresponding DNN network. Both loss functions were used to train the neural network in the ElectroNet. Computations were made to minimize the loss by updating the weights and biases of neural networks. The relationships between the objective evaluation scores of the two ElectroNets and those of the ACE strategy were compared.

2) Network architecture: Three types of neural networks were assessed to clarify the efficacy of the deep, convolutional, and recurrent network architectures on simulating the envelope detection of the ACE strategy. DNN, CNN, and LSTM networks in ElectroNets were trained using identical TMHINT(AS) sentences and the L_1 norm as the loss function. The speech intelligibility was estimated for each network. Errors and correlations were computed between the results of each ElectroNet and the ACE strategy.

3) Language: The performance of ElectroNet was examined using identical and distinct languages for training and testing datasets. Three datasets were used for training, including TMHINT, TIMIT, and TMHINT + TIMIT. DNN model based ElectroNets trained by various datasets were compared using objective evaluations. Matched and mismatched language datasets were used to investigate the model capability of covering sentences in similar and distinct compositions of consonants, vowels, and lexical tones. Additionally, the

Fig. 2. (a) Waveform and (b) spectrogram of trisyllabic Mandarin phrase “Gōng Chéng Shī” (“Engineer” in Chinese with tones 1, 2 and 1) uttered by a female speaker. The electrodegrams are processed by (c) the ACE strategy and ElectrodeNets based on (d) DNN, (e) CNN, and (f) LSTM models.

network model trained by the bilingual dataset, TMHINT + TIMIT, was examined.

4) Noise type: The robustness of deep learning based strategies against four types of noises were studied. Different types and intensities of noise were used to investigate the characteristics of ElectrodeNet compared to those of ACE. Speech shaped noise (SSN), white noise, street noise [63] and cafeteria babble [63] were mixed with clean speech at SNRs of -10, -5, 0, 5, 10, and 15 dB. The noisy sentences were normalized in sound intensities to 65 dB SPL as the clean speech. The correlations in objective speech intelligibility between the neural network model based ElectrodeNet and the ACE strategy were evaluated for these noise conditions.

G. Objective Speech Intelligibility

To efficiently evaluate the performance of the proposed ElectrodeNet, speech intelligibility was measured using two objective predictors. Objective prediction measures can be divided into either intrusive approaches using reference speech or non-intrusive approaches without reference speech. [64]–[66]. Some of the objective predictors for estimate the speech intelligibility, quality, and the perceptual SNR have been examined for assistive hearing devices [64], [67]–[70]. Among these predictors, short-time objective intelligibility (STOI) [71], and normalized covariance metric (NCM) [72], [73], were used in the present study. These two intrusive approaches estimate the perceived intelligibility using both a reference speech in quiet and a processed version of the speech. STOI has a high correlation with speech intelligibility and has

been used in CI studies on speech enhancement and sound coding strategies [24], [27], [28], [45], [64], [74]–[78]. NCM is another frequently used indicator for evaluating speech comprehension in CI sound processing [24], [28], [64], [74], [78]. It has been demonstrated that NCM is a reliable predictor for non-tonal languages as well as tonal languages, such as Mandarin [67].

H. Subjective Listening Test

Listening tests were conducted with two groups of normal-hearing listeners to investigate the DNN and CNN based ElectrodeNets. Forty native Mandarin speakers (age 20-29) were randomly divided into two test groups to evaluate the performance of the two ElectrodeNets compared to the ACE strategy. In the DNN group, 14 male and 6 female subjects were with an average age of 22.3 years, while 15 male and 5 female subjects with an average age of 22.6 were assigned to the CNN group. All the subjects had hearing thresholds no more than 25 dB HL at octave frequencies between 0.125 and 8 kHz in pure tone audiometry (PTA). The sentence recognition test for each subject took about 60-90 minutes. The experimental procedure was approved by an institutional review board (IRB).

In the listening test, the perceived intelligibility of each subject was measured for the ACE strategy and an ElectrodeNet. Each subject listened to the simulated CI sound simultaneously played to both ears via Telephonics TDH-39P headphones in a sound isolation room. The experiment was carried out with NCU-CI, a MATLAB-based CI experiment platform with

a graphical user interface in traditional Chinese [45], [46], [79]. In the practice session, each subject was instructed about the experimental procedure by being familiarized with vocoded sentences presented at a fixed and comfortable sound level that were generated from List 16 of the TMHINT material. Vocoded sentences processed by ACE and one of the ElectrodeNets, depending on the test group assignment of DNN or CNN, were presented at four SNR levels in the order of -5, 0, 5 dB and quiet. At each SNR level, the two coding strategies of ACE and ElectrodeNet were chosen in a random order. For the eight test conditions (2 strategies $\times$ 4 SNR levels), Lists 1-8 of the TMHINT dataset was selected at random without replacement, and the 20 sentences for each condition were also presented in random order. Each sentence was played once only and the subjects were encouraged to type the Chinese characters to the best of their ability. Under each test condition, the speech intelligibility was calculated as the average percentage of correct results for 200 syllables (10 syllables/sentence $\times$ 20 sentences/list). Scores were given for each correct syllable, including a homophone, a different Chinese character with the correct pronunciation.

I. Correlations between ElectrodeNet and ACE

Three relationships between ElectrodeNet and the ACE strategy under various experimental conditions were evaluated using the MSE, linear correlation coefficient (LCC) [80], and Spearman's rank correlation coefficient (SRCC) [81]. The three measures computed the errors or correlations between ElectrodeNet and ACE in STOI and NCM scores for objective and subjective speech intelligibility. Note that the MSE here was used to evaluate the differences in speech intelligibility between different coding strategies, instead of the L_2 loss function used for training network models. The closer to 0 the MSE scores, the smaller are the errors between the two strategies.

The LCC between ACE and ElectrodeNet was computed as

$$\text{LCC} = \rho = \frac{\text{cov}(SI_{ACE}, SI_{ElectrodeNet})}{\sigma_{SI_{ACE}} \sigma_{SI_{ElectrodeNet}}} \quad (1)$$

where SI is the speech intelligibility for a coding strategy, $\text{cov}(X, Y)$ is the covariance of X and Y , and σ is the standard deviation. Therefore, the LCC is the covariance of speech intelligibility for ACE and ElectrodeNet divided by the product of standard deviations of speech intelligibility for the two strategies. The SRCC is computed in a similar manner to LCC, except that the speech intelligibility scores were replaced by the ranks for each value as in

$$\text{SRCC} = \rho_s = \frac{\text{cov}(\text{Rank}(ACE), \text{Rank}(ElectrodeNet))}{\sigma_{\text{Rank}(ACE)} \sigma_{\text{Rank}(ElectrodeNet)}} \quad (2)$$

where $\text{Rank}(\cdot)$ is the resulting rank in speech intelligibility for one coding strategy. The ranks for n samples were generally distinct integer numbers 1, 2, ..., n with respect to the ascending order of speech intelligibility. Average ranks were used for repeated scores in intelligibility. For both (1) and (2), the order of the two coding strategies in equations

TABLE II
OBJECTIVE EVALUATION RESULTS FOR DNN MODELS USING L_1 AND L_2 LOSS FUNCTIONS.

Predictor	Loss Function	MSE	LCC	SRCC
STOI	L_1 norm	0.0004	0.9949	0.9920
	L_2 norm	0.0005	0.9919	0.9849
NCM	L_1 norm	0.0037	0.9927	0.9820
	L_2 norm	0.0045	0.9894	0.9712

does not change the values of LCC and SRCC. The closer the LCC and SRCC scores to 1, the higher correlations between the two strategies in performance. Usually, the LCC and SRCC values greater than 0.9 are interpreted as very high positive correlations [82]. Consequently, the desired results for ElectrodeNet are small in MSE and large in LCC and SRCC scores for a strong correlation to the ACE strategy.

III. RESULTS

The results of objective and subjective evaluations reveal that ElectrodeNet has a strong relationship to ACE in the STOI, NCM, and sentence recognition results.

A. Objective Prediction Results

The relationships between ElectrodeNets based on various models and the ACE strategy in STOI and NCM scores were compared to analyze the effects of loss function, neural network, language, and noise type.

1) *Loss function*: The objective correlations between the ACE strategy and two DNN-based ElectrodeNets built from distinct loss functions are shown in Table II, where the L_1 norm outperforms the L_2 norm. The MSEs for the L_1 norm, 0.0004 (STOI) and 0.0037 (NCM), were slightly closer to 0 compared to MSEs for L_2 norm, 0.0005 (STOI) and 0.0045 (NCM). The LCCs for the L_1 norm were 0.9949 (STOI) and 0.9927 (NCM), which were slightly greater than the LCCs for the L_2 norm, 0.9919 (STOI) and 0.9894 (NCM). Similarly, the SRCCs for the L_1 norm, 0.9920 (STOI) and 0.9820 (NCM), were slightly greater than the L_2 results, 0.9849 (STOI) and 0.9712 (NCM), respectively. The results for the L_1 norm in comparison to those of the L_2 norm showed that MSE scores were closer to 0, whereas the LCC and SRCC scores were closer to 1. That is, the relationship between the DNN based ElectrodeNet trained using the L_1 loss function and ACE was closer than that of the case using the L_2 norm. Consequently, the L_1 norm was chosen as the loss function used in the rest of this study.

2) *Network architecture*: Relationships between ACE and the DNN, CNN, and LSTM based ElectrodeNets in objective evaluations are illustrated in scatter plots for STOI in Fig. 3 and for NCM in Fig. 4 to demonstrate evidence of strong correlations. In Fig. 3, each dot on a rectangular coordinate system denotes a pair of ACE and ElectrodeNet results estimated by STOI for one sentence at $(STOI_{ACE}, STOI_{ElectrodeNet})$. Likewise, each dot in Fig. 4 denotes the results estimated by NCM for one sentence at $(NCM_{ACE}, NCM_{ElectrodeNet})$. Similar trends were demonstrated with the DNN, CNN, and

Fig. 3. Scatter plots of STOI scores for (a) DNN, (b) CNN, and (c) LSTM networks compared to ACE under SSN.

Fig. 4. Scatter plots of NCM scores for (a) DNN, (b) CNN, and (c) LSTM networks compared to ACE under SSN.

LSTM based ElectrodeNets. For an ideal ElectrodeNet identical to ACE, the dots of the scatter plot would be aligned in a straight line. The dots representing low SNR levels for STOI scores in Fig. 3 are distributed farther to the origin coordinates compared to the NCM results in Fig. 4.

The MSE, LCC and SRCC results for STOI and NCM in Fig. 3 and Fig. 4 indicate strong relationships between ACE and the three variants of ElectrodeNet. Based on the STOI measure, the MSE results were all close to 0 in a range of 0.0004–0.0007. The LCC and SRCC scores were close to the ideal result of 1 in the ranges of 0.9936–0.9949 and 0.9899–0.9920, respectively. As for the NCM results, the MSE scores were in a range of 0.0037–0.0060, and LCC and SRCC results were within ranges of 0.9911–0.9927 and 0.9791–0.9820, respectively. The MSEs for DNN (0.0004 with STOI and 0.0037 with NCM) were slightly smaller than those for CNN and LSTM (0.0006–0.0007 with STOI and 0.0051–0.0060 with NCM). Similarly, the LCCs and SRCCs for DNN were slightly greater than those for CNN and LSTM. Although a slightly stronger relationship with ACE was observed for the DNN based ElectrodeNet than those for CNN and LSTM based ElectrodeNets, the differences in the MSE, LCC, and SRCC scores were quite small. Consequently, strong relationships in objective speech intelligibility existed between the ACE strategy and each of the three ElectrodeNet architectures. More dots close to 0 in the scatter plots for NCM results in Fig. 4 compared to the STOI results in Fig. 3 indicates different estimations between the two objective intelligibility predictors.

3) Language: Strong relationships between ACE and the DNN based ElectrodeNets under various training and test dataset combinations of spoken languages are shown in Table III. For the STOI predictor, the MSE scores were between 0.0002 and 0.0006 for all datasets. The LCCs and SRCCs were greater than 0.9907 and 0.9735, respectively. For the NCM predictor, the MSE scores were in a range of 0.0022–0.0051, LCC scores were above 0.99, and SRCC scores were in a range of 0.9735–0.9875. Therefore, all ElectrodeNets were highly correlated with ACE for all the language combinations of training and test materials.

Small differences may be observed between the matched and mismatched datasets in Table III. In the matched condition, the MSE scores with the STOI predictor were 0.0004 for both Mandarin and English datasets, and the scores with NCM were 0.0037 for TMHINT and 0.0034 for TIMIT, respectively. In addition, LCCs were 0.9949 (TMHINT) and 0.9953 (TIMIT), and SRCCs were 0.9920 (TMHINT) and 0.9933 for (TIMIT). For the NCM predictor, the LCCs were 0.9927 (TMHINT) and 0.9922 (TIMIT) and the SRCCs were 0.9820 (TMHINT) and 0.9847 (TIMIT). As to the mismatched datasets, slightly larger errors and smaller correlation coefficients were obtained in comparison to the matched datasets. For the model trained with TIMIT and tested with TMHINT, the MSE scores were 0.0006 for STOI and 0.0051 for NCM, the LCC scores were 0.9926 for STOI and 0.9907 for NCM, and the SRCC scores were 0.9868 for STOI and 0.9735 for NCM. On the contrary, the results using the TMHINT training dataset and the TIMIT test dataset (MSE = 0.0004, LCC = 0.9946, and SRCC = 0.9915 for STOI, and MSE = 0.0034, LCC = 0.9924, and SRCC = 0.9834 for NCM) were closer to the results under the matched conditions and the scores for ideal linear relationships (MSE = 0 and LCC = SRCC = 1). Therefore, DNN models within ElectrodeNet trained with matched languages of datasets had stronger relationships with ACE compared to the mismatched datasets, especially when using English speech for training and Mandarin speech for testing.

As for the bilingual TMHINT + TIMIT training materials in Table III, stronger correlation scores are shown compared to the conditions using training datasets of only one spoken language. For the STOI predictor in Table III, MSEs (0.0002 and 0.0003) were smaller and LCCs (0.9961 and 0.9970) and SRCCs (0.9939 and 0.9965) were greater than the results of all the other matched and mismatched conditions. The differences between the matched and mismatched languages in training and test datasets were relatively small.

4) Noise type: The DNN based ElectrodeNet exhibits strong correlations with the ACE strategy in the STOI and NCM scores under four types of noises as listed in Table IV. The MSE, LCC, and SRCC scores were quite close between SSN, white noise, street noise, and cafeteria babble for the seven SNR levels aforementioned. For STOI, the MSE results were in a range of 0.0004–0.0009 with all the four types of noises. In addition, the LCCs and SRCCs were all greater than 0.98. As to NCM, the MSE results were in a range of 0.0033–0.0066, while the LCC and SRCC results were above 0.9729 and 0.9627, respectively. Hence, ElectrodeNet

TABLE III

OBJECTIVE EVALUATION RESULTS FOR VARIOUS TRAINING AND TEST SPEECH MATERIALS USING DNNs WITH SSN

Predictor	Training Materials	Test Materials	MSE	LCC	SRCC
STOI	TMHINT	TMHINT	0.0004	0.9949	0.9920
		TIMIT	0.0004	0.9946	0.9915
	TIMIT	TMHINT	0.0006	0.9926	0.9868
		TIMIT	0.0004	0.9953	0.9933
	TMHINT+TIMIT	TMHINT	0.0003	0.9961	0.9939
		TIMIT	0.0002	0.9970	0.9965
NCM	TMHINT	TMHINT	0.0037	0.9927	0.9820
		TIMIT	0.0034	0.9924	0.9834
	TIMIT	TMHINT	0.0051	0.9907	0.9735
		TIMIT	0.0032	0.9922	0.9847
	TMHINT+TIMIT	TMHINT	0.0028	0.9939	0.9847
		TIMIT	0.0022	0.9938	0.9875

Note: TMHINT-AS sentences were used for training and TMHINT-NCU sentences were used for testing.

TABLE IV

OBJECTIVE EVALUATION RESULTS FOR THE DNN MODEL UNDER SSN, WHITE NOISE, STREET NOISE, AND CAFETERIA BABBLE.

Predictor	Noise Type	MSE	LCC	SRCC
STOI	SSN	0.0004	0.9949	0.9920
	white noise	0.0006	0.9879	0.9869
	street noise	0.0009	0.9898	0.9837
	cafeteria babble	0.0005	0.9932	0.9887
NCM	SSN	0.0037	0.9927	0.9820
	white noise	0.0033	0.9872	0.9852
	street noise	0.0066	0.9729	0.9627
	cafeteria babble	0.0056	0.9895	0.9791

and ACE were correlated in the presence of all the four types of noises. The largest error and smallest correlation coefficients were observed with the street noise, except for the LCC results predicted using STOI. Small differences in the MSE, LCC, and SRCC results suggested strong relationships between ElectrodeNet and ACE in speech intelligibility under different noises.

B. Subjective Listening Test

The listening test scores in Fig. 5 suggest that the DNN and CNN based ElectrodeNets strongly correlate with ACE. The MSE results for DNN and CNN were 0.0163 and 0.0196, respectively. The LCC for DNN, 0.9666, was greater than that for CNN, 0.9594, while the SRCC for DNN, 0.9194, was smaller than that for CNN, 0.9371. Small MSEs close to 0 and large correlation coefficients above 0.9 indicated strong relationships with ACE for both DNN and CNN based ElectrodeNets.

The scatter plots in Fig. 5 demonstrate positive relationships between ACE and the two ElectrodeNets based on DNN and CNN. Each dot at $(SI_{ACE}, SI_{ElectrodeNet})$ represents the subjective speech intelligibility for two distinct lists separately processed by ACE and ElectrodeNet, while each dot in Fig. 3 and Fig. 4 represents a pair of objective scores for an identical sentence processed by the two coding strategies. About 50% of the dots concentrate in the top-right corner correspond to sentences at 5 dB SNR and in quiet with recognition scores over 80%. Another 25% of dots for low recognition results at

Fig. 5. Scatter plots of listening test results for (a) DNN and (b) CNN compared to ACE under SSN.

an SRN of -5 dB are concentrated at the bottom-left corner. The rest of the dots in the middle regions are primarily for noisy speech at 0 dB SNR. Most of the dots are located at the bottom-right side of the dotted diagonal line, which indicates more sentences are scored higher in sentence recognition with ACE than with DNN and CNN based ElectrodeNets. The correlations between ACE and ElectrodeNet in subjective listening test results (MSEs ≤ 0.0196 , LCCs ≥ 0.9594 , SRCCs ≥ 0.9194) were strong.

IV. DISCUSSION

Findings of utilizing deep learning based ElectrodeNet to emulate the conventional ACE coding strategy are discussed in this section.

A. Objective Results

The two DNN based ElectrodeNets trained using different loss functions show better results with the L_1 norm than with the L_2 norm. The L_2 norm is sensitive to exceptional results or outliers in the training data. However, the L_1 loss function helps shape a more generalized network model, which might better mimic the overall electrode magnitudes of ACE. Therefore, the L_1 norm is an appropriate objective function in training a deep learning based CI coding strategy.

Strong correlations between DNN, CNN, and LSTM based ElectrodeNets and the ACE strategy implies that the three nonlinear neural networks in ElectrodeNets can function comparable to the linear processing. In contrast, the envelope detection of ACE can be considered as a simplified version of neural network. The L-to-M envelope detection process can be regarded as a special case of a fully-connected network with fixed weights, no hidden layers, no biases, and bypassed activation functions. Because all the three network models containing dense layers can be regarded as general and extended forms of the ACE strategy, they can perform comparable to ACE using nonlinear operations. A slightly stronger relationship to ACE for DNN based ElectrodeNet compared to CNN and LSTM based ElectrodeNets may be owing to the closer architecture of its neural network to the envelope detection of ACE.

All the results for ElectrodeNets using the matched, mismatched, and bilingual datasets were closely related to those of the ACE strategy. Because the envelope detection of ACE is relative simple in computation complexity, it can be modeled as a DNN network using an adequate quantity of training data. In addition to the experiments using matched languages, close relationships between ElectrodeNet and ACE were also shown in the mismatched language condition using Mandarin speech for training and English speech for testing. Indeed, English language uses intonation and prosody, but its changes in tone are relatively small compared to tonal languages, such as Mandarin. Therefore, the observations showed that when ElectrodeNet was trained by the TIMIT corpus, slightly larger gaps existed in the correlation to ACE compared to that of the TMHINT trained ElectrodeNet for mismatched test datasets. In addition, when ElectrodeNet was trained using the TMHINT + TIMIT dataset, slightly better scores were obtained for both Mandarin and English sentences than those of the models trained by a single language. A possible reason is that the bilingual dataset provided more speech cues, such as suprasegmental information for the resulting model to better cover diverse speech features. Therefore, future deep learning based coding strategies are suggested to be trained using a variety of spoken languages.

The objective evaluation results revealed that ElectrodeNet had a strong positive linear relationship with ACE under four different types of noises. The DNN based ElectrodeNet trained only with clean speech was shown to have high positive correlations to ACE under various noise conditions. The scores with SSN generally showed a closer relationship between ElectrodeNet and ACE than others because the SSN, essentially generated from speech, is closer to the training speech materials than white and street noises. Compared to SSN, slightly lower correlations between ElectrodeNet and ACE were observed for cafeteria babble, which includes not only speech components, but also non-vocal noises unlike speech-based training materials. White noise with a flat spectrum had similar influences across the M electrodes and produced close effects on ElectrodeNet and ACE, whereas street noise with an uneven power spectrum caused the two strategies to behave dissimilarly. Correlations between ElectrodeNet and ACE under different noises were not identical, but still small in differences. Therefore, ElectrodeNet was capable to model the general characteristics of ACE under various noise conditions.

B. Subjective Results

The sentence recognition results of positive correlations between ACE and ElectrodeNet proved that it is possible for a deep learning based coding strategy to imitate the ACE strategy. At 5 dB SNR and in the quiet condition, sentence recognition scores close to 100% indicated that ElectrodeNet was capable of encoding sufficient information to provide speech intelligibility comparable with that of ACE in low-noise conditions. Reasonably high correlations were still obtained between ACE and ElectrodeNets for both DNN and CNN groups. The similar performance between DNN and CNN may be owing to the present method of model training.

If more temporal and spectral information can be referenced during training, CNN and the other advanced networks may be able to take advantage of their strengths. In summary, the use of DNN and CNN networks in ElectrodeNets may adequately model the envelope detection function of the ACE strategy in CI simulations with normal-hearing subjects.

C. General Discussions

The utilization of deep learning in the CI coding strategy has provided new insights into the ‘AI + CI’ research. AI not only transforms hearing healthcare and research [9], but also has a significant potential for breakthroughs to the CI coding strategy limited by present knowledge. Therefore, this study attempts to replace the traditional signal processing algorithm of the ACE strategy with artificial neural networks. ElectrodeNets using DNN, CNN, and LSTM models have strong correlations with the ACE strategy in terms of objective and subjective performance using CI simulations. Therefore, the feasibility of applying deep learning in encoding speech for CI listening has been confirmed. With more advanced network architectures and training methods, deep learning based coding strategies may outperform coding strategies based on traditional signal processing techniques. For example, the present envelop detection is implemented using the conventional power-sum and vector-sum methods to combine the L bins into M channels [47], which may be surpassed by a more advanced neural network with proper interconnections. Furthermore, the complexity of deep learning algorithms will not continue to be an issue for hearing assistive devices with the acceleration of computing power. Using the analogy of commercial hearing aids with DNN models [21], CI sound processors will also be capable of utilizing deep learning. Therefore, the availability of AI-powered CIs and other related neural prosthetic devices can be expected in the near future.

On the journey towards an AI based sound coding strategy, safety and ethics are important issues for researchers. For example, the channel selection, presentation level, and pulse patterns must comply with safety limits to prevent biological damage and uncomfortable over stimulations [83]. These configurations are crucial for conducting clinical experiments on deep learning based coding strategies. In addition, AI raises ethical issues, such as privacy concerns in collecting user data for re-training models and further research. Consequently, equal and effective communications are required between researchers, practitioners, and CI users in collecting personal information and expectations in using AI technologies. Present guidelines related to research ethics and IRB approvals may be further developed to reinforce moral values, considerations, and regulations for researchers and institutes in the ‘AI + CI’ field [84]–[86].

More investigations on ElectrodeNet is ongoing. Traditional signal processing stages of CI coding strategies such as preprocessing, FFT, and channel selection not covered in this study will also be employed using neural networks. More advanced network architectures and methods of supervised training need to be evaluated. Appropriate preparation of datasets, such as noisy speech with clean electrode stimulation patterns, may be used

to train models supporting speech enhancement. Additional multimodal information, such as contralateral sounds, visual cues, tactile vibrations, and auditory responses, may also be processed in the neural networks [28], [87]–[90]. Furthermore, clinical experiments for a deep learning based coding strategy with CI listeners need to be conducted. The innovation of ElectroNet will be continued to make some differences to CI listening.

V. CONCLUSION

This study proposes ElectroNet, a deep learning based sound coding strategy for the cochlear implant. The performance of ElectroNet is comparable to that of the ACE strategy in signal processing with the exception of using an artificial neural network to replace the conventional envelope detection. Objective evaluations using the STOI and NCM intelligibility predictors have shown the following results: (1) Positive linear relationships between the ACE strategy and the ElectroNets based on DNN, CNN, and LSTM models. (2) The L_1 norm outperformed the L_2 norm in training the DNN based ElectroNet. (3) The diversity in training materials, such as tonal cues and multiple languages, increased the correlations between ElectroNet and ACE. (4) Strong relationships were observed between ACE and the DNN based ElectroNet with SSN, white noise, street noise, and cafeteria babble at seven SNR levels. Even the lowest LCC and SRCC scores in objective evaluations estimated by NCM for street noise were greater than 0.97 and 0.96, respectively. In addition to the objective results, the listening tests using CI simulations revealed that both DNN and CNN based ElectroNets had strong linear relationships to ACE in Mandarin sentence recognition scores, where all the LCC and SRCC scores were above 0.9 and MSEs were below 0.02. Both objective and subjective results have suggested the substantial feasibility of using ElectroNet as a replacement for the ACE strategy based on traditional signal processing. ElectroNet may be further improved by including more signal processing stages. The approach and findings in this study provide some insights and research directions toward the inevitable trend of ‘AI + CI’ innovation.

REFERENCES

- [1] F.-G. Zeng, S. Rebscher, W. Harrison, X. Sun, and H. Feng, “Cochlear implants: system design, integration, and evaluation,” *IEEE Rev. Biomed. Eng.*, vol. 1, pp. 115–142, 2008.
- [2] G. M. Clark, “The multi-channel cochlear implant: Multi-disciplinary development of electrical stimulation of the cochlea and the resulting clinical benefit,” *Hear. Res.*, vol. 322, pp. 4–13, 2015.
- [3] B. S. Wilson, “Getting a decent (but sparse) signal to the brain for users of cochlear implants,” *Hear. Res.*, vol. 322, pp. 24–38, 2015.
- [4] F.-G. Zeng, “Challenges in improving cochlear implant performance and accessibility,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1662–1664, 2017.
- [5] B. L. Hamel, K. Vasil, V. Shafiro, A. C. Moberly, and M. S. Harris, “Safety-relevant environmental sound identification in cochlear implant candidates and users,” *Laryngoscope*, vol. 130, no. 6, pp. 1547–1551, 2020.
- [6] A. Taddei, E. A. López, and R. A. R. Reyes, “Children with Hearing Disabilities during the Pandemic: Challenges and Perspectives of Inclusion,” *Education Sciences & Society-Open Access*. 12(1). pp.178–196, 2021.
- [7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [9] N. A. Lesica, N. Mehta, J. G. Manjaly, L. Deng, B. S. Wilson, and F.-G. Zeng, “Harnessing the power of artificial intelligence to transform hearing healthcare and research,” *Nat. Mach. Intell.*, vol. 3, no. 10, pp. 840–849, 2021.
- [10] M. G. Crowson, V. Lin, J. M. Chen, and T. C. Chan, “Machine learning and cochlear implantation—a structured review of opportunities and challenges,” *Otol. Neurotol.*, vol. 41, no. 1, pp. e36–e45, 2020.
- [11] G. Guerra-Jiménez, Á. R. De Miguel, J. C. F. González, S. A. B. Barreiro, D. P. Plasencia, and Á. R. Macías, “Cochlear implant evaluation: Prognosis estimation by data mining system,” *J. Int. Adv. Otol.*, vol. 12, no. 1, pp. 1–7, 2016.
- [12] X. Gao, D. B. Grayden, and M. D. McDonnell, “Modeling electrode place discrimination in cochlear implant stimulation,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2219–2229, 2016.
- [13] J. Pile, G. B. Wana, and N. Simaan, “Robot-assisted perception augmentation for online detection of insertion failure during cochlear implant surgery,” *Robotica*, vol. 35, no. 7, pp. 1598–1615, 2017.
- [14] J. Torresen, A. H. Iversen, and R. Greisiger, “Data from past patients used to streamline adjustment of levels for cochlear implant for new patients,” in *Proc. of 2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2016, pp. 1–7.
- [15] M. Meeuws, D. Pascoal, I. Bermejo, M. Artaso, G. De Ceulaer, and P. J. Govaerts, “Computer-assisted CI fitting: Is the learning capacity of the intelligent agent fox beneficial for speech understanding?” *Cochlear Implants Int.*, vol. 18, no. 4, pp. 198–206, 2017.
- [16] J. M. Desmond, L. M. Collins, and C. S. Throckmorton, “Using channel-specific statistical models to detect reverberation in cochlear implant stimuli,” *J. Acoust. Soc. Amer.*, vol. 134, no. 2, pp. 1112–1120, 2013.
- [17] K. Chu, C. Throckmorton, L. Collins, and B. Mainsah, “Using machine learning to mitigate the effects of reverberation and noise in cochlear implants,” in *Proc. Mtgs. Acoust.*, vol. 33, no. 1, pp. 1–13, 2018.
- [18] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Ganot, and C.-A. Deledalle, “Machine learning in acoustics: Theory and applications,” *J. Acoust. Soc. Amer.*, vol. 146, no. 5, pp. 3590–3628, 2019.
- [19] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, “Deep learning for audio signal processing,” *IEEE J. Sel. Top. Signal. Process.*, vol. 13, no. 2, pp. 206–219, 2019.
- [20] D. Wang, “Deep learning reinvents the hearing aid,” *IEEE Spectr.*, vol. 54, no. 3, pp. 32–37, 2017.
- [21] D. Beck, “Hearing, listening and deep neural networks in hearing aids,” *J. Otolaryngol.-ENT Res.*, vol. 13, no. 1, pp. 5–8, 2021.
- [22] F. Bolner, T. Goehring, J. Monaghan, B. Van Dijk, J. Wouters, and S. Bleack, “Speech enhancement based on neural networks applied to cochlear implant coding strategies,” in *2016 Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2016, pp. 6520–6524.
- [23] T. Goehring, F. Bolner, J. J. Monaghan, B. Van Dijk, A. Zarowski, and S. Bleack, “Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users,” *Hear. Res.*, vol. 344, pp. 183–194, 2017.
- [24] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, “A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1568–1578, 2017.
- [25] Y.-H. Lai, Y. Tsao, X. Lu, F. Chen, Y.-T. Su, K.-C. Chen, Y.-H. Chen, L.-C. Chen, L. P.-H. Li, and C.-H. Lee, “Deep learning-based noise reduction approach to improve speech intelligibility for cochlear implant recipients,” *Ear Hear.*, vol. 39, no. 4, pp. 795–809, 2018.
- [26] N. Mamun, S. Khorram, and J. H. Hansen, “Convolutional neural network-based speech enhancement for cochlear implant recipients,” in *Interspeech*, 2019.
- [27] N. Y.-H. Wang, H.-L. S. Wang, T.-W. Wang, S.-W. Fu, X. Lu, H.-M. Wang, and Y. Tsao, “Improving the intelligibility of speech for simulated electric and acoustic stimulation using fully convolutional neural networks,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 184–195, 2021.
- [28] R.-Y. Tseng, T.-W. Wang, S.-W. Fu, C.-Y. Lee, and Y. Tsao, “A study of joint effect on denoising techniques and visual cues to improve speech intelligibility in cochlear implant simulation,” *IEEE Trans. Cogn. Develop. Syst.*, vol. 13, no. 4, pp. 984–994, 2021.
- [29] T. Goehring, M. Keshavarzi, R. P. Carlyon, and B. C. Moore, “Using recurrent neural networks to improve the perception of speech in non-

- stationary noise by people with cochlear implants,” *J. Acoust. Soc. Amer.*, vol. 146, no. 1, pp. 705–718, 2019.
- [30] K. Chu, L. Collins, and B. Mainsah, “A causal deep learning framework for classifying phonemes in cochlear implants,” in *2021 Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 6498–6502, 2021.
 - [31] W. Nogueira, T. Gajecski, B. Krüger, J. Janer, and A. Büchner, “Development of a sound coding strategy based on a deep recurrent neural network for monaural source separation in cochlear implants,” in *Proc. of the 12th ITG Symp. on Speech Comm.*, 2016.
 - [32] J. Pons, J. Janer, T. Rode, and W. Nogueira, “Remixing music using source separation algorithms to improve the musical experience of cochlear implant users,” *J. Acoust. Soc. Amer.*, vol. 140, no. 6, pp. 4338–4349, 2016.
 - [33] T. Gajecski and W. Nogueira, “Deep learning models to remix music for cochlear implant users,” *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. 3602–3615, 2018.
 - [34] W. Nogueira, A. Nagathil, and R. Martin, “Making music more accessible for cochlear implant listeners: recent developments,” *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 115–127, 2018.
 - [35] J. Wouters, H. J. McDermott, and T. Francart, “Sound coding in cochlear implants: From electric pulses to hearing,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 67–80, 2015.
 - [36] 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, 1991.
 - [37] A. E. Vandali, L. A. Whitford, K. L. Plant, G. M. Clark, “Speech perception as a function of electrical stimulation rate: using the Nucleus 24 cochlear implant system,” *Ear Hear.*, vol. 21, no. 6, pp. 608–624, 2000.
 - [38] W. Nogueira, A. Büchner, T. Lenarz, and B. Edler, “A psychoacoustic “NofM”-type speech coding strategy for cochlear implants,” *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 18, pp. 3044–3095, 2005.
 - [39] A. Büchner, W. Nogueira, B. Edler, R.-D. Battmer, and T. Lenarz, “Results from a psychoacoustic model-based strategy for the Nucleus-24 and Freedom cochlear implants,” *Otol. Neurotol.*, vol. 29, no. 2, pp. 189–192, 2008.
 - [40] D. Riss, J.-S. Hamzavi, A. Selberherr, A. Kaider, M. Blineder, V. Starlinger, W. Gstoettner, and C. Arnoldner, “Envelope versus fine structure speech coding strategy: A crossover study,” *Otol. Neurotol.*, vol. 32, no. 7, pp. 1094–1101, 2011.
 - [41] J. Müller, S. Brill, R. Hagen, A. Moeltner, S.-J. Brockmeier, T. Stark, S. Helbig, J. Maurer, T. Zahner, C. Zierhofer et al., “Clinical trial results with the MED-EL fine structure processing coding strategy in experienced cochlear implant users,” *ORL*, vol. 74, no. 4, pp. 185–198, 2012.
 - [42] M. Brendel, A. Buechner, B. Krueger, C. Frohne-Buechner, and T. Lenarz, “Evaluation of the Harmony sound processor in combination with the speech coding strategy HiRes 120,” *Otol. Neurotol.*, vol. 29, no. 2, pp. 199–202, 2008.
 - [43] J. B. Firszt, L. K. Holden, R. M. Reeder, and M. W. Skinner, “Speech recognition in cochlear implant recipients: comparison of standard HiRes and HiRes 120 sound processing,” *Otol. Neurotol.*, vol. 30, no. 2, pp. 146–152, 2009.
 - [44] D. Schramm, J. Chen, D. P. Morris, N. Shoman, D. Philippon, P. Cayé-Thomasen, M. Hoen, C. Karoui, A. Laplante-Lévesque, and D. Gnansia, “Clinical efficiency and safety of the Oticon Medical Neuro cochlear implant system: a multicenter prospective longitudinal study,” *Expert Review of Medical Devices*, vol. 17, no. 9, pp. 959–967, 2020.
 - [45] E. H.-H. Huang, K.-H. Hung, Y. Tsao, and H.-C. Lin, “Combination and Comparison of Sound Coding Strategies using Cochlear Implant Simulation with Mandarin Speech,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2407–2416, 2021.
 - [46] E. H.-H. Huang, K.-H. Hung, Y. Tsao, and C.-M. Wu, “ElectrodeNet – artificial intelligence based sound coding strategy for cochlear implants,” in *Proc. of the 12th Asia Pacific Symposium on Cochlear Implants and Related Sciences (APSCI2019)*, Tokyo, Japan, pp. O2-5, 2019.
 - [47] B. A. Swanson, “Pitch perception with cochlear implants,” Ph.D. dissertation, Dept. Otol., Univ. Melbourne, VIC, Australia, 2008.
 - [48] M. F. Dorman, P. C. Loizou, A. J. Spahr, and E. Maloff, “A comparison of the speech understanding provided by acoustic models of fixed-channel and channel-picking signal processors for cochlear implants,” *J. Speech Lang. Hear. Res.*, vol. 44, pp. 1–6, 2002.
 - [49] 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, 1995.
 - [50] B. Swanson and H. Mauch, *Nucleus Matlab Toolbox 4.20 software user manual*, Cochlear Ltd, Lane Cove NSW, Australia, 2006.
 - [51] P. J. Werbos, “Applications of advances in nonlinear sensitivity analysis,” in *Proc. of the 10th IFIP Conf.*, pp. 762–770, 1981.
 - [52] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural netw.*, vol. 2, no. 5, pp. 359–366, 1989.
 - [53] Y. LeCun, “Generalization and network design strategies,” Technical Report CRG-TR-89-4, Univ. Toronto, pp. 1–19, 1989.
 - [54] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
 - [55] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 9, 533–536, 1986.
 - [56] L. L. Wong, S. D. Soli, S. Liu, N. Han, and M. -W. Huang, “Development of the Mandarin hearing in noise test (MHINT),” *Ear Hear.*, vol. 28, no. 2, pp. 70S–74S, 2007.
 - [57] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1,” *NASA STI/Recon Tech. Rep. N*, vol. 93, no. 27403, 1993.
 - [58] S. E. King, J. B. Firszt, R. M. Reeder, L. K. Holden, and M. Strube, “Evaluation of TIMIT sentence list equivalency with adult cochlear implant recipients,” *J. Am. Acad. Audiol.*, vol. 23, no. 05, pp. 313–331, 2012.
 - [59] R. H. Gifford, L. Loiselle, S. Natale, S. W. Sheffield, L. W. Sunderhaus, M. S. Dietrich, and M. F. Dorman, “Speech understanding in noise for adults with cochlear implants: Effects of hearing configuration, source location certainty, and head movement,” *J. Speech Lang. Hear.*, vol. 61, no. 5, pp. 1306–1321, 2018.
 - [60] F. Chen, and Y. Hu, Y. “Segmental contributions to cochlear implant speech perception,” *Speech Commun.*, vol. 106, pp. 79–84, 2019.
 - [61] S. L. Nissen, R. W. Harris, and K. B. Slade, “Development of speech reception threshold materials for speakers of Taiwan Mandarin,” *Int. J. Audiol.*, vol. 46, no. 8, pp. 449–458, 2007.
 - [62] J. Y. Jeng, “Butong yanyu sudu, fayu danwei he fayu weizhi dui guoyu induan shichang de yingxiang” [The influence of speaking rate, utterance unit and position on segmental duration of Mandarin], *J. Natl. Univ. Tainan*, vol. 39, no. 1, pp. 161–185, 2006.
 - [63] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed., Boca Raton, FL, USA: CRC Press, 2013.
 - [64] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, “Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 114–124, 2015.
 - [65] N. Mamun, M. S. Zilany, J. H. Hansen, and E. E. Davies-Venn, “An intrusive method for estimating speech intelligibility from noisy and distorted signals,” *J. Acoust. Soc. Amer.*, vol. 150, no. 3, pp. 1762–1778, 2021.
 - [66] Y. Feng, and F. Chen, “Nonintrusive objective measurement of speech intelligibility: A review of methodology,” *Biomed. Signal Process. Control*, vol. 71, 103204, pp. 1–14, 2022.
 - [67] F. Chen and P. C. Loizou, “Predicting the intelligibility of vocoded and wideband Mandarin Chinese,” *J. Acoust. Soc. Amer.*, vol. 129, no. 5, pp. 3281–3290, 2011.
 - [68] J. F. Santos, S. Cosentino, O. Hazrati, P. C. Loizou, and T. H. Falk, “Objective speech intelligibility measurement for cochlear implant users in complex listening environments,” *Speech Commun.*, vol. 55, no. 7-8, pp. 815–824, 2013.
 - [69] G. D. Watkins, B. A. Swanson, and G. J. Suaning, “An evaluation of output signal to noise ratio as a predictor of cochlear implant speech intelligibility,” *Ear Hear.*, vol. 39, no. 5, pp. 958–968, 2018.
 - [70] G. D. Watkins, B. A. Swanson, and G. J. Suaning, “Prediction of Individual Cochlear Implant Recipient Speech Perception With the Output Signal to Noise Ratio Metric,” *Ear Hear.*, vol. 41, no. 5, pp. 1270–1281, 2020.
 - [71] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.
 - [72] I. Holube and B. Kollmeier, “Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model,” *J. Acoust. Soc. Amer.*, vol. 100, no. 3, pp. 1703–1716, 1996.
 - [73] R. L. Goldsworthy and J. E. Greenberg, “Analysis of speech-based speech transmission index methods with implications for nonlinear operations,” *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3679–3689, 2004.

- [74] H. Hu, M. E. Lutman, S. D. Ewert, G. Li, and S. Bleeck, "Sparse nonnegative matrix factorization strategy for cochlear implants," *Trends Hear.*, vol. 19, pp. 1–16, 2015.
- [75] E. A. Lopez-Poveda, and A. Eustaquio-Martín, "Objective speech transmission improvements with a binaural cochlear implant sound-coding strategy inspired by the contralateral medial olivocochlear reflex," *J. Acoust. Soc. Amer.*, vol. 143, no. 4, pp. 2217–2231, 2018.
- [76] E. A. Lopez-Poveda, A. Eustaquio-Martín, M. J. Fumero, J. M. Gorospe, R. P. López, M. A. G. Revilla, R. Schatzer, P. Nopp, and J. S. Stohl, "Speech-in-noise recognition with more realistic implementations of a binaural cochlear-implant sound coding strategy inspired by the medial olivocochlear reflex," *Ear Hear.*, vol. 41, no. 6, pp. 1492–1510, 2020.
- [77] F. Langner, A. Büchner, and W. Nogueira, "Evaluation of an adaptive dynamic compensation system in cochlear implant listeners," *Trends Hear.*, vol. 24, pp. 1–13, 2020.
- [78] G. L. Mourão, M. H. Costa, and S. Paul, "Speech intelligibility for cochlear implant users with the MMSE noise-reduction time-frequency mask," *Biomed Signal Process Control*, vol. 60, p. 101982, 2020.
- [79] C.-M. Wu, K.-Y. H. Huang, and H.-C. Lin, "Effects of channel number, stimulation rate, and electroacoustic stimulation of cochlear implant simulation on Chinese speech recognition in noise," in *Proc. of the 7-th Asia Pacific Symposium on Cochlear Implants and Related Sciences (APSCI2009)*, Singapore, pp. RS2B-7, 2009.
- [80] K. Pearson, "Notes on The History of Correlation," *Biometrika*, vol. 13, no. 1, pp. 25–45, 10 1920.
- [81] C. Spearman, "The proof and measurement of association between two things," *Am. J. Psychol.*, vol. 15, no. 1, pp. 72–101, 1904.
- [82] M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi Med. J.*, vol. 24, no. 3, pp. 69–71, 2012.
- [83] R. Y. Litovsky, M. J. Goupell, A. Kan, and D. M. Landsberger, "Use of research interfaces for psychophysical studies with cochlear-implant users," *Trends Hear.*, vol. 21, pp. 1–15, 2017.
- [84] J. Parvizi, T. D. Tarity, K. Conner, and J. B. Smith, "Institutional review board approval: why it matters," *J. Bone Jt. Surg.*, vol. 89, no. 2, pp. 418–426, 2007.
- [85] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *Minds Mach.*, vol. 30, no. 1, pp. 99–120, 2020.
- [86] J. W. Wasmann, C. Lanting, W. Huinck, E. Mylanus, J. van der Laak, P. Govaerts, D. W. Swanepoel, D. R. Moore and D. L. Barbour, "Computational Audiology: New Approaches to Advance Hearing Health Care in the Digital Age," *Ear Hear.*, vol. 42, no. 6, pp. 1499–1507, 2021.
- [87] G. Kurien, E. Hwang, K. Smilsky, L. Smith, V. Y. Lin, J. Nedzelski, and J. M. Chen, "The benefit of a wireless contralateral routing of signals (CROS) microphone in unilateral cochlear implant recipients," *Otol. Neurotol.*, vol. 40, no. 2, e82–e88, 2019.
- [88] J. Huang, B. Sheffield, P. Lin, and F. G. Zeng, "Electro-Tactile Stimulation Enhances Cochlear Implant Speech Recognition in Noise," *Sci. Rep.*, vol. 7, no. 1, pp. 1–5, 2017.
- [89] M. D. Fletcher, "Using haptic stimulation to enhance auditory perception in hearing-impaired listeners," *Expert Rev. Med. Devices*, vol. 18, no. 1, pp. 63–74, 2021.
- [90] M. J. Fumero, A. Eustaquio-Martín, J. M. Gorospe, R. P. López, M. A. G. Revilla, L. Lassaletta, L., R. Schatzer, P. Nopp, J. S. Stohl, and E. A. Lopez-Poveda, "A state-of-the-art implementation of a binaural cochlear-implant sound coding strategy inspired by the medial olivocochlear reflex," *Hear. Res.*, vol. 409, 108320, pp. 1–14, 2021.