Leibniz Universität Hannover

Medizinische Hochschule Hannover

Institut für Mechatronische Systeme
Leibniz Universität Hannover

Institut für Mechatronische Systeme

Towards end-to-end deep learning sound coding strategies for cochlear implants

Masterarbeit

Yichi Zhang
Matr.-Nr: 10027133

First Examiner: Dr.-Ing. habil. Hans-Georg Jacob
Second Examiner: Prof. Dr.-Ing. Waldo Nogueira
Supervisor: Sontje Ihler, M.Sc.
Co-Supervisor: Tom Gajecki, M.Sc.

June 2022

Abstract

Cochlear implants (CI) can provide improved sound perception for patients with severe sensorineural hearing loss, but the speech understanding for CI users is still greatly challenged in noisy environments. To overcome this problem, front-end speech enhancement technologies have been developed, but there is still much room to improve the speech intelligibility of CI users.

In this thesis, an end-to-end deep learning speech denoising and coding strategy is presented. Deep neural networks (DNN) are built to simulate the advanced combination encoder (ACE) sound coding strategy and perform speech denoising. Inspired by previous studies, which showed that CI users are more sensitive to distortion of the band selection compared to the distortion of signal envelopes, the DeepACE was optimized by training with a custom loss function and architecture aiming to suppress the incorrect band selection. The commercially used Wiener filter and a deep learning based front-end Conv-TasNet model were selected as baseline models. The speech enhancement performance was investigated through objective instrumental measures such as the signal-to-noise-ratio (SNR) improvement, the short-time objective intelligibility (STOI) measure and linear correlation coefficients between predicted electrodograms through selected algorithms and target electrodograms through ACE. Furthermore, listening tests with 8 CI users were conducted to assess the potential benefits in speech intelligibility for different algorithms.

Results show that electrodogram domain the optimized DeepACE models obtained great benefits in SNR improvement and the enhanced electrodograms had strong correlation with the clean ones. At the same time, the STOI scores rose remarkably in vocoded audio domain. Moreover, results in CI users show that optimized DeepACE provided 20% more speech understanding. All the improvement was consistent across different speech and noise datasets, which confirmed the generalization ability and practicality.

Keywords: Cochlear Implants, Sound Coding Strategy, Deep Learning, Speech Enhancement, Source Separation

Ehrenwörtliche Erklärung

Ich versichere, dass

- ich diese Masterarbeit selbständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich sonst keiner unerlaubten Hilfe bedient habe.

Datum: 17.06.2022 Unterschrift: Yichi Zhang

Diese Arbeit wurde betreut von:

1. Prüfer: Dr.-Ing. habil. Hans-Georg Jacob

2. Prüfer: Prof. Dr.-Ing. Waldo Nogueira

Contents

List of Figures

List of Tables

List of Abbreviations

ACE Advanced Combination Encoder

ANN Artificial Neural Network

ANOVA Analysis of Variance

BCE Binary Cross-entropy

CI Cochlear Implant

CNN Convolutional Neural Network

Conv-TasNet Convolutional Time-domain Audio Separation Network

CSR Channel Stimulation Rate

DL Deep Learning

DNN Deep Neural Network

FFT Fast Fourier Transform

LCC Linear Correlation Coefficients

LGF Loudness Growth Function

LSTM Long Short-Term Memory

MCL Most Comfortable Level

MSE Mean Squared Error

NH Normal Hearing

RNN Recurrent Neural Network

SCS Sound Coding Strategy

SE Speech Enhancement

SI-SDR Scale-Invariant Source-to-Distortion Ratio

SNR Signal-to-Noise Ratio

SS Source Separation

STFT Short-Time Fourier Transform

STOI Short-Time Objective Intelligibility

TCN Temporal Convolutional Network

THL Threshold Level

TF Time-Frequency

TFS Temporal Fine Structure

WF Wiener Filter

Task of this Study

Thema: Towards end-to-end deep learning sound coding strategies for cochlear implants

Cochlear impants (CIs) allow many hearing impaired to understand speech, but CI users struggle in difficult listening conditions with significant background noise. Since most speech enhancement techniques use spectrograms as front-end processing and discard the phase, the goal of this project is to develop an end-to-end deep learning sound coding strategy for speech enhancement. The framework is based on DeepACE prototype, which is previously developed by APG Group.

Throughout the project, further investigations on the architectures and loss functions of DeepACE will be conducted respectively, in order to exceed the performance of other state of the art methods while maintaining low latency and fewer parameters. After optimization of relevent hyper-parameters, the generalization ability of the model is to be evaluated as the main concern at different voice, noise and SNR conditions by means of objective instrumental metrics. Furthermore, subjective evaluation will be done on more patients. And it will be investigated whether a new architecture that mixes clean speech and noise in a certain level will benefit subjective perception.

1 Introduction

1.1 Motivation

A cochlear implant (CI) is a surgically implanted neuroprosthetic device, which provides a patient with moderate to profound sensorineural hearing loss a modified sense of sound with an improved speech intelligibility.

In order to overcome the shortcomings of CIs including the fact that performance degenerates drastically in noisy and reverberant environments, many speech enhancement (SE) techniques have been proposed to improve speech intelligibility, e.g. (1) classic single-channel SE algorithms for CIs including spectral contrast enhancement [7][8], spectral subtraction (SS) [9], Wiener filter (WF) [10] and time-frequency (T-F) masking [11], (2) binaural sound coding strategies (BSCSs) [12][13] and (3) data-driven models, particularly the deep-learning (DL) ones which mainly contain spectral mapping-based [14][15] and masking based models [16][17].

Previously, general speech enhancement and source separation methods commonly used TF representations or spectrograms [18][19]. However, some potentially valuable information, e.g., the phase information is discarded in this progress. To solve this problem, many deep learning based approaches that directly work in time-domain are investigated. Luo et al. [20] proposed a fully-convolutional time-domain audio separation network (Conv-TasNet), a deep learning framework for end-to-end time-domain speech separation. The framework addresses the shortcomings of separation in the STFT domain, achieves state-of-the-art performance and is suitable for low latency applications.

Most recently, inspired by the aforementioned Conv-TasNet, Gajecki et al. [21] proposed a deep learning-based end-to-end CI sound coding strategy. DeepACE estimates denoised CI electrodograms from raw audio data and enhances the speech intelligibility of CI users. An electrodogram can be seen as the CI electrode stimulation response. Although DeepACE has achieved encouraging results for CI users. There is still much room for further research and improvement in terms of architecture and details. The generalization ability of the model also deserves further investigation for future application to CIs.

Besides, CI users obtain hearing perception that is much different from the acoustic hearing of normal hearing (NH) people. The main reasons are the low number of electrodes, the spread of excitation caused by electric stimulation and limitations in the perception of mid to high temporal modulations [22]. NH people are more sensitive to distortion than noise [23]. Some studies show that CI recipients are more tolerant of distortion

but very sensitive to noise [24][25][26]. But distortion are also perceived by CI users and the speech understanding is obviously affected [27]. This should also be considered in designing SE models for CI users.

In this work, an empirical study of DeepACE model was conducted. Various combinations of hyperparameters were investigated, and a custom loss function and architecture based on the hearing characteristics of CI users were proposed, expecting to achieve better performance especially for CI users and outperform other baseline models. In addition, an optimized dataset setting was adopted by using more realistic noise types and making test noise 100% unseen to the model to explore the generalization ability. Furthermore, except the conventional objective evaluation indicators, a listening test for CI users was also conducted to assess the potential benefits of DeepACE, Conv-TasNet+ACE, Wiener filter+ACE w.r.t. the clinical ACE sound coding strategy.

The deep learning algorithms were optimized based on objective instrumental measures. Among them, the signal-to-noise-ratio (SNR) improvement and linear correlation coefficients (LCC) were evaluated in the electrodogram domain, while the short-time objective intelligibility (STOI) score was calculated in the vocoded audio domain. In this work, the electrodograms were resynthesized with a sine vocoder [2]. In the end, the optimized DeepACE model was compared with the other aforementioned methods at different conditions to prove its practicality.

1.2 Aims of the study

This project aims to optimize the DeepACE model in terms of speech intelligibility for CI users and to evaluate the performance through objective instrumental measures and listening tests. The model framework is based on the previous work conducted at the Auditory Prosthetic Group [21]. The Following objectives are defined in this study:

1. Investigate more architectures and loss functions for DeepACE. Optimize the performance of other state-of-the-art methods for speech enhancement in CI users while maintaining low latency and fewer parameters.

2. Develop an optimized DeepACE model which considers the hearing characteristics of CI users.

3. Generate more realistic training and testing datasets for deep learning based speech enhancement algorithms for CIs, get more solid results about the denoising performance and the generalization ability of the models.

4. Evaluate the model objectively at different conditions in both the electrodogram domain and the vocoded audio domain. Implement the listening tests on more patients.

1.3 Structure of the Master Thesis

This thesis is organized as follows:

- Chapter 2 introduces the fundamental knowledge and the state-of-the-art techniques regarding CIs, speech enhancement for CI, objective instrumental measures and discusses the performance of the prototype model;
- Chapter 3 is about the methods and materials that are used for optimizing DeepACE, and focuses on the process of the experiment with varied settings;
- Chapter 4 presents the results in both objective measures and subjective listening tests. Some focal points are then discussed in Chapter 5;
- Chapter 6 draws the conclusions and describes the expected future work.

2 Fundamentals

2.1 Hearing Loss

Hearing is the ability to perceive sounds by detecting vibrations through the ear. The hearing system contains many components. The outer ear collects the sound that vibrates the eardrum of the middle ear while the inner ear receives these vibrations and sends them to the auditory nerve. These impulses eventually reach the brain [28].

A very important part of the hearing system is cochlea. The cochlea is a spiraled, hollow, conical chamber of bone found in the inner ear that plays a crucial role in the sense of hearing and participates in auditory transduction. Sound waves are transduced into electrical impulses that can be interpreted by the brain as individual frequencies of sound. The spiral shape of the cochlea allows for differing frequencies to stimulate specific areas along the spiral. Specific areas along the cochlea are stimulated by vibrations carried within a fluid called endolymph. The vibrations are then converted to electrical impulses in the cochlear duct through the mechanical stimulation of hair cells within the organ of Corti. These nerve impulses are carried from the cochlea to the brain for interpretation [29][30].

Hearing loss is a partial or total inability to hear, and it could be diagnosed with hearing screening or hearing tests. Clinically, loudness is expressed in the decibel's hearing level (HL), representing the sound pressure level produced at a specific frequency. The threshold for perceiving a sound at a given frequency by normal persons is 0 dB HL, and the HL is 45 to 60 dB for normal conversation. The diagnosed threshold set by World Health Organization (WHO) for a hearing loss, which refers to the minimum sound intensity that a hearing loss person could not detective as normal people, is 20 dB or more in both ears. The most recent WHO estimate report suggests that approximately 466 million people have some degree of hearing loss- almost the 5.5% of the world's population. This number is expected to rise to over 2.5 billion by 2050 [31].

Hearing loss could be classified as mild, moderate, severe, profound, or total hearing loss. It could be characterized as one of three types: conductive hearing loss, sensorineural hearing loss and mixed hearing loss (combination of the two peripheral hearing losses). The conductive hearing loss is usually caused by any possible condition that interferes with the sound wave transmission (involves outer or middle ear), and it is generally correctable. However, the sensorineural hearing loss is caused by dysfunction in the inner ear or the auditory nerve, therefore the sensorineural hearing loss is usually permanent and makes it a challenging problem [32].

2.2 Cochlear Implants

2.2.1 Specifications

A possible treatment for sensorineural hearing loss is neural prostheses. Neural prostheses for hearing restoration, particularly the cochlear implant (CI), have achieved immense success beyond original expectations. Modern CIs have multiple electrodes integrated into an array, and they will set along the scala tympani. Each electrode can potentially electrically stimulate a different set of auditory neurons in the cochlear by injecting current through the electrode. As the most successful neural prosthesis, CI has provided a partial hearing to more than hundred-thousand patients worldwide. Many of them report that the CI has substantial benefits in speech recognition and spoken language understanding after cochlear implantation [31]. However, the restored hearing is unfortunately still far from the normal hearing.

As Figure 2.1 displays, a CI consist of an external part worn beside an ear and an internal part implanted through surgery. The corresponding components along with their mechanisms are as follows:

- Microphones in the sound processor convert sound into digital information.
- Digital information is transferred to the coil and then to the implant which sits just under the skin.
- The implant converts digital information into electrical signals which are then transferred to the cochlea (inner ear).
- In the cochlea, the electrode array stimulates nerve fibres which transfers signals along the auditory nerve to the brain to be interpreted as sound.

In some cases, an additional acoustic component also delivers amplified sound to the ear canal. The sound is then transmitted along the normal auditory pathway – from the outer ear to the middle ear and inner ear.

In other cases, the sound processor and the coil can be combined into a single component that sits above and behind the ear.

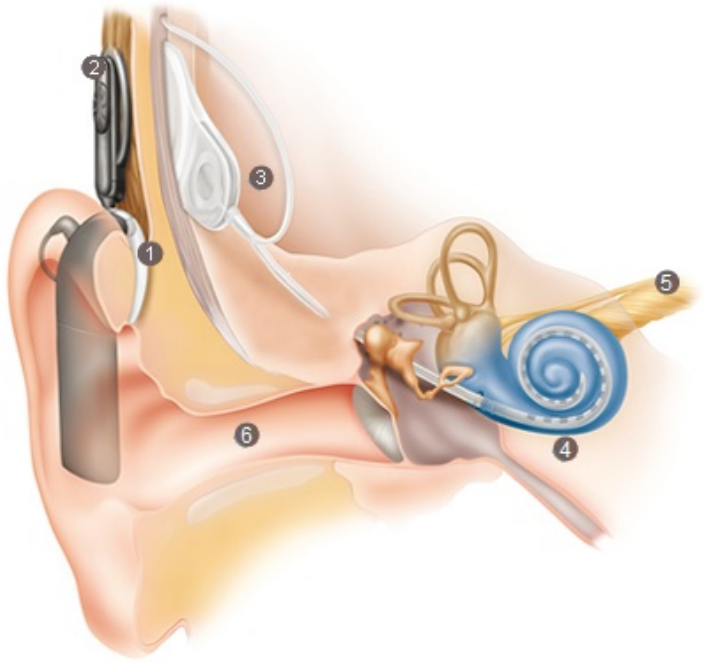

Figure 2.1: The structure of cochlear implant [1]. CI is comprised of two parts: external – sound processor (1) and coil (2) and internal – implant (3), electrode and electrode array (4). Besides, (5) is the auditory nerve and (6) is the ear canal.

2.2.2 Sound Coding Strategy

Cochlear implants convert sound to simulate the process of natural hearing. This conversion of sound is called sound coding. A sound coding strategy (SCS) is an algorithm that converts an audio signal into the electrodogram, which will be applied through the electrode array to the auditory nerve. An electrodogram can be seen as the CI electrode stimulation response, which is similar to a spectrogram, but the vertical axis indicates channel number rather than frequency, and biphasic current pulses are represented as vertical lines with amplitudes between 0 and 1 [33]. An example of electrodogram can be seen in Figure 2.2.

Current CI manufacturers offer several strategies to users [34] to provide more details about acoustic signals to CI users. For instance, the high-resolution strategy (HiRes), the spectral resolution strategy (SpecRes), the continuous interleaved sampling strategy (CIS) and the advanced combination encoder strategy (ACE). The strategy used in this work is ACE. The ACE converts acoustic signals into electrical pulses through a series of transformations as Figure 2.3 shows.

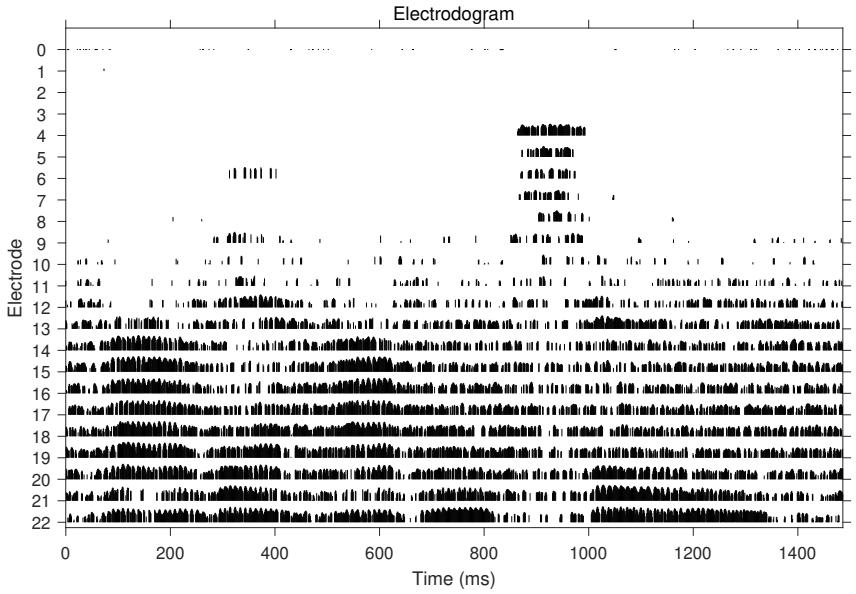

Figure 2.2: Electrodogram of a German phrase 'War der Abend schön?' ('Was the evening nice?') uttered by a male speaker with CCITT 5 dB noise and processed by the ACE strategy. Higher electrode numbers represent lower frequency bands.

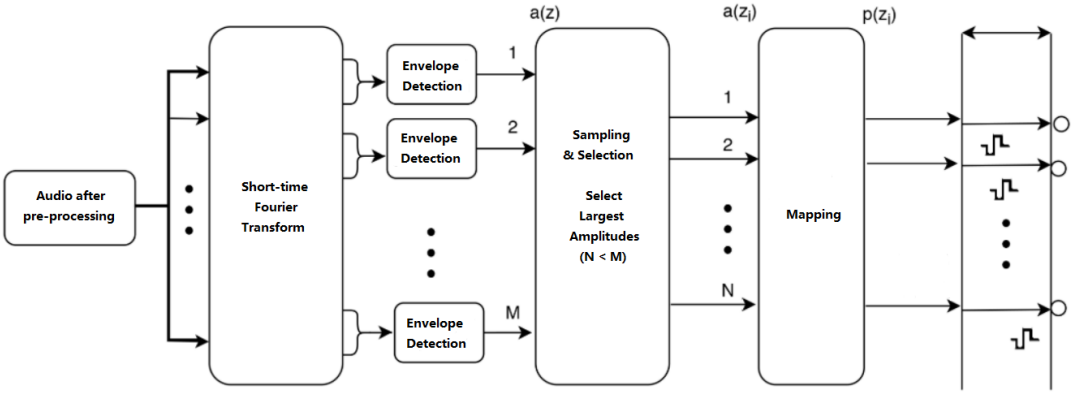

Figure 2.3: Block diagram of ACE [2]

The audio signal from the microphone is first pre-emphasized by a filter that amplifies the high-frequency components. Adaptive-gain control is used subsequently to limit distortion of loud sounds. Then the digitized signal is sent through the filter bank which is composed of the windowing, the short-time Fourier transform (STFT) and part of the envelope detector. The specific process is as follows:

- The signal sampled at $f_s = 16$ kHz is fragmented into frames of length $L = 128$

samples using a Hanning window.

- The windowed signal is shifted over a time interval. Shifting length N_s, namely the number of samples in the block shift, is defined through $N_s = f_s/CSR$ where the CSR stands for channel stimulation rate and equals 1000 pulses per second in this work.
- The center frequencies of each bin are linearly spaced at multiples of 125 Hz (0 Hz for 0th bin and 8 kHz for 64th bin). Since the input signals do not contain any complex number, only bins 0 to 64 are required due to Hermitian symmetry.

In the following envelope detection, the linearly-spaced $L/2 + 1$ coefficients obtained at the output of the FFT are grouped to form M bands. This grouping is performed considering the frequency resolution of the human auditory system as described by the critical band partition. The weighted sums of frequency bins are regrouped following the design of Table 2.1.

Table 2.1: Number of FFT bins, center frequencies and gains per filter hand for M=22 and a 128-FFT.

Band number z	1	2	3	4	5	6	7	8	9	10	11
Start bin n_{start_z}	3	4	5	6	7	8	9	10	11	13	15
Number of bins N_z	1	1	1	1	1	1	1	1	1	2	2
Center freqs (Hz)	250	375	500	625	750	875	1000	1125	1250	1437	1687
Gains g_z	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.68	0.68
Band number z	12	13	14	15	16	17	18	19	20	21	22
Start bin n_{start_z}	17	19	22	25	28	32	36	41	45	51	58
Number of bins N_z	2	2	3	3	4	4	5	5	6	7	8
Center freqs (Hz)	1937	2187	2500	2875	3312	3812	4375	5000	5687	6500	7437
Gains g_z	0.68	0.68	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65

The envelope magnitude $a(z_i)$ is calculated as Eq. 2.1. The gain $g(z)$ serves to normalize the envelope power to 1, when the signal is pure tone exactly at the center frequency with a magnitude equal to 1.

$$a(z) = \sqrt{\sum_{n=n_{start_z}}^{n_{end_z-1}} g(z) \cdot r^2(n)} \quad z = 1, \ldots, M \tag{2.1}$$

In the followed 'Sampling and Selection' block, N bands with the largest amplitude are selected for stimulation. N is the specified number of electrodes that release stimuli in a certain interval to complete one circulation of stimulation.

The channel stimulation rate represents the temporal resolution of the implant, while the total number of electrodes M represents the frequency resolution. By default $N = 8$ and

$M = 22$ are set.

Afterward, in the "Mapping" block, the selected N envelopes will go through the loudness growth function (LGF) that maps the acoustic envelope amplitude $a(z_i)$ to an electrical magnitude $p(z_i)$, and limits the stimuli magnitude within the most comfortable level (MCL, the upper limit of electrical stimulation judged to be most comfortable, or loud but comfortable) and the threshold level (THL, the minimal amount of electrical stimulation required for the auditory system to perceive sound) as the following Eq. 2.2:

$$p\left(z_i\right) = \begin{cases} \frac{\log\left(1+\rho\left(\frac{a(z_i)-s_{base}}{m_{sat}-s_{base}}\right)\right)}{\log(1+\rho)} & s \leq a\left(z_i\right) \leq m_{sat} \\ no\ output & a\left(z_i\right) < s_{base} \\ 1 & a\left(z_i\right) \geq m_{sat} \end{cases} \tag{2.2}$$

In this work, $s_{base} = 4/255$, $m_{sat} = 150/255$ are thresholds for the floor and ceiling of the energy level. And ρ is a constant parameter that determines the steepness of LGF and can be deducted according to the requirement. The value of ρ was then numerically solved and equals 416.21.

The magnitude of the excitation pattern denoted by $p(z_i)$ determines the dynamic range of output signals denoted by $Y(z_i)$ as:

$$Y\left(z_i\right) = THL(z) + \text{round}\left[\left(MCL(z) - THL(z) \cdot p\left(z_i\right)\right]\right] \quad i = 1, \ldots, N \tag{2.3}$$

In the end, the selected channels will activate the corresponding electrodes with the calculated stimulation levels $Y(z_i)$, using biphasic pulses in a sequence of high-to-low frequencies.

2.2.3 Limitation of Sound Coding Strategies

In general, CI users achieve excellent speech intelligibility in quiet conditions by using different sound coding strategies but struggle to understand speech in noisy environments. Reasons could be the channel interactions created by the broad spread of electrical fields in the cochlea and the low number of electrodes which causes a limited information transfer for the perception of fine spectro-temporal details in many types of sound. Many methods have been proposed to alleviate the limited speech understanding.

Several improvements to the sound coding strategies have been investigated. A strategy named psychoacoustic ACE (PACE) was proposed by [2], which incorporates a psychoacoustic model with ACE. Frequency bands that are most important to normal hearing people in terms of hearing perception are selected through a psychoacoustic-masking model, which improve the speech understanding to some extent.

Since common sound coding strategies can not encode the temporal fine structure details

which contribute to the perception of speech in noisy listening situations and CI users are more sensitive to noise, in addition to optimizing the sound coding strategy itself, it is also widespread to add speech pre-processing steps to achieve speech enhancement after receiving the audio signals. Related algorithms are discussed in detail in the next section.

2.3 Speech Enhancement

Speech enhancement aims to improve speech intelligibility by using various algorithms. The enhancement objective is improvement in intelligibility and overall perceptual quality of degraded speech signals using audio signal processing techniques. An important category in speech enhancement is source separation, which is the extraction of a set of source signals from a group of mixed signals without the aid of information about the source signals or the mixing process. It is most commonly applied in digital signal processing and involves the analysis of mixtures of signals; the objective is to recover the original component signals from a mixture signal. The typical example of a source separation problem is the cocktail party problem, where several people are talking simultaneously in a room, and a listener is trying to follow one of the discussions. The human brain can handle this auditory source separation problem, but it is a complex problem in digital signal processing.

The set of individual source signals, $s(t) = (s_1(t), \ldots, s_n(t))^T$ is mixed using a matrix, $A = [a_{ij}] \in \mathbb{R}^{m \times n}$, to produce a set of mixed signals, $x(t) = (x_1(t), \ldots, x_m(t))^T$, as follows. Usually, n is equal to m. If $m > n$, then the system of equations is overdetermined and thus can be unmixed using a conventional linear method. If $n > m$, the system is underdetermined and a non-linear method must be employed to recover the unmixed signals. The signals themselves can be multidimensional. The mixed signals are obtained using the mixing rule as shown in Eq. 2.4:

$$x(t) = A \cdot s(t) \tag{2.4}$$

The above equation is effectively inverted as follows. Blind source separation separates the set of mixed signals, $x(t)$, through the determination of an 'unmixing' matrix, $B = [B_{ij}] \in \mathbb{R}^{n \times m}$, to 'recover' an approximation of the original signals, $y(t) = (y_1(t), \ldots, y_n(t))^T$ [35]. The mixed signal are segregated using the formula as shown in Eq. 2.5:

$$y(t) = B \cdot x(t) \tag{2.5}$$

The overall structure of the source separation system is shown in Figure 2.4. The first process of source separation is to obtain the magnitude spectrum and phase information of mixed speech through STFT. The estimated targets using speech magnitude spectrum have been shown to suppress noise significantly and improve the speech intelligibility and its perceived quality [36]. Then, the magnitude spectrum information is used as

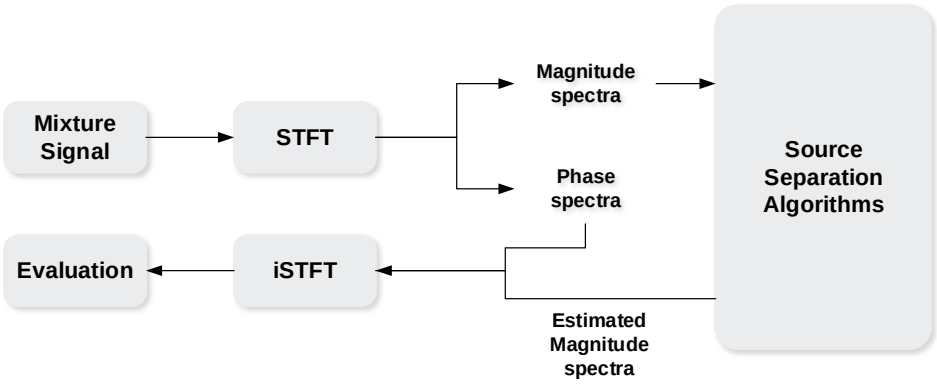

Figure 2.4: The modeling of source separation

the input of the source separation model. After the mixed signal has been processed by the separation algorithms, the target speech is obtained by the combination of the magnitude spectrum information estimated by the separation model and the previous phase information through the overlap-add method [37]. The source separation ability of the model is evaluated based on the comparison between the estimated speech and the clean speech. Not just speaker separation, this technique can also be used to separate speech and non-speech noise for noise reduction and speech enhancement, which is a potential approach to enhance the auditory perception of CI users.

2.3.1 Classic Speech Enhancement Methods

A blind approach to source separation would assume strictly no prior knowledge about either source properties or type of mixture [38]. However, the source separation problem cannot be solved in such conditions. Therefore, the most classical source separation approaches, which are called 'blind methods', are based on generic priors. Three types of such priors have, in particular, been used in the literature, thus yielding three main classes of source separation methods, i.e. statistical independence of the sources, which leads to Independent Component Analysis (ICA); positivity, which leads to Non-negative Matrix Factorization (NMF) and sparsity, which results in Sparse Component Analysis (SCA) and allows one to solve the source separation problem for more sources than observations (underdetermined case).

Based on the mechanisms of the human auditory system, Computational Auditory Scene Analysis (CASA) is also used for speech separation. Early models focused on extracting some acoustic properties such as pitch [39], onset [40], and amplitude modulation [41]. They are then based on proximity, similarity, or common destiny of these properties. Most of these models are biologically plausible and easy to interpret. However, the current understanding of auditory neuroscience is insufficient to develop systems as intelligent as

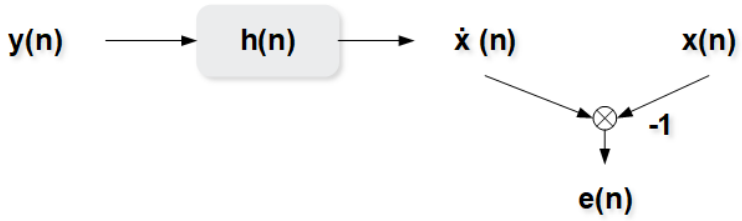

Figure 2.5: Diagram of Wiener filter

humans. These models are generally effective for simple stimuli but cannot accommodate natural sources in complex acoustic scenes.

Wiener Filter Different from the other aforementioned methods, Wiener filter a widely used technique for noise reduction that relies on a priori SNR estimation [42]. As Figure 2.5 shows, $y(n)$ is the noisy signal, $\hat{x}(n)$ and $x(n)$ are the enhanced and clean signal, $\hat{x}(n) = h(n) * y(n)$, the filter $h(n)$ is designed to minimize the mean squared error of $e(n)$. After frequency domain analysis, the error $E(\omega_k)$ can be defined as:

$$E(\omega_k) = X(\omega_k) - \hat{X}(\omega_k) = X(\omega_k) - H(\omega_k) * Y(\omega_k) \tag{2.6}$$

According to the minimum mean square error criterion:

$$\begin{aligned} \mathrm{E}[|E(\omega_k)|^2] = \mathrm{E}[|X(\omega_k)|^2] - H(\omega_k)\,\mathrm{E}[X^*(\omega_k)\,Y(\omega_k)] \\ -H^*(\omega_k)\,\mathrm{E}[X(\omega_k)\,Y^*(\omega_k)] + |H(\omega_k)|^2\,\mathrm{E}\left[|Y(\omega_k)|^2\right] \end{aligned} \tag{2.7}$$

Let $P_{yy}(\omega_k) = E\left[|Y(\omega_k)|^2\right]$, $P_{yx}(\omega_k) = E[Y(\omega_k)\,X(\omega_k)^*]$ and $\mathrm{J} = \mathrm{E}\left[|X(\omega_k)|^2\right] - H(\omega_k)\,P_{yx}(\omega_k) - H^*(\omega_k)\,P_{yx}^*(\omega_k) + |H(\omega_k)|^2\,P_{yy}(\omega_k)$, derivative with respect to J gives the expression for the Wiener filter:

$$H(\omega_k) = \frac{P_{yx}^*(\omega_k)}{P_{yy}(\omega_k)} \tag{2.8}$$

Combine Eq. 2.8 with $Y(\omega_k) = X(\omega_k) + N(\omega_k)$, where $N(\omega_k)$ is the Fourier transform of noise signal, X and N are independent and the expectation of N is 0, the expression of the filter can be written as:

$$H(\omega_k) = \frac{P_{xx}(\omega_k)}{P_{xx}(\omega_k) + P_{nn}(\omega_k)} \tag{2.9}$$

Define $\xi_k = \frac{P_{xx}(\omega_k)}{P_{nn}(\omega_k)}$ as priori SNR estimation, then:

$$H(\omega_k) = \frac{\xi_k}{\xi_k + 1}, 0 < H(\omega_k) < 1 \tag{2.10}$$

The physical meaning of Eq. 2.10 is: when the signal-to-noise ratio is large, the signal

will be allowed to pass, otherwise it will be suppressed. It is worth noting that Wiener filter works well for stationary noise, but for complex non-stationary noise, the boost is very limited.

2.3.2 Deep Learning Based Speech Enhancement

In contrast to classic methods, source separation can also be treated as a supervised learning problem and solves it by developing statistical models [43]. This approach develops supervised models based on task optimization on a specific training dataset. Recently proposed deep source separation systems have made significant progress due to the development of deep learning [44], which has started to perform natural source separation at the human level. The second method has attracted considerable attention due to its excellent performance, and it is also chosen to achieve speech enhancement in this project.

2.3.2.1 Background

Deep learning is part of machine learning methods based on artificial neural networks (NN) with representation learning. Learning can be supervised, semi-supervised or unsupervised [44]. Deep-learning architectures such as deep neural networks (DNN), recurrent neural networks (RNN), and convolutional neural networks (CNN) have been applied to fields including computer vision, speech recognition, speech enhancement, and natural language processing, where they have produced results comparable to human experts' performance [45].

NNs are based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron receives signals, then processes them and can signal neurons connected to it. The "signal" at a connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. Neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses that threshold. Typically, neurons are aggregated into layers. Different layers may perform various transformations on their inputs. Signals travel from the first layer (the input layer) to the last layer (the output layer), possibly after traversing the layers multiple times.

A deep neural network is a neural network with multiple layers between the input and output layers [46]. DNNs can model complex nonlinear relationships. The architecture generates compositional models in which objects are represented as hierarchical compositions of primitives [47]. Additional layers can combine features from lower layers, potentially using fewer units to model complex data than similarly performing shallow

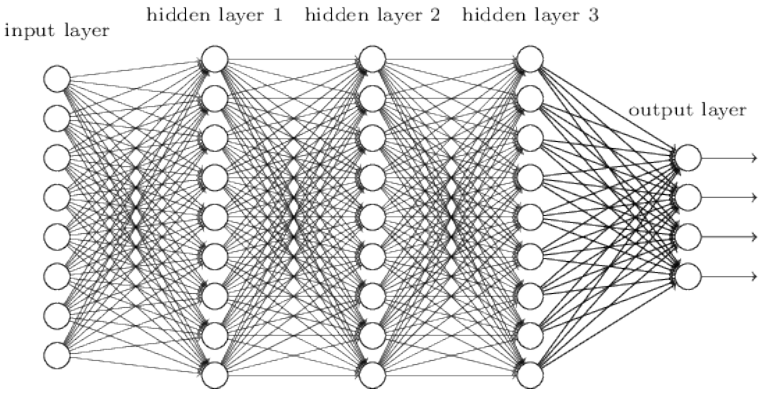

Figure 2.6: The architecture of deep neural network. [3]

networks [48]. DNNs are usually feedforward networks, where data flows from the input layer to the output layer and does not loop back. First, the DNN creates a map of virtual neurons and assigns weights to the connections between them. The weights are multiplied by the input and return an output. The algorithm will adjust the weights to make certain parameters more influential until it determines the correct mathematical manipulation to fully process the data. The architecture of DNN is shown in Figure 2.6. Then some important terminology of neural network are introduced.

Activation Function An activation function $g(x)$ is used to introduce non-linear properties into neural networks. To achieve it, at least one of the activation functions used must be non-linear. Another important property of the activation function is differentiability, since this is a prerequisite for the application of gradient descent. A variety of different activation functions have been explored, with the sigmoid function σ traditionally being a popular choice. Similarly, another widely used activation function is the tangens hyperbolicus (tanh). It shares its asymptotic properties with Sigmoid, but tanh is symmetric around the origin, which let your model converge faster by producing the least weight swings [49]. Recently, the Rectified Linear Unit (ReLU) and its variants (Leaky ReLU, Parametric ReLU, Gaussian Error Linear Unit, etc.) are perhaps the most popular in modern deep learning approaches because of the simple formula, low computational cost and good generalization ability [50]. Figure 2.7 visualizes a comparison of these three functions.

Backpropagation Backpropagation [51] is a neural network training algorithm. For supervised learning, target classes are essential for error calculation. The error is afterwards backpropagated to every node in previous layers. This error e is obtained as a gradient of the loss function L with respect to each layer's weights w_{kj} given input of the node x and activation function ϕ, like Eq. 2.11 shows:

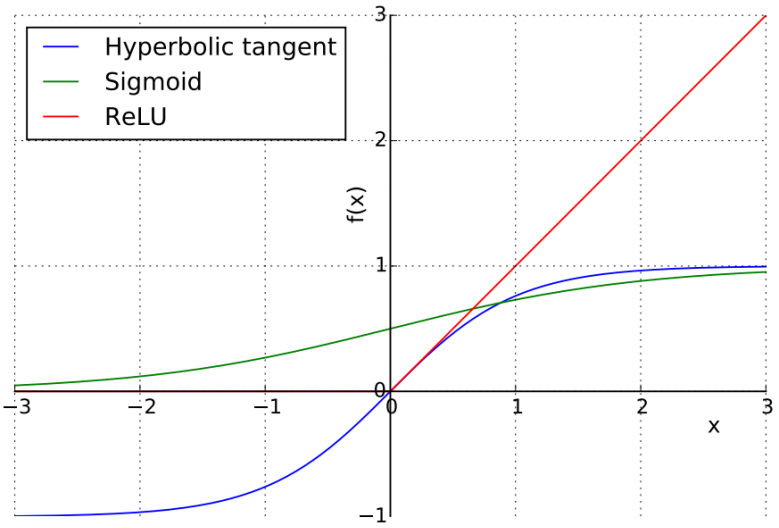

Figure 2.7: Visual comparison of the three most relevant activation functions of DNN: hyperbolic tangent, sigmoid and ReLU.

$$e = \frac{\partial L}{\partial w_{kj}} = \frac{\partial L}{\partial a_j} \frac{\partial a_j}{\partial w_{kj}} \tag{2.11}$$

where $a_j = \phi\left(\sum_{k=1}^{n} w_{kj} x_k\right)$. Gradient computation demands application of the chain rule in order to compute partial derivative of the loss function L with respect to particular weight w_{kj}. Using the error, weights are updated by an optimization algorithm such as gradient descent.

Loss Functions Loss functions are used to calculate the error between the estimated and true/desired outputs. So they serve as a measure of how good the predictions of the NN are. For regression tasks, the Mean Squared Error (MSE), Mean Absolute Error (MAE) and Huber Loss are commonly used. MAE and MSE are also called $L1$ and $L2$ losses, respectively, as shown in Eq. 2.12 and Eq. 2.13 :

$$\text{MAE} = \frac{1}{M} \sum_{i=1}^{M} |y_i - \hat{y}_i| \tag{2.12}$$

$$\text{MSE} = \frac{1}{M} \sum_{i=1}^{M} (y_i - \hat{y}_i)^2 \tag{2.13}$$

where $\hat{y}_i$ and y_i are the predicted and target signals, respectively. The disadvantage of MSE is that it imposes a large penalty on outliers and is not robust enough. MAE has the property of resisting outlier interference, but it is relatively more difficult to optimize. To overcome these tradeoffs, Huber Loss uses a hyper-parameter δ and combines the advantanges from MSE and MAE. MSE is used when the error is small, which makes

the loss function derivable and the gradient more stable. MAE is used when the error is large, which can reduce the influence of outliers and make training more robust to outliers.

$$\text{Huber Loss} = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}_{|y_i-\hat{y}i|\leq\delta}\frac{(y_i-\hat{y}_i)^2}{2} + \mathbb{I}_{|y_i-\hat{y}i|>\delta}\left(\delta\,|y_i-\hat{y}_i| - \frac{1}{2}\delta^2\right) \tag{2.14}$$

As for classification tasks, Cross-entropy Loss (CE) is widely used, which is essentially a log-likelihood function. For binary classification problem it is described as Eq. 2.15. When using sigmoid as the activation function, CE can perfectly solve the problem that the weight updates too slow.

$$\text{CE} = -\frac{1}{M}\sum_{i}^{M}\left[y_i \log \hat{y}_i + (1-y_i)\log\left(1-\hat{y}_i\right)\right] \tag{2.15}$$

Gradient Descent The goal of machine learning models is to minimize the loss function. And gradient descent is an efficient optimization algorithm that attempts to find a local or global minimum of a function, which is a first-order approximation algorithm that updates the weights of the model. The algorithm approaches a local minimum in the direction of the negative gradient of the loss function with respect to the weights. The size of the step is called the learning rate. It is a scalar in the range $(0, 1)$, controlling the magnitude of the network's parameters change. The whole training dataset has to be used for one update of the weights, the whole training set has to be used, which might be computationally expensive.

A more time efficient gradient descent based optimization method is the stochastic gradient descent or SGD [52] as shown in Eq. 2.16:

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta\nabla_{\boldsymbol{w}} L\left(\boldsymbol{w}^{(t)}\right) \tag{2.16}$$

where weights w are being updated by the negative of the gradient of the loss function with respect to the weights. This change is limited by the learning rate η. SGD needs only one observation to update model parameters w. As the name suggests, at each weight update a random observation is used. Furthermore, SGD does not tend to end up stuck in a local minima. A disadvantage of SGD is a slower convergence rate than the convergence rate of batch gradient descent. Due to its stochasticity, a wrong choice of starting observations may cause the algorithm to move further from global minima and make converge problematic. The Adam Optimizer [53] was developed to address these issues of SGD. It converges faster than most other optimizers in a wide variety of scenarios, while staying relatively simple in implementation and execution. The method computes individual adaptive learning rates for different parameters from estimates of the first and second moments of the gradients. It also combines the advantages from adaptive gradient algorithm (AdaGrad) and root mean square propagation (RMSProp) so that problems with sparse gradients and non-stationary situations can be better solved.

2.3.2.2 State of the Art Algorithms

For source separation, various deep networks have made significant progress, many of them are based on the three most commonly used architectures in the field of speech processing: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Transformer, respectively. CNN [54] is one typical feed-forward architecture, which is especially suited for visual pattern recognition due to the properties of translation invariance and weight sharing. RNN [55] processes the input as a sequence based on recurrent connections, useful for learning temporal dynamics. Transformer [56] is a recently proposed architecture, which is one sequence transduction model based on multi-headed self-attention. It is more parallelizable and requires less training time than CNN and RNN. In addition, the introduction of the residual path, skip-connection path, and dual-path makes networks more complex and flexible, incorporating more differences. These parts will be introduced in details for specific models later.

Most of the previous speech denoising techniques have used magnitude spectrograms as a front end. However, they discard potentially valuable information (phase) and utilize a generic feature extractor (magnitude spectrogram analysis) instead of learning specific feature representations for given data distribution. To overcome these disadvantages, Dario Rethage et al. [57] have proposed an end-to-end learning method for speech denoising based on Wavenet, which can learn multi-scale hierarchical representations from raw audio instead of TF representations. The model achieves better predictions through non-causality, i.e., using a few milliseconds of future samples, which can be afforded even in real-time scenarios. At the same time, since overlapping samples can be used to predict adjacent samples, this noise reduction model can reduce a lot of redundant calculations by predicting the target block, so parallelizing the prediction will save a lot of memory and time.

Another study employed generative adversarial networks (GANs) to perform temporal mapping [58], so-called end-to-end speech enhancement GAN (SEGAN). The generator is a fully convolutional network that performs denoising. The discriminator follows the same convolutional structure as G and it transmits information of generated waveform signals versus clean signals back to G. D can be viewed as providing a trainable loss function for G. The results show that the method is viable compared to current approaches.

Generative methods using variational autoencoders (VAE) or generative adversarial networks (GANs) have been increasingly used in recent years, but normalizing flow (NF) based systems are still rare. Therefore, Martin Strauss et al. [59] have proposed an end-to-end speech enhancement method based on normalizing flow, which allows for density estimation of clean speech samples given their noisy counterparts and signals enhancement via generative inference. The flow-based DNN architectures in speech synthesis can be modified to perform SE directly in the time domain without pre-defined

features or T-F transformations.

Conv-TasNet In particular, a fully-convolutional time-domain audio separation network (Conv-TasNet) is a recently proposed model [20] that made significant progress in the task of speech separation relative to previously state-of-the-art models. Conv-TasNet contains three parts: (1) a linear 1-D convolutional encoder that encapsulates the input mixture waveform into an adaptive 2-D representation; (2) a separator that estimates masking matrices for target sources, the masks are found using a temporal convolutional network (TCN) consisting of stacked 1-D dilated convolutional blocks, which allows the network to model the long-term dependencies of the speech signal while maintaining a small model size; and (3) a linear 1-D transposed convolutional decoder that converts the masked 2-D representations back to waveforms.

TCN The use of TCN has solved many shortcomings of the LSTM used in TasNet, such as the difficulty of convergence for long sequence LSTM, the large amount of LSTM parameters and the uncertainty of long dependencies. The TCN model is based on the CNN model and uses causal convolution to make it suitable for sequential models. Meanwhile, the use of dilated convolution and residual block gives the network the ability to obtain historical memory.

Dilated Causal Convolution A new CNN model - causal convolution is used to deal with sequence problems. The sequence problem can be transformed into predicting $y_1, y_2, \ldots, y_t$ from $x_1, x_2, \ldots, x_t$. Given filter $F = (f_1, f_2, \ldots, f_K)$ and sequence $X = (x_1, x_2, \ldots, x_T)$, the causal convolution in x_t is $(F * X)_{(x_t)} = \sum_{k=1}^{K} f_k x_{t-K+k}$.

In causal convolution, future information is not considered. That is, when predicting y_t, only the observed sequence $x_1, \ldots, x_t$ can be used, not $x_{t+1}, x_{t+2}, \ldots$. At the same time, the longer the historical information is traced back, the more hidden layers there are.

The standard CNN obtains a larger receptive field by adding a pooling layer, and there must be a problem of information loss after that. Dilated convolution injects holes into standard convolution to increase the receptive field. The dilated convolution has an additional hyperparameter - dilation rate, which refers to the number of intervals of the kernel. In standard CNN the dilatation rate is equal to 1. The advantage of holes is that the receptive field is increased without pooling loss information, so that each convolution output contains a larger range of information. With the same filter F and sequence X mentioned above, the dilated convolution at x_t with dilation rate d is defined as $(F *_d X)_{(x_t)} = \sum_{k=1}^{K} f_k x_{t-(K-k)d}$. The dilated causal convolutions for dilation rates 1, 2, 4 and 8 is depicted as shown in Figure 2.8.

Identity Mapping CNN can extract low/mid/high-level features. The more layers of the network, the richer the features that can be extracted at different levels. Moreover,

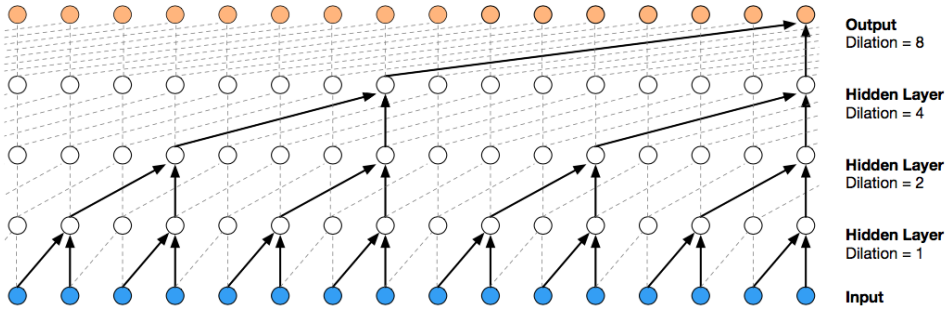

Figure 2.8: Visualization of a stack of dilated causal convolutional layers [4]

the features extracted by the deeper network are more abstract and have more semantic information.

Simply increasing the depth will result in vanishing gradients or exploding gradients. The solution is to initialize the weight parameters and use a regularization layer (Batch Normalization), which can train dozens of layers of network. After solving the gradient problem, another problem will arise: the problem of network degradation. As the number of network layers increases, the accuracy on the training set tends to saturate or even decline.

In order to solve the problem of network degradation, an identity map for the redundant layers of the network need to be generated to make a deep network equivalent to a shallow network. Usually it is difficult to let a layer of the network learn the identity mapping $H(x) = x$. However, if the network is designed as $H(x) = F(x) + x$, learning an identity mapping function can be transformed into learning a residual function $F(x) = H(x) - x$. As long as $F(x) = 0$, which is easier to achieve, the identity mapping is constituted.

Residual Block The structure of the residual module is shown in Figure 2.8. This module provides two selection methods, namely identity mapping x, which is called shortcut connection, and residual mapping $F(x)$. If the network has reached the optimal state, continuing to deepen the network will push the residual mapping to 0, leaving only the identity mapping. As a result, theoretically the network has always been in an optimal state, and the performance of the network does not degrade as the depth increases.

The separator in Conv-TasNet is a fully convolutional network, which can be seen as the core of the whole network. At first, input is processed by a global layer normalization (LayerNorm) and a 1×1 convolutional layer, indicating a 1-D convolution with a kernel size of 1. It determines the channels for the following modules and is regarded as the bottleneck layer. The main body of Conv-TasNet contains R repeated modules , and each module is composed of N stacked Dilated Conv-Blocks with exponentially increased

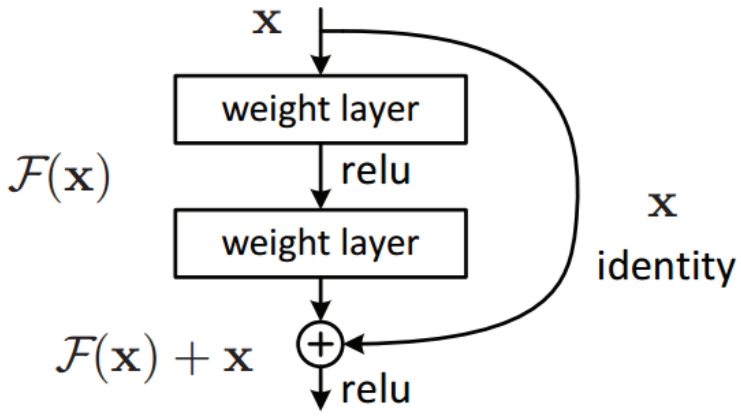

Figure 2.9: Structure of single residual block [5]

dilation factors. Each Dilated Conv-Block is a cascade of 1×1 convolution, PReLU, normalization, depth-wise dilated convolution, PReLU, normalization layer. Two 1×1 convolutional layers serve as the residual path and the skip-connection path, where the output of the residual path is the input of the next block, and the skip-connection paths of all blocks are summed up and used as the input to a 1×1 convolutional layer and a nonlinear activation layer to estimate two masks. The use of skip-connection facilitates training deep models by allowing them to explicitly incorporate features extracted at several hierarchical level into its final prediction. To further decrease the number of parameters, depthwise separable convolution is used to replace standard convolution in each convolutional block. The structure of the separator is shown in Figure 2.10.

The objective of training Conv-TasNet is maximizing the scale-invariant source-to-distortion ratio (SI-SDR), which has commonly been used as the evaluation metric for source separation. SI-SDR has solved the potential problem of using SNR as loss function, that is, the model may improve the SNR value by simply increasing the amplitude of the output, rather than improving the intelligibility of the speech.

Conv-TasNet significantly outperforms previous time-frequency masking methods in source separation tasks. Additionally, Conv-TasNet surpasses several ideal time-frequency magnitude masks in speech separation as evaluated by both objective distortion measures and subjective quality assessment by human listeners.

Source Separation for CIs The different hearing characteristics of CI users with normal hearing people and the practical feasibility of cochlear implant applications put forward higher requirements for speech enhancement and source separation technologies. Better enhancement performance should be achieved with fewer parameters and smaller structure while maintaining low latency. In CI research, an end-to-end audio latency of less

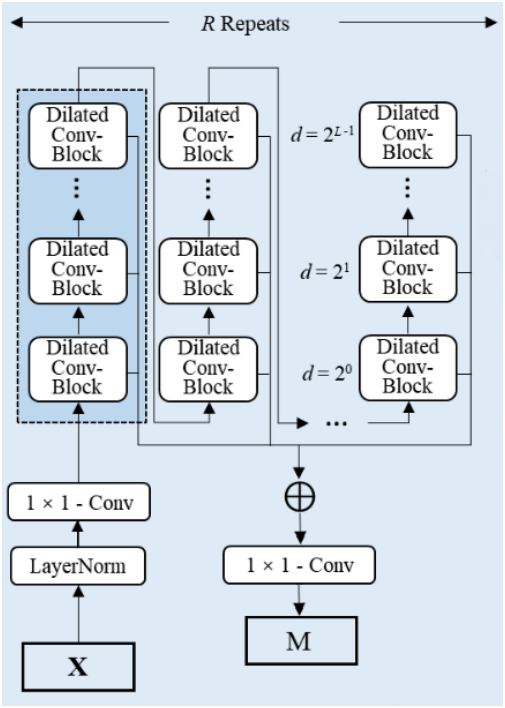

Figure 2.10: Architecture of separator in Conv-TasNet [6]

than 10 ms is required [60].Therefore, many source separation models were modified for CI.

Bolner et al. [61] have proposed a neural network based speech enhancement (NNSE) applied to cochlear implant coding strategies. The model decomposes the noisy input signal into time-frequency units, extracts a set of auditory-inspired features and feeds them to a feed-forward NN with two hidden layers to estimate which CI channels contain more important perceptual information. This estimation is then integrated in ACE, used accordingly to reserve a subset of channels for electrical stimulation, as in a traditional 'n-of-m' coding strategy. The results confirmed a significant improvement in speech intelligibility. But how well it works under more challenging conditions is questionable.

To improve speech enhancement methods for CI users, Mamun et al. [62] proposed to perform speech enhancement in a cochlear filter-bank feature space, which is a feature-set specifically designed for CI users based on CI auditory stimuli . The speech features are extracted from noisy signal based on CI auditory features. These 22-dimensional features before 'n-of-m' are then fed to different versions of convolutional neural network (CNN) to perform noise reduction. Afterwards, they are directly used to stimulate electrodes. A better performance is achieved in lower SNR conditions compared to other baseline models.

More recently, Zheng et al. [63] have proposed a noise-robust NN-based signal processing strategy for CI. Unlike other research, the neural network is built and trained to simulate

Table 2.2: Hyper-parameters used for training the models

Description	Value
Number of filters in the autoencoder	64
Length of the filters	32
Number of channels in the bottleneck blocks	64
Number of channels in the skip-connections	32
Number of channels in the convolutional blocks	128
Kernel size in the convolutional blocks	128
Number of convolutional blocks in each repeat	3
Number of repeats	2

the advanced combination encoder (ACE), instead of working as an extra module. This NN-based ACE (NNACE) is composed of a long-short term memory (LSTM) network followed by two DNNs, and trained with designated loss functions to extract noise-robust envelopes and temporal fine structure (TFS) as much as possible. The perceptually weighted error of the log-magnitude spectra, the estimation error of envelopes and the error between the vocoded and the clean speech waveform are all taken into account with different weights. According to the Objective and subjective evaluations, clinical CI recipients can get great benefits even with a very light NN.

2.3.3 Speech Enhancement for CIs: DeepACE

The combination of high accuracy, short latency, and small model size makes Conv-TasNet a suitable choice for real-time low-latency speech processing applications such as wearable hearing devices. Inspired by the success of Conv-TasNet, Gajecki et al. [21] have proposed a deep learning speech denoising sound coding strategy - DeepACE that estimates electrodograms directly from raw audio data, performing end-to-end (audio-to-electrodogram) CI processing. Conv-TasNet separates different speakers, while DeepACE achieves noise reduction by separating the speech and non-speech noise and completely bypasses ACE.

DeepACE and Conv-TasNet are similar in implementation. The separator, as the core of the model, basically retains the original structure. The hyper-parameter settings used for training can be seen in Table 2.2. The biggest difference is the output dimension of 1-D transposed convolutional layer in the decoder, because the output and the target signals of DeepACE are electrodograms instead of audio files. The structure of DeepACE is shown in Figure 2.11.

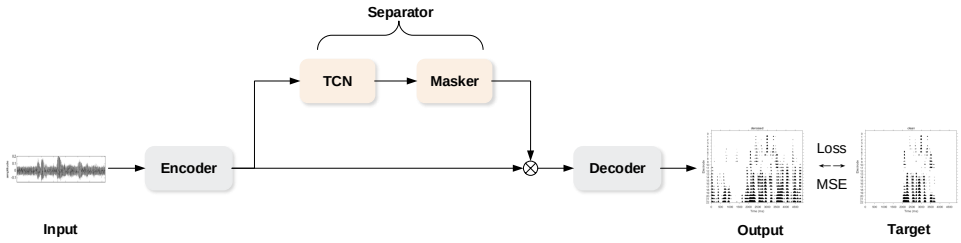

Figure 2.11: Structure of DeepACE

For the same reason, SI-SDR - the previous loss function of Conv-TasNet, is also no longer suitable for DeepACE. The optimizer is used to minimize the MSE between the predicted and target electrodograms. MSE is defined before as Eq. 2.13. In DeepACE, $\hat{y}_i$ and y_i represent the estimated and clean electrodogtrams, and M represents the number of bands, which is usually set to 22 for CI users.

The performance of DeepACE was evaluated objectively in vocoded audio domain. The estimated electrodograms were resynthesized using a sine vocoder. Afterwards, the SNR improvement and STOI improvement between processed vocoded audio and the unprocessed audio were calculated. Furthermore, a subjective listening test for CI patients was held to obtain an intuitive perception of the improved speech intelligibility of CI users by this algorithm.

In the context of speech enhancement, DeepACE proves great improvement compared to the unprocessed signals. But its objective performance is slightly worse than the front-end Conv-TasNet approach, although this end-to-end technique reduces the algorithmic latency with respect to the front-end methods by bypassing the ACE sound coding strategy. The reason might be potentially related to a sub-optimal loss function used to minimize the error between the input and target electrodograms. As is known to all, SI-SDR fits well for speech enhancement tasks in time domain, but MSE is just a common used loss function, which may not bring extra benefits to the training of DeepACE.

Besides, limited by hardware conditions, DeepACE did not conduct research on larger model sizes. Although the size of Conv-TasNet was kept the same to make the evaluation fairer, we still don't know whether increasing the size will further improve the performance of the model, and whether there is a 'saturated' upper limit.

Finally, DeepACE also has room for improvement in the selection and setting of datasets. In order to make the experimental scene closer to the real situation and get more solid results about generalization ability. According to all the concerns mentioned above, the corresponding improvement measures will be discussed in detail in the next chapter.

3 Methods and Materials

3.1 Audio Material

3.1.1 Speech Datasets

Totally there are three different audio datasets that were used for training and testing the DeepACE model.

Since most deep learning-based speech enhancement tasks use English datasets as training datasets, it would be interesting to try out a German speech dataset and observe the training performance. Thus, LibriVoxDeEn Corpus [64] is chosen for training and testing materials. This dataset is a corpus of German audio based on numerous German audiobooks. The corpus consists of over 100 hours of audio material. The speech data are low in disfluencies because of the audiobook setup. In this work, audio files from thirty male and thirty female speakers are used in the training period, 20% of them are chosen for validation. The total length reached eighteen hours. At the same time, the speech of three male and three female speakers was used for testing.

As with other CI-related studies, the HSM sentence test (Hochmair, Schulz and Moser) is used to assess speech intelligibility in CI users [65], which consists of 30 lists of 20 (German) everyday sentences (106 words per lists). All sentences are spoken once by a male and a female speaker, yielding 600 speech files for each gender.

For English sentences the TIMIT corpus [66] of read speech was used. It is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 people speaking the eight major dialects of American English, each reading ten phonetically-rich sentences. Corpus design is a joint effort of the Massachusetts Institute of Technology (MIT), SRI International (SRI), and Texas Instruments (TI). The test and training subsets of TIMIT corpus that are balanced for phonetic and dialect coverage are specified, which allows the results to be easily compared with other studies. In this work, files from 112 male and 56 female speakers in the test set were selected.

3.1.2 Noise Datasets

To bring this research much closer to reality and demonstrate the generalization ability of the DeepACE model, there are three different noise types used in this work.

The Consultative Committee for International Telephony and Telegraphy (CCITT) noise [67] was used, which is a fluctuating speech noise simulating average spectral and temporal features of fluent speech. So was the multiple-speaker-modulated speech-weighted noise source (ICRA7) [68], which has been developed for the International Collegium of Rehabilitative Audiology. The purpose was to establish collection of noise signals to be used as background noise in clinical tests of hearing aids. The signals are composed with well defined spectral and temporal characteristics similar to those typically found in real life speech signals and babble noise.

Furthermore, the Diverse Environments Multichannel Acoustic Noise Database (DEMAND, in short DM) [69] - a database of 16-channel environmental noise recordings was also implemented in this work. The noise is divided into 6 categories, 4 of which are "inside" and 2 of which are open air. The inside environments are classified as Domestic, Office, Public, and Transportation; the open air environments are Street and Nature. There are 3 environment recordings in each category. The shape and the spectrograms of these three noise types can be seen in Figure 3.1 and 3.2.

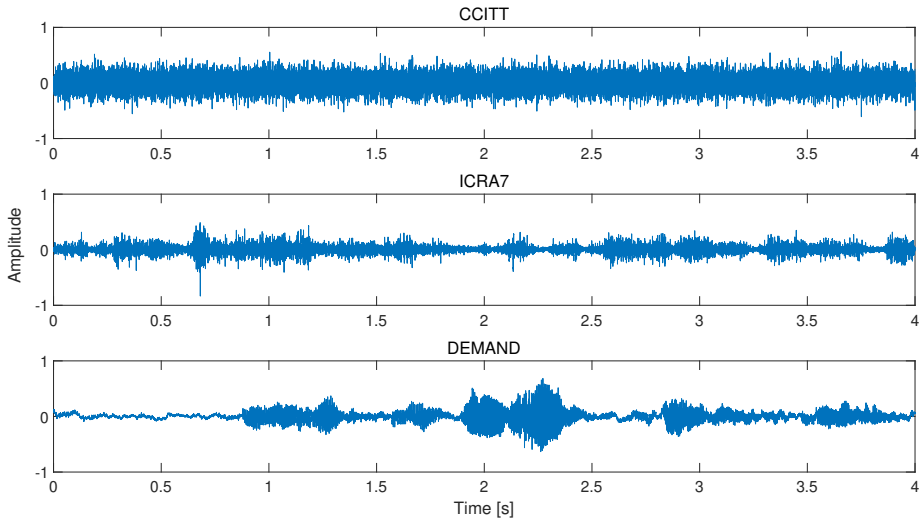

Figure 3.1: Shape of three noise types

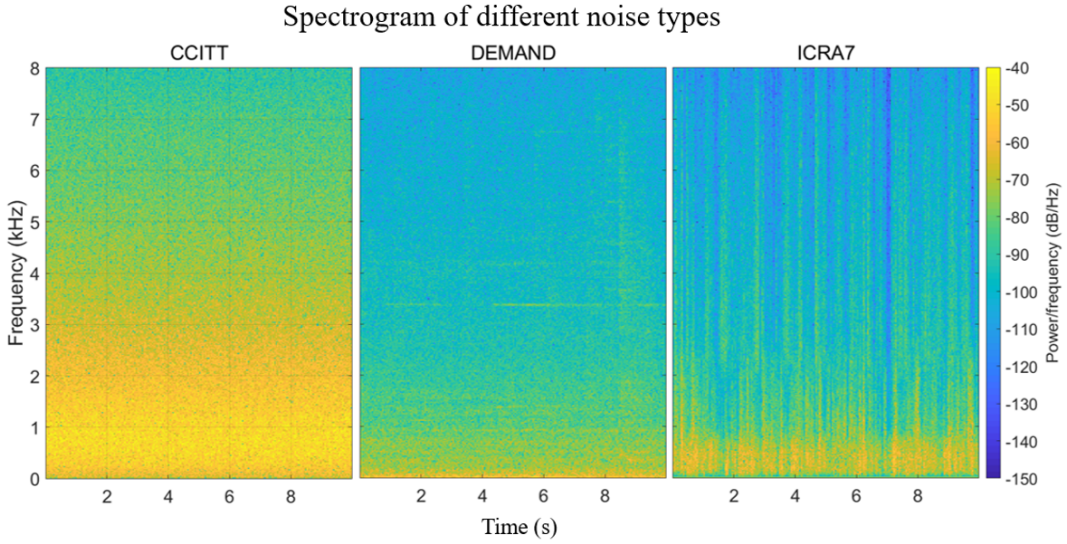

Figure 3.2: Spectrogram of CCITT, DM and ICRA7 noise.

3.1.3 Preprocessing

In this work, only LibriVoxDEEN corpus as speech and DEMAND as noise were chosen for the training progress of the model. The whole three speech and three noise datasets were used in test. It is worth noting that all speakers used in training were not present in test again. At the same time, for each category of DEMAND noise there were two environment recordings used for training and the rest one for test. This ensures that all speech and noise in test are unseen, which can improve the confidence of the results and allows for a better study of the model's generalization ability.

Before these data were put into the model, all samples were set to mono files and resampled to 16 kHz in Matlab. All speech was mixed with noise in a proper way, the SNR values ranged from -5 to 15 dB, in 5 dB steps. The corresponding electrodograms were obtained using the ACE strategy with 1000 pulses per second (pps) per electrode. It is worth noting that some clean speech is also used as test data to investigate whether our model can achieve performance close to ACE in quiet conditions without introducing artifacts.

Some papers [57] report that deep learning-based speech enhancement models had difficulties producing silence, this result was also demonstrated in our earlier trials. Therefore, a noise-only data augmentation was implemented in this work. Some samples with random lengths only contain background noise at the beginning and end of each training sentence. For test this length was fixed as two seconds. For the convenience of subsequent use, we first declared all the datasets utilized, including their specific noise settings, number of speakers, and total duration, as Table 3.1 shows.

Table 3.1: Datasets used with different types and sizes

Usage	Speech	Number of Speakers	Duration	Noise
Training	LibriVox	30 male, 30 female	18 h.	DM
Testset 1	LibriVox	3 male, 3 female	4.5 h.	DM/CCITT/ICRA7
Testset 2	TIMIT	112 male, 56 female	4.2 h.	DM/CCITT/ICRA7
Testset 3	HSM	1 male, 1 female	1.5 h.	DM/CCITT/ICRA7
Listening testset	HSM	1 male	6.5 h.	CCITT/ICRA7

Table 3.2: Modified hyper-parameters of the network

Symbol	Description
N	Number of filters in autoencoder
R	Number of repeats
L	Number of convolutional blocks in each repeat
K	Kernel size in convolutional blocks
P	Length of the filter (im samples)

3.2 Optimization of DeepACE

In order to make all optimization results comparable, experiment configurations were determined and unified for all experiments. The networks were trained for 100 epochs on 4-second long segments. The initial learning rate was set to $1e^{-3}$. The learning rate was halved if the accuracy of validation set was not improved in 5 consecutive epochs. Adam was used as the optimizer. A 50% stride size is used in the convolutional autoencoder (i.e. 50% overlap between consecutive frames).

3.2.1 Optimization in Hyper-Parameters

Limited by hardware conditions, DeepACE [21] did not conduct research on larger model sizes. Thus, it is worth experimenting with more combinations of hyper-parameters to see if they can improve the performance of the model. The optimization in hyper-parameters mainly focuses on two parts, the separator and the encoder/decoder. All the hyper-parameters that were modified in the experiments are shown in Table 3.2.

In this work, the size of TCN was continuously increased by adjusting the hyper-parameters mentioned above. To make the comparison fairer, The receptive field of the models was all set within a similar and reasonable range. The receptive field is defined as the maximum number of steps back in time from current sample. It was computed as Eq. 3.1:

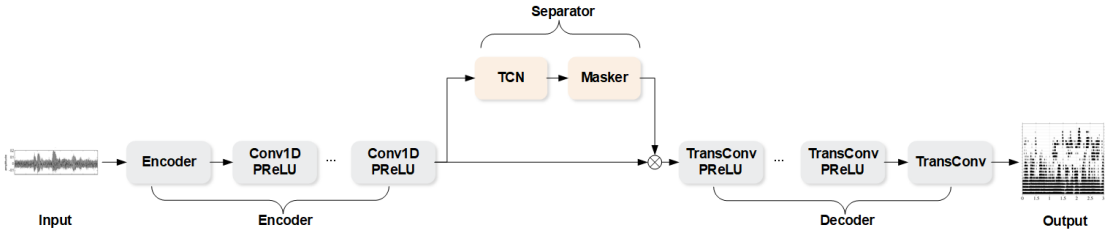

Figure 3.3: Structure of deep encoder/decoder

$$R_{field} = \frac{\left[K + (K-1) * \left(2 + 2^2 + \ldots + 2^{L-1}\right) * R\right] * P}{f_s} \tag{3.1}$$

where R_{field} is the receptive field of the model in seconds, P is the length of the filter and f_s is the sample rate of the input audio files.

Besides, the dimension of the feature extracted by the encoder is determined by N, which might affect the performance of separation. For Conv-TasNet, increasing the dimension of the feature improved the SI-SDR slightly [20]. Thus, the different choices of N were also tested here.

Most of the investigations about Conv-TasNet focused on the separator, leaving its encoder/decoder as a (shallow) linear operator. Recently, an enhancement to the encoder/decoder was achieved based on a (deep) non-linear variant of them [70]. A nonlinear deep encoder with I layers was utilized. The first layer is equivalent to the original Conv-TasNet encoder. The second part consists of a stack of $I-1$ 1D convolutional layers, with each layer having N kernels of size 3 and a PReLU:

$$\mathbf{E}_i = \text{PReLU}\left(\mathbf{U}_i * \mathbf{E}_{i-1}\right) \tag{3.2}$$

where $\mathbf{U}_i \in \mathbb{R}^{N \times N \times 3}$ are the kernels, $\mathbf{E}_i \in \mathbb{R}^{N \times K}$ is the layer output and $i = 2, 3, \ldots, I$. The input audio files were transfered hierarchically into a nonlinear latent space. The deep decoder directly mirrors the encoder architecture by implementing 1-D transposed convolutional layers with PReLU activations.

Better performance was reported, however, the downside is that more different combinations have not been studied for this variant. A similar deep encoder/decoder model was also implemented in this work, but the final 1-D transposed convolutional layer in decoder was applied to generate the electrodograms, not the time domain separated signals. Figure 3.3 depicts the diagram of the proposed deep encoder/decoder.

3.2.2 Optimization in Loss Functions

Loss functions are equations that give a curve of loss generated by the predictions of your model. The aim is to minimize the loss function to enhance the model's accuracy for better predictions. Since this work is about regression task, many related loss functions were tested.

First of all, mean squared error (MSE) was implemented as in DeepACE. The MSE is described in Eq. 2.13. The results were also used as baseline to compare with other loss functions.

The scale-invariant source-to-distortion ratio (SI-SDR), as used in Conv-TasNet, was used in this work to investigate if it fits the source separation task for electrodogram signals instead of audio signals. However, it is worth noting that when generating the electrodogram signals, a logarithmic operation was performed in loudness growth functions (LGF) in Eq. 2.2. Meanwhile, in original SI-SDR, there is also a logarithmic operation. Logarithmic scales are useful for quantifying the relative change of a value as opposed to its absolute difference. Moreover, because the logarithmic function $log(x)$ grows very slowly for large x, logarithmic scales are used to compress large-scale scientific data. But calculating the logarithm twice makes no mathematical sense. Thus, for electrodogram signals a linear version of SI-SDR was introduced:

$$\text{SI-SNR}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = 10 \cdot \log_{10}\left(\frac{\|\gamma \boldsymbol{x}\|^2}{\|\gamma \boldsymbol{x} - \hat{\boldsymbol{x}}\|^2}\right), \gamma = \frac{\hat{\boldsymbol{x}}^\top \boldsymbol{x}}{\|\boldsymbol{x}\|^2} \tag{3.3}$$

$$\text{SI-SDR}_{\text{lin}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\|\gamma \boldsymbol{x}\|^2}{\|\gamma \boldsymbol{x} - \hat{\boldsymbol{x}}\|^2}, \gamma = \frac{\hat{\boldsymbol{x}}^\top \boldsymbol{x}}{\|\boldsymbol{x}\|^2} \tag{3.4}$$

where $\hat{\boldsymbol{x}} \in \mathbb{R}^{1\times T}$ and $\boldsymbol{x} \in \mathbb{R}^{1\times T}$ are the estimated and original clean sources, respectively.

SNR loss was proved to be more suitable than SI-SDR loss for the perceptual quality metrics in Conv-TasNet [71]. So SNR is also considered as a loss function. Similarly, for electrodogram signals a linear version of SNR was used:

$$\text{SNR}_{\text{lin}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\boldsymbol{x}^2}{\boldsymbol{x} - \hat{\boldsymbol{x}}^2} \tag{3.5}$$

Besides modifying the loss function from logarithmic to linear, there is another way to keep the original form of the loss function. The target signal can be modified from electrodograms to evelope signals that do not include logarithmic operations, i.e. using the signals before LGF in the sound coding strategy. Then the loss functions SI-SDR and SNR can be described as:

$$\text{SI-SDR}(\boldsymbol{x_{lin}}, \hat{\boldsymbol{x_{lin}}}) = 10 \cdot \log_{10_{lin}}\left(\frac{\|\gamma \boldsymbol{x_{lin}}\|^2}{\|\gamma \boldsymbol{x_{lin}} - \hat{\boldsymbol{x_{lin}}}\|^2}\right), \gamma = \frac{\hat{\boldsymbol{x_{lin}}}^\top \boldsymbol{x_{lin}}}{\|\boldsymbol{x_{lin}}\|^2} \tag{3.6}$$

$$\mathrm{SNR}(\boldsymbol{x_{lin}}, \hat{\boldsymbol{x_{lin}}}) = 10 \cdot \log_{10} \left(\frac{\|\boldsymbol{x_{lin}}\|^2}{\|\boldsymbol{x_{lin}} - \hat{\boldsymbol{x_{lin}}}\|^2} \right) \tag{3.7}$$

where $\hat{\boldsymbol{x_{lin}}}$ and $\boldsymbol{x_{lin}}$ are the estimated and original clean envelope signals, respectively.

As a variant of the MSE, mean squared logarithmic error (MSLE) was implemented in envelope domain. Advantage is that MSLE doesn't penalize the large error significantly more than the small ones, especially when the target variable has large values. The MSLE is shown in Eq. 3.8:

$$\mathrm{MSLE}(x_{lin}, \hat{x_{lin}}) = \frac{1}{M} \sum_{i=0}^{M} \left(\log\left(x_{lin_i} + 1\right) - \log\left(\hat{x}_{lin_i} + 1\right) \right)^2 \tag{3.8}$$

The value is incremented by 1 to account for the $log(0)$ undefined problem. MSLE helps to separate the penalties for overestimates and underestimates whereas there is no categorization in MSE. MSLE penalizes underestimates more than overestimates.

All the loss functions mentioned above are relatively common in the field of deep learning. MSE is suitable for many different kinds of tasks, but from a semantic point of view, the flaws of MSE are obvious, as the semantic information represented by low frequency bands is often more important than it by the high frequency bands. MSE does not restrict the band selection, to some extent, a smaller MSE does not necessarily mean higher speech intelligibility. For the unique auditory characteristics of CI users, a new MSE-based loss function is proposed in this work.

From Qazi et al. we could know that, compared to distortion of the signal envelopes, CI users are more sensitive to distortion of the band-selection [26], i.e. CI users can tolerate large distortions in speech segments, if the band-selection is not corrupted.

We hope that our DeepACE model can ensure that the data will be reconstructed on the same bands in original clean electrodograms as much as possible. So in this work, we modify the MSE and add a punishment for incorrectly selected frequency bands, which is a weighted MSE (wMSE). The formula of wMSE is shown as Eq. 3.9:

$$\mathrm{wMSE} = \frac{1}{M} \left(\sum_{i \in Sel} (y_i - \hat{y}_i)^2 + w * \sum_{i \in \overline{Sel}} (y_i - \hat{y}_i)^2 \right) \tag{3.9}$$

where Sel and $\overline{Sel}$ are sets about the frequency band number i, all the frequency band that are selected in the clean electrodograms are classified as Sel set, the rest are classified as $\overline{Sel}$ set, i.e. the complementary set of Sel, and w refers to the weight of unselected part. In this way, in the process of gradient descent, the value generated in the frequency band that should not be selected will be significantly suppressed. The appropriate weight value needs to be determined through experiments. All the loss functions that were

Table 3.3: Loss functions used for optimizing the model

Domain	Selected Loss Functions
Electrodogram	MSE, SI-SDR, SI-SDR$_{lin}$, SNR$_{lin}$, weighted MSE
Envelope	MSLE, SI-SDR, SNR

tested in this work are shown in Table 3.3.

3.2.3 Optimization in Architecture

Based on the same idea [26] like the weighted MSE that suppresses the output of incorrect bands, a masked DeepACE was also developed.

Instead of modifying the MSE, two loss functions were introduced in this model. Compared to the original architecture of DeepACE, the output of TCN no longer generates the separated feature masks, but a ideal binary mask (IBM), which identifies speech dominated and noise dominated units. A binary mask was computed and applied to the noisy input features to get the noise-suppressed mask, which only contains values of 0 or 1. The target of the mask came from the corresponding clean electrodograms.

Another difference was that the feature extracted by the encode is not directly multiplied with the mask to obtain a denoised feature. Several Conv-1D layers were added to extract the feature further and convert the matrix into the same dimension as the mask has. According to the original idea of ACE, the part was recognized as Deep Envelop Detection. The architecture of the masked DeepACE is shown in Figure 3.4.

The core network is constructed and trained to obtain denoised electrodograms. To do so, the loss functions were computed on outputs at two stages of the training. The error between estimated and the clean target electrodograms was measured by MSE, just like the original DeepACE. The reason why not using weighted MSE here is that the suppression was already done through the second loss Binary Cross-entropy (BCE), which counts the binary classification error of the ideal binary mask. The formula of BCE is shown as Eq. 3.10:

$$\mathrm{BCE} = -\frac{1}{M}\sum_{i=1}^{M} y_i \cdot \log\left(p\left(y_i\right)\right) + \left(1 - y_i\right) \cdot \log\left(1 - p\left(y_i\right)\right) \tag{3.10}$$

where y is the label and $p\left(y\right)$ is the predicted probability for all N values.

Finally, the loss function for optimization is given by Eq. 3.11:

$$\mathrm{Loss} = w_{MSE} * \mathrm{MSE} + w_{BCE} * \mathrm{BCE} \tag{3.11}$$

where the weights w_{MSE} and w_{BCE} were determined through the experiment. The first

loss force the output to be as close to the clean electrodogram as possible, therefore, denoising could be achieved. The second loss forces the mask output to be as close to the clean ideal binary mask as possible, assuring that only the desired bands will be selected to stimulate.

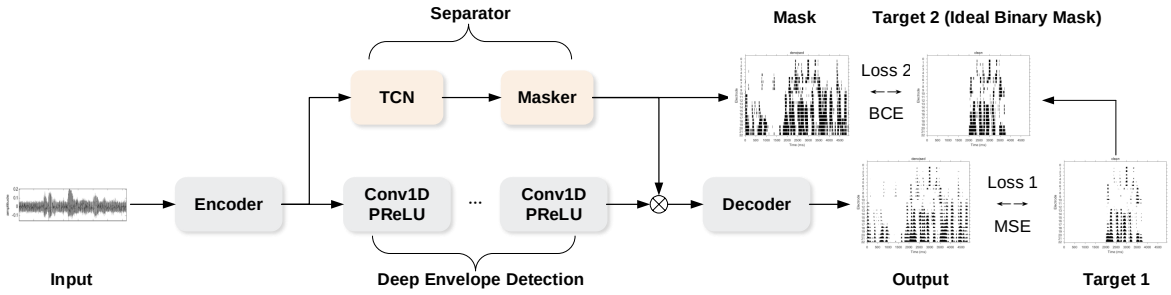

Figure 3.4: Architecture of masked DeepACE

3.3 Baseline Models

The performance of optimized DeepACE was evaluated in comparison with some baseline systems, i.e. the original ACE, Wiener filter as a front-end to ACE (Wiener+ACE), Conv-TasNet as a front-end to ACE (TasNet+ACE in short) and the DeepACE with original architecture. As an optimization of the original DeepACE model, our model undoubtedly needs to surpass the original DeepACE and remove the noise while preserving speech information as much as possible. It was ensured that the two models always have the same combination of hyper-parameters during comparison.

ACE Original ACE was also used for evaluation. It generated all the unprocessed signals. The improvement of other models can then be easily measured. Figure 3.5 shows all the different signal processing systems that used in this work. The band selection part was removed out of the DeepACE model, which provides more flexibility in actual use.

Wiener+ACE Furthermore, a classic front-end signal processing method based on Wiener filtering was implemented in this work. This algorithm is used in many commercially available single channel noise reduction systems included in CIs [72]. Thus, this is a suitable baseline when developing new speech enhancement methods for CI users.

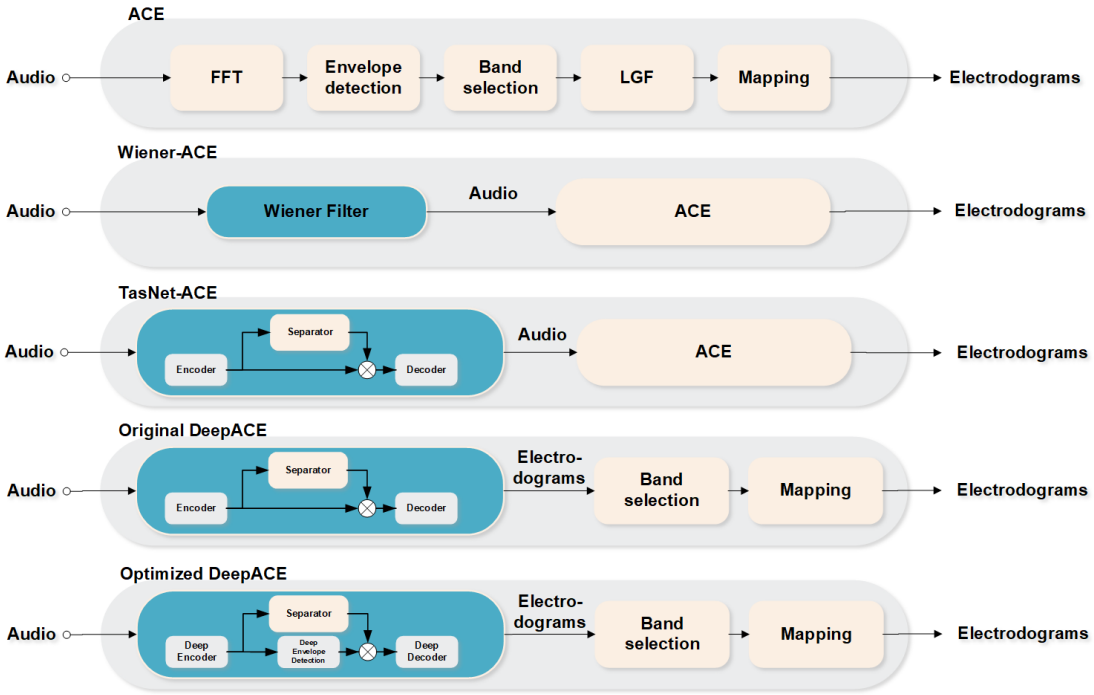

Figure 3.5: Block Diagrams of Different Signal Processing Systems

TasNet+ACE Conv-TasNet is a successful end-to-end speech enhancement method commonly used in single-speaker tasks. It was selected because it offered a causal implementation and had few parameters, which showed potential in real-time wearable devices. In previous research it has been proven that the TasNet+ACE even outperforms the original DeepACE a bit objectively [21]. Therefore, it is very critical whether the optimized model can surpass TasNet-ACE in both objective evaluation and listening tests. DeepACE and Conv-TasNet both introduce an algorithmic latency of 2 ms because of the filter length in encoder, but TasNet-ACE gets another 2 ms latency from ACE as a front-end technique.

3.4 Experimental Design

3.4.1 Objective Metrics

Since the DeepACE model aims to generate denoised electrodogram which can be directly fed to the implant to stimulate the auditory nerve, this work mainly focuses on the objective evaluation in electrodogram domain.

SNRi In electrodogram domain, the SNR improvement with respect to the unprocessed signal (noisy electrodogram) was taken into account, which compare the SNR between denoised and clean signals to the SNR between noisy and clean signals. This can be computed as Eq. 4.2 shows:

$$\text{SNRi} = 20 \log \frac{LGF_{noisy} - LGF_{clean}}{LGF_{denoised} - LGF_{clean}} \tag{3.12}$$

where LGF means the electrodogram signals. The SNR improvement was calculated for every sentence.

LCC In electrodogram domain, the correlations between denoised and clean signals were also evaluated using the linear correlation coefficient (LCC). The channel-wise calculation gave the possibility to evaluate the performance in different frequency bins. The LCC was computed as Eq. 3.13:

$$\text{LCC} = \rho = \frac{\text{cov}\left(LGF_{clean}, LGF_{\text{denoised}}\right)}{\sigma_{LGF_{clean}} \sigma_{LGF_{denoised}}} \tag{3.13}$$

where LGF means the electrodogram signals, $\text{cov}\left(X, Y\right)$ is the covariance of X and Y, and σ is the covariance of the values of denoised and clean electrodogram divided by the product of standard deviations of the values of them.

STOI On the other hand, in order to estimate the improvement in speech intelligibility, metrics in time domain were also implemented. In the field of audio signal processing, short-time objective intelligibility (STOI) score [73] is more closely related to the human auditory perception and has a high correlation with speech intelligibility and has been used in CI studies on speech enhancement and sound coding strategies. The range of STOI score is between 0 to 1. The higher the score is, the better the intelligibility is. In this work, the perceived intelligibility was estimated using both a reference speech in clean and a vocoded denoised speech resynthesized from the electrodograms using a sine vocoder.

A downside of STOI may be that the score is too abstract to intuitively understand. In contrast, the value of Word Recognition Score (WRS) can be easily understood. WRS is sometimes also referred to as Speech Discrimination Score (SDS) and represents the number of words correctly repeated. The number of correct words is scored out of the number of presented words to give the WRS. A relationship between STOI and WRS can be established through a mapping function [74]. Word Recognition Score $f(d)$ is calculated through a logistic function as Eq. 4.5 shows:

$$f(d) = \frac{100}{1 + \exp(ad + b)} \tag{3.14}$$

where d is the STOI score, a and b are free parameters, which are fitted to the data with

a nonlinear least-squares procedure. And as the logistic function is also monotonic, the monotonicity between STOI and WRS will not be influenced.

In general, different mapping functions are needed for different speech corpus. But as we just want to show the speech intelligibility more intuitively than STOI score by the estimates of WRS, in this work, we did not fit the data from our speech dataset but directly used parameter pairs for the English IEEE database, where $a = -17.4906$ and $b = 9.6921$ [74]. In this work, WRS was used as a supplement to STOI score to evaluate the speech understanding.

3.4.2 Test Setup

In the process of optimizing the model, different structures were tested only with speech dataset HSM and noise dataset CCITT and DEMAND, and evaluated by SNR improvement in electrodogram domain to make a clear comparison and save time.

In the final objective evaluation, the optimized, best performed DeepACE model and all the baseline models were tested with the whole test dataset in both electrodogram domain and vocoded audio domain.

3.5 Listening Test

It is worth noting that STOI as an metric for evaluating speech intelligibility, may not accurately reflect the hearing quality of CI users. Thus, unlike other papers that only implemented objective evaluation and subjective evaluation based on vocoded speech towards normal hearing (NH) people, we applied a listening test for real CI users.

3.5.1 Introduction of Participants

Ten postlingually deafened CI users participated in this listening test . All participants were native German speakers and have been implanted for several years. They were invited to come to German Hearing Center of the Hannover Medical School (MHH) for a 3-hours test and the travel cost were all covered. The experiment was granted with ethical approval by the MHH ethics commission. A synopsis of the pertinent patient-related data is shown in Table 3.4.

3.5.2 Test Setup

Since all the participants were German speaking, the HSM dataset mixed with CCITT and ICRA7 noise in SNR range between 0 and 10 dB was applied for the listening

Table 3.4: Listener demographics and etiology

ID	Age	Gender	Best Side	CCITT SNR [dB]	ICRA7 SNR [dB]
BI01	70	M	R	0	5
BI02	40	M	R	0	5
BI03	74	M	R	0	0
BI04	70	M	R	5	5
BI05	36	F	R	0	0
BI06	62	M	R	5	5
BI07	86	M	R	10	10
BI08	57	F	R	5	5

Table 3.5: Listening test setup

Elements	Setting
Model	ACE, Wiener+ACE, TasNet+ACE, Optimized DeepACE
Testset	HSM male, HSM female
Noise Type	CCITT, ICRA7
SNR [dB]	0, 5, 10, quiet

experiments in CIs.

There were four models selected for this test. The optimized DeepACE model with best performance was chosen. Meanwhile, the original ACE progress, the two front-end baseline models TasNet and the wiener filter were also selected. The detailed test setup is shown in Table 3.5

The test materials first went through these four models and the related electrodogram were generated. Except for the files passed through ACE, the rest of them were all denoised. The channel stimulation rate (CSR) used in this work to train and evaluate the models was 1000 pps. Afterwards, the stimuli were delivered to the participants' CI via direct stimulation through the RF GeneratorXS interface (Cochlear Ltd.) with MATLAB via the Nucleus Implant Communicator V.3 (Cochlear Ltd.).

During the test, the participants were asked to sit in a quiet room and remove their own implants, the stimulation went straight to their auditory nerve, so they were not disturbed by outside sounds. The result was obtained by conducting the HSM sentence test. Subjects were required to repeat the sentence they've heard out loud as accurately as possible. All the correctly repeated word were recorded, and for each test list there was a final score that represented the percentage of understood words. Each listening condition was tested twice with different sentence lists, and the list which was used will not appear again for the same subject. All the conditions were blinded to the subjects.

3.5.3 Statistics

To further investigate if the three selected algorithms achieved significant improvement in speech intelligibility compared to the unprocessed ACE and which algorithm had the best performance, some statistical analyses are implemented. For the first purpose, the percentage of correctly understood words were directly used. For the second one, a benefit score compared to ACE was first calculated as Eq. 3.15, where 'unprocessed' refers to ACE and 'processed' refers to the other three algorithms.

$$\text{Benefit}[\%] = \text{Understood Word}_{processed} - \text{Understood Word}_{unprocessed} \tag{3.15}$$

Significance tests can be divided into parametric and non-parametric analysis of variance (ANOVA) [75]. The parametric ANOVA requires that the samples follow a normal distribution, and these normal populations have the same variance. When the data do not satisfy the assumptions of normality and homogeneity of variances, parametric ANOVA may give incorrect answers, and rank-based non-parametric ANOVA should be used. First of all, a statistical test for determining the normality of the data distribution was implemented. In this method, the Skewness and Kurtosis were calculated to analyse the characteristics of the data distribution:

$$\text{Skew}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{k_3}{\sigma^3} = \frac{k_3}{k_2^{3/2}} \tag{3.16}$$

$$\text{Kurt}[X] = \text{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} = \frac{\text{E}\left[(X-\mu)^4\right]}{\left(\text{E}\left[(X-\mu)^2\right]\right)^{2'}} \tag{3.17}$$

where μ is the mean, σ is the standard deviation, E is the expectation operator, μ_t is the t-th central moment, and k_t are the t-th cumulants. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. For a unimodal distribution, negative skew commonly indicates that the tail is on the left side of the distribution, and positive skew indicates that the tail is on the right. kurtosis is a measure of the 'tailedness' of the probability distribution of a real-valued random variable. A high kurtosis means that the increased variance is caused by low frequency extreme differences greater or less than the mean.

Besides, the Shapiro–Wilk test [76] was also used to test the normality since the sample size of the data was less than 50. The null-hypothesis of this test is that the population is normally distributed. Thus, if the p-value is less than the chosen significance level $\alpha = 0.05$, then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed. As for the homogenity of variance, the Levene's test [77] with a null-hypothesis that the population variances are equal was selected.

If the data are shown to be normally distributed and have homogeneity of variance, one way ANOVA will be selected for significance test. The null-hypothesis of this test is that there is no significant difference between groups. Afterwards, t-test [78] can be also

used to calculate pairwise comparisons between group levels with corrections for multiple testing.

As long as one of the two conditions is not met, non-parametric ANOVA is required for further analysis. In this case, Kruskal–Wallis test [79] will be chosen. The idea of the Kruskal-Wallis test is to mix n groups of samples into a dataset, then rank the data from small to large to give each one a rank in the mixed dataset. If the rank of each group is not much different from the total average rank of the mixed data, it means that each group comes from the same population. The null-hypothesis of this test is that there is no significant difference between groups. If this hypothesis is rejected, the Wilcoxon signed rank test [80] will be implemented for the pairwise multiple comparisons.

4 Results

4.1 Results of Optimization

4.1.1 Hyper-Parameter Tuning

First of all, the performance of DeepACE with different hyper-parameters was investigated. To unsure a fair comparison, all the models here were causal and trained with the same training dataset described in Table 3.1 and tested only with speech dataset HSM combined with noise dataset CCITT and DEMAND in 0 dB condition. The loss function chosen for all was MSE. The model with best SNR improvement in the electrodogram domain was selected for further optimization.

Table 4.1 shows the performance of the models with different parameters, from which we can conclude the following statements:

(i) Effect of hyper-parameters in TCN: Within a certain range, increasing the number of repeats and the convolutional blocks in each repeat increased the performance, that is, deeper networks lead to better performance, perhaps attributable to a larger model capacity. But further increasing the model size leads to a slight drop in performance. The introduction of residual blocks and skip connections can alleviate the degradation (accuracy saturation) problem to a certain extent, but cannot completely eliminate it. TCN with 3 repeats and 8 convolutional blocks in each repeat might be the best architecture for this task.

(ii) Effect of the receptive field size: Increasing the size of receptive field leads to better performance. It proves that in speech related tasks the modeling of temporal dependencies is very important. But it is worth noting that when the receptive field is too large (greater than 4s), the model performance begins to decline, which may also be the combined effect of the model size and the receptive field.

(iii) Effect of encoder/decoder: Increasing the number of N, i.e. increasing the dimension of the feature extracted by encoder did not improve the performance. This doesn't match the observation in [20], where the performance was improved by increasing the overcompleteness of the basis signals. A similar result was reported by [70], but without further explanation. This might also denote the importance of the architecture itself, because the objective metrics did not improve by simply increasing the capacity of the model.

Table 4.1: The effect of different configurations in DeepACE

N	R	L	K	P	Receptive Field (s)	Model Size	SNRi (dB) CCITT	SNRi (dB) DEMAND
64	2	2	128	32	1.272	168k	7.5582	6.4379
64	2	3	128	32	3.304	162k	8.2213	7.0377
512	2	3	128	32	3.304	800k	7.9752	6.9085
1024	2	3	128	32	3.304	2000k	7.7231	6.6623
64	3	4	16	32	1.292	293k	8.1326	6.7530
64	3	8	3	8	0.763	549k	8.8801	7.5781
64	3	8	3	16	1.527	551k	8.9523	7.7071
64	**3**	**8**	**3**	**32**	**3.054**	**552k**	**9.0199**	**7.9753**
64	3	8	3	48	4.581	553k	8.7501	7.5306
64	3	9	3	32	6.126	617k	8.8211	7.3512
64	4	4	16	32	1.712	468k	8.3471	7.0133
64	4	6	8	32	3.488	549k	8.7375	8.0791
64	4	8	3	32	4.070	724k	8.7932	7.7524

Besides the structure of TCN, a deep encoder/decoder structure was also studied. The number of layer I means that the encoder has a single 1D convolutional layer and $I-1$ 1D convolutional layers combined with a PReLU. As for Decoder, the convolutional layers were replaced by transpose convolutional layers. The original DeepACE model was selected as the baseline. So the TCN with 2 repeats and 3 convolutional blocks in each repeat was fixed in this part. The loss function chosen was MSE. The results are shown in Table 4.2.

Table 4.2: The effect of different depth in deep encoder/decoder

I	Model Size	SNRi (dB) CCITT	SNRi (dB) DEMAND
1	162k	8.2213	7.0377
3	196k	8.7920	7.7461
4	**234k**	**8.9732**	**8.1015**
5	275k	8.9336	8.0801

Encoder/decoder with 4 layers achieved the best performance. Surprisingly, deepening the encoder/decoder can reach similar performance with much fewer parameters than deepening the TCN. However, combining the best performed deep encoder/decoder and TCN structure did not lead to further improvement. On the contrary, the SNR improvement for CCITT noise only reached 8.3089 dB, much less than both of two separately.

Based on all the results above, an architecture with shallow encoder/decoder and a TCN

Table 4.3: The effect of different loss functions

Loss Function	w	Signal Domain	SNRi (dB) CCITT	SNRi (dB) DEMAND
MSE	NA	Electrodogram	8.2213	7.0377
SI-SDR	NA	Electrodogram	7.3390	6.1294
SI-SDR$_{lin}$	NA	Electrodogram	6.3467	5.4476
SNR$_{lin}$	NA	Electrodogram	2.6538	0.7835
MLSE	NA	Envelope	6.5792	5.2501
SI-SDR	NA	Envelope	1.5488	0.9257
SNR	NA	Envelope	5.6739	4.4483
weighted MSE	2	Electrodogram	8.3817	7.0428
weighted MSE	5	Electrodogram	8.6632	7.2997
weighted MSE	**10**	**Electrodogram**	**9.0720**	**7.9343**
weighted MSE	20	Electrodogram	7.9594	6.8933
weighted MSE	100	Electrodogram	7.3409	6.0510

with 3 repeats and 8 convolutional blocks was selected for further optimization in other aspects.

4.1.2 Loss Function Tuning

In this part, different loss functions mentioned in 4.1.2 were implemented in the DeepACE model. For those non-logarithmic loss functions, the estimated signals were in electrodogram domain. For logarithmic loss functions, the estimated signals were in envelope domain and a extra loudness growth function was calculated to generate electrodogram. As for the weighted MSE that suppress the incorrect band selection, different weights were also tested. To speed up the progress, a small TCN structure used in the original DeepACE model was selected. All the training and evaluation setups remained the same as in 4.1.1. Table 4.3 shows the performance of the models with different loss functions.

In addition to weighted MSE, the original MSE outperforms the other loss functions in both electrodogram and envelope domain. Meanwhile, SI-SDR also achieves good performance. The reason why it is not as good as MSE might be that SI-SDR is optimal for audio domain signals, not for electrodogram domain signals since the logarithm is calculated twice. Loss functions for the envelope domain generally achieve poor results, which shows the error might be further amplified through the extra loudness growth function.

Comparing the original MSE ($w = 1$) and weighted MSE, as the weight increases, the performance of the model has a trend of first increasing and then decreasing. Although

in essence the weighted MSE is aimed at the subjective auditory characteristics of CI users to suppress incorrect frequency band selection, objectively it also allows the model to reconstruct signals accurately in frequency bands that are more likely to contain clean signals. However, with the further increase of the weight, the requirement for the accuracy of the frequency band selection greatly exceeds the accuracy of the numerical reconstruction of a single signal, which leads to a certain degree of distortion. Based on the results, the weighted MSE with a $w = 10$ was selected for further investigation.

4.1.3 Investigated DeepACE Architecture

The masked DeepACE model was optimized from two perspectives: the structure of Deep Envelope Detection and the weight combination of two loss functions of MSE and BCE. In order to be able to compare with previous results, the original DeepACE model was chosen as baseline model. For masked DeepACE model, a small TCN structure used in the original DeepACE model was also selected. The other training and evaluation setups remained the same as in 4.1.1. Table 4.4 shows the performance of the models with different depth of Deep Envelope Detection module (DED). I_{DED} refers to the number of 1D convolutional layers combined with a PReLU for each layer. The weight of MSE and BCE was fixed here and set to 10 and 1,respectively.

Table 4.4: The effect of different depth in Deep Envelope Detection in masked DeepACE

I_{DED}	SNRi (dB) CCITT	SNRi (dB) DEMAND
1	8.2579	7.0101
2	8.3904	7.1206
3	**8.4210**	**7.1358**
4	8.3556	7.0528

According to the results, DED part itself can also bring a tiny improvement. A reason might be that although it was mainly implemented for dimensional transformation, it also further extracted the features of the input audio based on the encoder output. Finally, a DED with a stack of 3 convolutional layers and PReLUs was selected.

The weight combination of MSE and BCE actually decides the importance of the regression part and the classification part. From the results of the previous subsection, it can be seen that increasing the weight of weighted MSE finally led to a decrease of performance because the numerical distortion was amplified. The value of MSE and BCE were generally in the order of magnitude of 10^{-3} and 10^{-2} respectively, based on the above analysis, the weights of the two loss functions should act on the total loss function in a balanced manner without introducing a large bias. Table 4.5 shows the performance of the models with different combination of weights. The weight combination 15/1 was

Table 4.5: The effect of different combination of weights in masked DeepACE

w_{MSE}	w_{BCE}	SNRi (dB) CCITT	SNRi (dB) DEMAND
5	1	7.6738	6.8901
10	1	8.5311	7.1206
15	**1**	**8.6974**	**7.4512**
20	1	8.0231	6.9527

selected for further evaluation.

4.1.4 Final Models

The results in this section demonstrate that the idea about suppressing erroneous band selection is effective. At the same time, the two different implementations of this idea, i.e. DeepACE with weighted MSE and masked DeepACE, have also been proved to surpass the performance of the baseline structure and bring different degrees of improvement.

Since the results in this section were only evaluated from a small testset and only the SNR improvement in electrodogram domain was used as objective metric, it's not entirely clear which implementation has an absolute advantage. Therefor, both of them were selected for further objective evaluation and listening test. Table 4.6 shows the used hyper-parameters and configurations of these two implemented models and the baseline model TasNet. In subsequent chapters, DeepACE with weighted MSE will be called DeepACE-wMSE and masked DeepACE will be called DeepACE-Mask for clarity.

Table 4.6: Configurations for selected models

Configuration	DeepACE-wMSE	DeepACE-Mask	TasNet
N	64	64	64
R	3	3	3
L	8	8	8
K	3	3	3
P	32	32	32
Model Size	552k	553k	526k
Loss Function	wMSE	MSE+BCE	SI-SDR
Weight w	10	15/1	1
Signal Domain	Electrodogram	Electrodogram	Audio

4.2 Objective Results

Figure 4.1 shows the electrodograms for the celan and noisy speech with 0 dB CCITT noise produced by ACE and the electrodograms of denoised speech produced by different algorithms. The objective instrumental evaluation concentrated mainly in the electrodogram domain, as well as the vocoded audio domain where the signals were synthesized by electrodograms. For the first second signal there was only noise and this part was ignored during the whole objective evaluation.

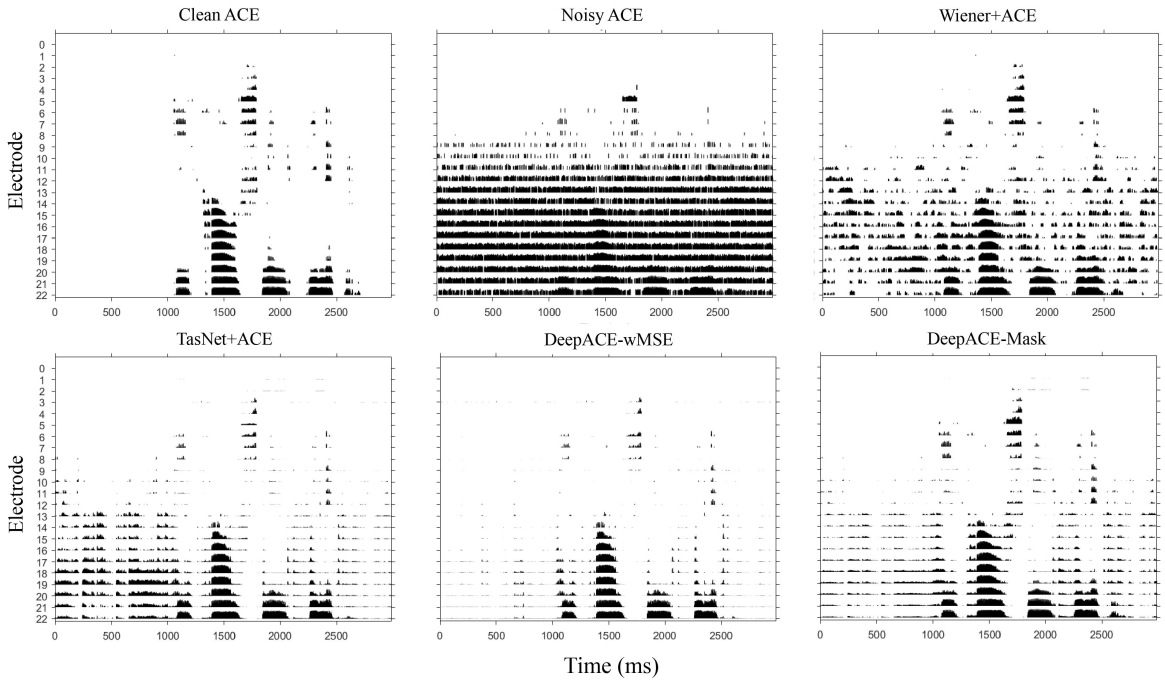

Figure 4.1: Electrodograms for clean speech and speech with noise produced by different algorithms, electrode numbers increase as the corresponding frequency range increases.

4.2.1 SNR Improvement

Figure 4.2 illustrates the SNR improvement of the investigated algorithms in the electrodogram domain with respect to unprocessed ACE, for three different speech and noise dataset and four SNR inputs. From the perspective of different speech datasets, the three deep learning based algorithms achieved the best performance for LibriVox speech dataset. There was nearly 2 dB improvement compared to other datasets. Reason might be that LibriVox was used in training process. Although the speakers and sentences used in testing were totally different and unseen, the models might still have learned some hidden features from this dataset. The result also shows the performance will not be affected across languages. The English TIMIT dataset almost got the same performance

as the German HSM dataset. As for the wiener filter, the performance was very similar in the three speech datasets.

From the perspective of noise, the differences between different datasets were more obvious. For stationary noise CCITT and babble noise ICRA7, all three deep learning based algorithms can bring steady improvement. And the fluctuation of the performance for each sentence was not very large. But for the environmental record noise dataset DEMAND (DM), the fluctuation was obvious and the average performance decreased. A reason might be that it is difficult for the network to learn features faced with this kind of realistic and challenging noise. Meanwhile, in the condition with less noise, the algorithms had a very small probability to bring negative improvement. But in general, they still had a relatively stable improvement compared to unprocessed ACE. Wiener filter only achieved good performance for CCITT noise. For more complex noise, the improvement was very limited. In general, as the SNR value increased, the improvement gradually decreased.

It can be seen our DeepACE-wMSE and DeepACE-Mask models outperformed TasNet+ACE and Wiener+ACE in almost every condition, especially in low SNR (-5 and 0 dB) environment. However, it is hard to tell which of the two algorithms is better from this box plot. More analysis were then carried out. From Figure 4.1 it can be seen that our DeepACE-wMSE model achieved the best denoising performance. Almost all the noise was removed even though it is 100% unseen to the network. However, some speech information were also removed during the denoising process, which may disrupt speech structure to some extent and affect speech understanding. The impact were accessed through subsequent STOI analysis in the vocoded speech domain and the listening tests. Our another DeepACE-Mask model, on the contrary, can not remove all the noise, but retained more speech information. It is also worth noting that the wiener filter achieved poor denoising performance compared to the other deep learning based algorithms, but the speech information was not corrupted. Since our final goal is to improve the speech understanding for CI users, it is difficult to draw conclusions from the SNR improvement in electrodogram domain alone.

4.2.2 Linear Correlation

In order to further evaluate how much the electrodograms after denoising with different algorithms are related to the clean electrodograms, linear correlation coefficients were calculated. In this subsection, electrode numbers increase as the corresponding frequency range increases. To provide corresponding objective instrumental evaluation results for the subsequent listening test, the LCC of three deep learning based algorithms and Wiener+ACE were calculated in the listening test conditions, i.e. the HSM speech dataset with 0,5,10 dB CCITT and ICRA7 noise.

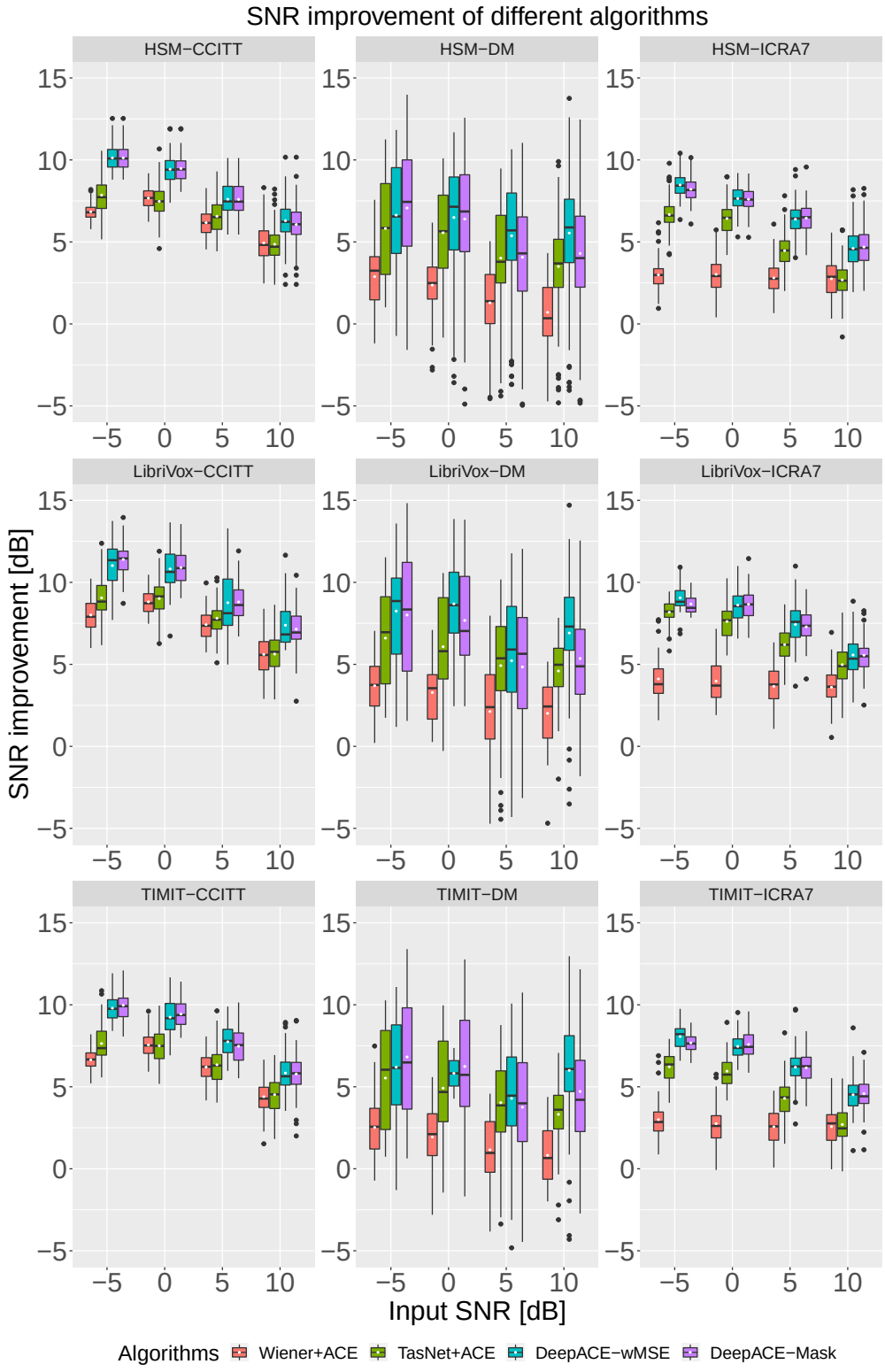

Figure 4.2: Box plots of SNR improvement in dB for the tested algorithms in different speech and noise conditions. The black horizontal bars within each of the boxes represent the median for each condition, the diamond-shaped marks indicate the mean improvement, the top and bottom extremes of the boxes indicate the 75% and 25% quartiles, respectively, and whiskers indicate the variability outside the upper and lower quartiles.

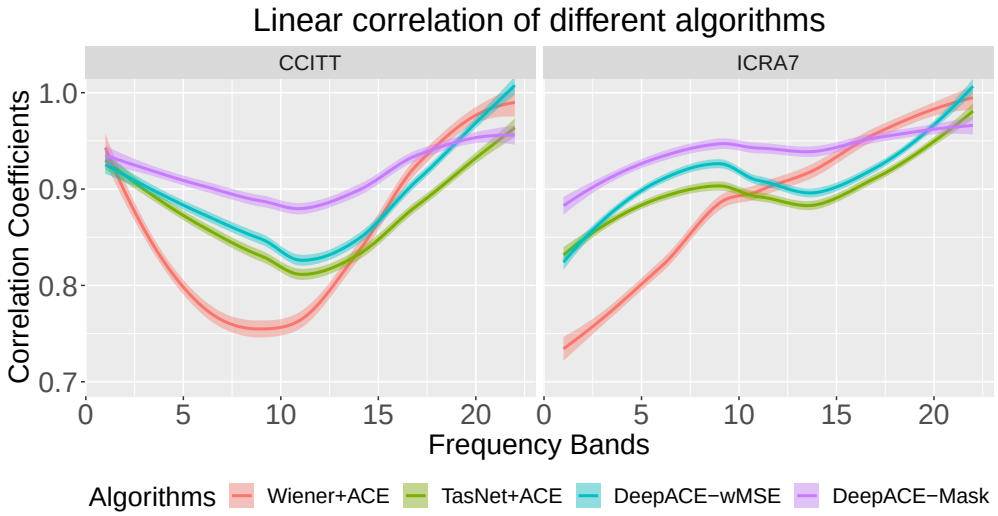

Figure 4.3: Plot of linear correlation coefficients between processed and clean electrodograms across bands of different algorithms in HSM speech with 0,5,10 dB CCITT and ICRA7 noise. Shaded area represents standard deviations.

Figure 4.3 shows that three deep learning based algorithms all got higher correlation than Wiener+ACE in low frequency parts. This is consistent with the analysis of SNR improvement in the previous subsection that the Wiener+ACE got poor performance for denoising. On the contrary, Wiener+ACE raised rapidly with increasing frequency and outperforms the other three algorithms in the highest frequency bands. The possible reason was that the other algorithms removed some high-frequency speech information as noise components, and at the same time introduced a small amount of artifacts. However, since the noise is mainly concentrated in the middle and low frequency bands, the Wiener+ACE did not need to remove too much noise in the high frequency bands and the speech structure was then kept.

Despite the numerical differences, for the same noise type, the fitting curves of different algorithms had similar trends. But when making comparisons across noise types, the trends were very different. Figure 3.2 shows the spectrograms of CCITT DM and ICRA7 noise. For CCITT noise, the noise was mainly concentrated in the 500-3000Hz frequency range, so the models were less correlated from 3rd to 12th bands. ICRA7 noise mainly occurred in the frequency range below 1000Hz, so the models were less correlated in the first few bands. Three deep learning based algorithms had similar performance across different noise types, which also proved the generalization ability. Wiener filter, on the other hand, had difficulty dealing with more complex noise. Overall, our DeepACE-Mask model had the highest correlation relative to clean electrodograms. Coupled with the similar performance of DeepACE-Mask and DeepACE-wMSE in terms of SNR improvement, DeepACE-Mask was selected in the listening test.

4.2.3 STOI

In addition to the objective evaluation in the electrodogram domain, STOI was used to evaluate the objective intelligibility of speech in the vocoded audio domain. Figure 4.4 illustrates the STOI scores obtained by the tested algorithms in different speech and noise conditions. All four algorithms have showed improvement with respect to ACE. Three deep learning based algorithms outperformed the wiener filter. Among them, our DeepACE-Mask and DeepACE-wMSE models were better than TasNet+ACE and they achieved similar performance. The results for speech datasets TIMIT and LibriVox showed the same trend as HSM dataset, due to space limitations, their results were not presented. In general, the results of STOI in vocoded audio domain were consistent with the previous SNR improvement in the electrodogram domain. As for clean speech, the mean STOI scores of HSM speech dataset obtained by ACE, baseline model TasNet+ACE, our DeepACE-wMSE and DeepACE-Mask were 0.8074, 0.7891,0.7955 and 0.8038, respectively. This means that our model objectively did not affect the speech understanding of CI users in quiet conditions.

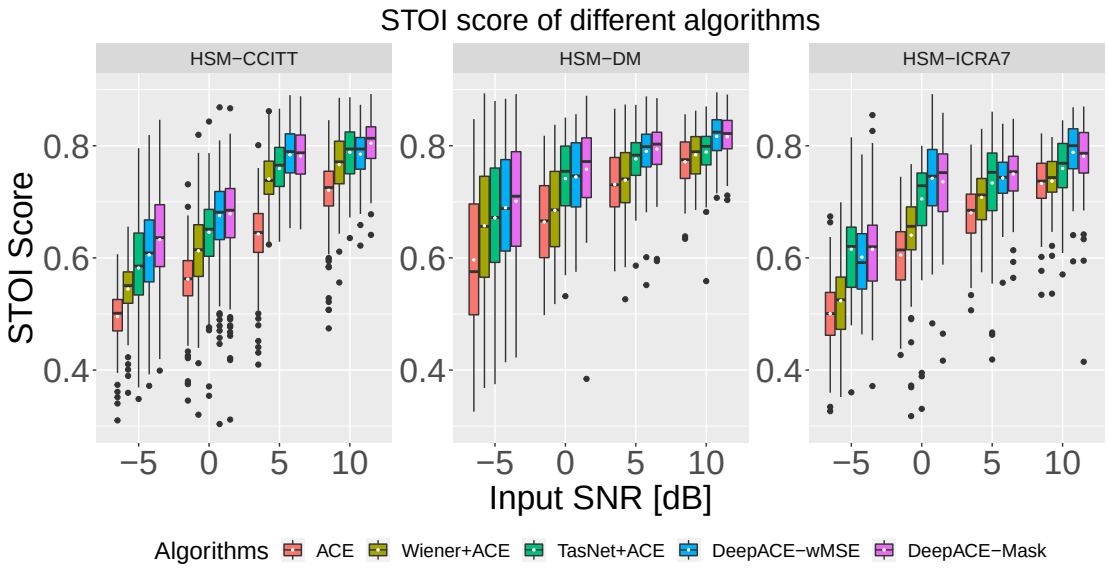

Figure 4.4: Box plot of STOI scores obtained by the tested algorithms in different noise conditions. The black horizontal bars within each of the boxes represent the median for each condition, the diamond-shaped marks indicate the mean score, the top and bottom extremes of the boxes indicate the 75% and 25% quartiles, respectively, and whiskers indicate the variability outside the upper and lower quartiles.

In order to provide corresponding objective instrumental evaluation results for the listening tests, WRS was calculated by STOI as estimiated speech understanding in the listening test conditions. The results of WRS are shown in Figure 4.5.

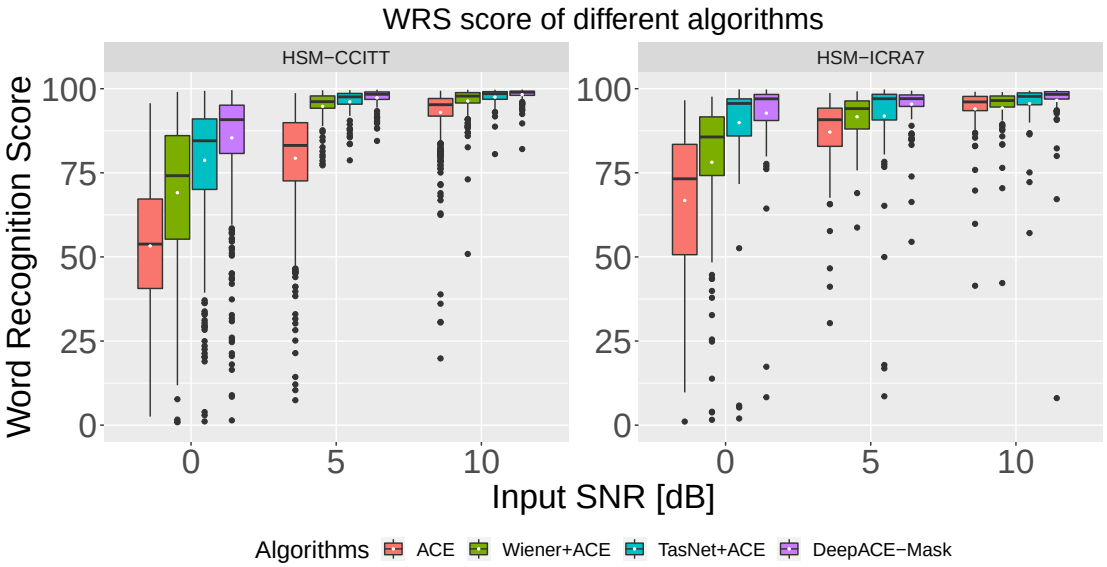

Figure 4.5: Box plot of WRS estimated from the STOI scores of tested algorithms in listening tests conditions. The black horizontal bars within each of the boxes represent the median for each condition, the diamond-shaped marks indicate the mean score, the top and bottom extremes of the boxes indicate the 75% and 25% quartiles, respectively, and whiskers indicate the variability outside the upper and lower quartiles.

4.2.4 Generalization ability

Since the speech and noise samples of our test dataset were significantly different from the training dataset, all the above mentioned objective instrumental evaluation results have already demonstrated that our optimized DeepACE models had good generalization ability. Furthermore, in order to compare the results from [21], the same dataset setting was also implemented. 30% of HSM speech dataset mixed with ICRA7 and CCITT was used for training the DeepACE-wMSE model, the rest 70% was used for the evaluation. As shown in Figure 4.6, when the speech and noise for testing were seen in training, the seen noise can be easily removed. However, compared to DeepACE-Mask, it still eliminated some speech information. Figure 4.7 shows the SNR improvement and LCC of DeepACE-wMSE, DeepACE-wMSE with seen testing data and DeepACE-Mask. The SNR value was significantly improved when the noise types were familiar to the network. But DeepACE-Mask still achieved the highest correlation related to the clean electrodograms. For the other unseen speech and noise dataset, there was no significant improvement from DeepACE-wMSE with seen testing data.

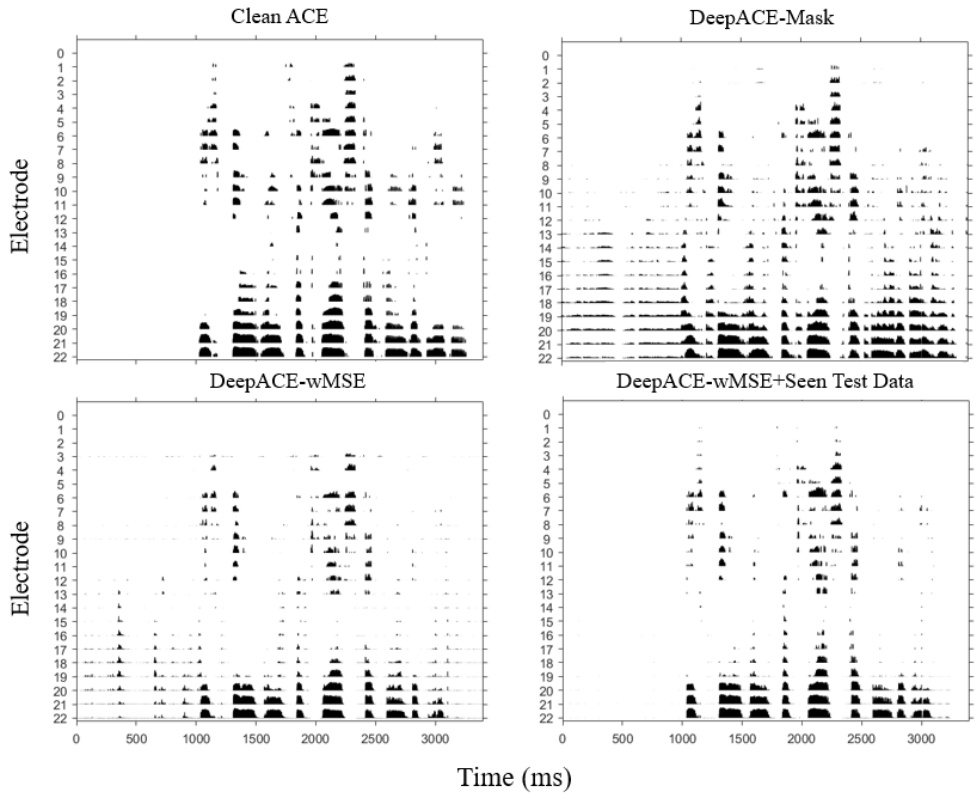

Figure 4.6: Electrodograms for clean speech and speech with noise produced by different algorithms with seen/unseen testing dataset, electrode numbers increase as the corresponding frequency range increases.

DeepACE-wMSE with seen testing data greatly enhanced the noise reduction ability for the same kind of noise, but the problem of disrupted speech structure still existed. Due to time constraints, the same training for DeepACE-Mask was not performed.

4.3 Results in subjects

4.3.1 Listening Test Results

Figure 4.8 shows the percentage of understood words by subjects in quiet. The unprocessed ACE and DeepACE-Mask were tested for this condition. Since there is no noise to reduce, the main purpose is to test if DeepACE-Mask will introduce artifacts on the clean speech. Results show that in most cases, the performance of both models is quite similar. Besides, for each subject the fluctuations in performance are not significant. Thus, DeepACE-Mask does not affect speech understanding in quiet when compared to ACE.

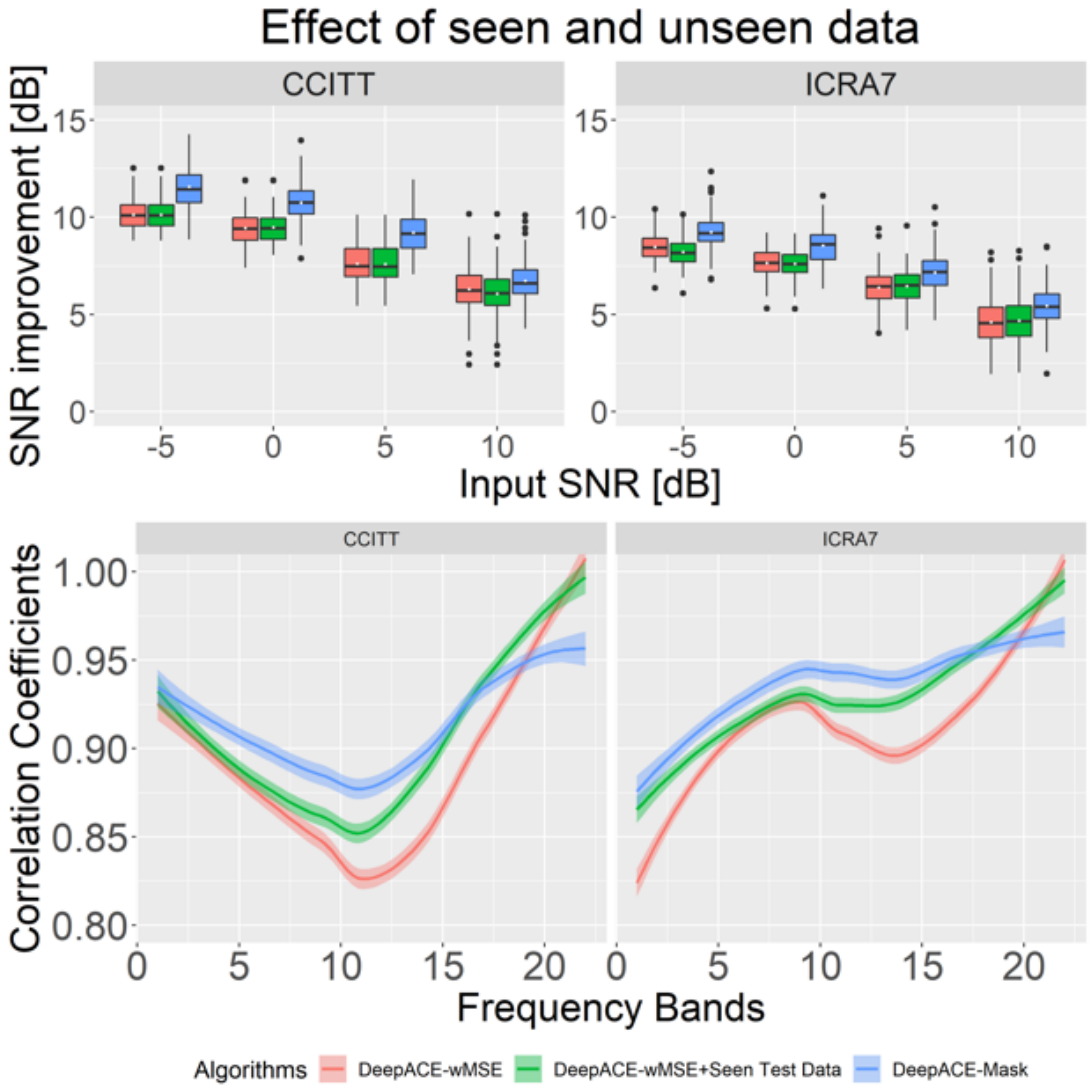

Figure 4.7: Box plot of SNR improvement and plot of linear correlation coefficients between processed and clean electrodograms across bands of different DeepACE models trained with seen/unseen testing dataset. In the upper part, the black horizontal bars within each of the boxes represent the median for each condition, the diamond-shaped marks indicate the mean score, the top and bottom extremes of the boxes indicate the 75% and 25% quartiles, respectively, and whiskers indicate the variability outside the upper and lower quartiles. In the lower part, shaded area represents standard deviations.

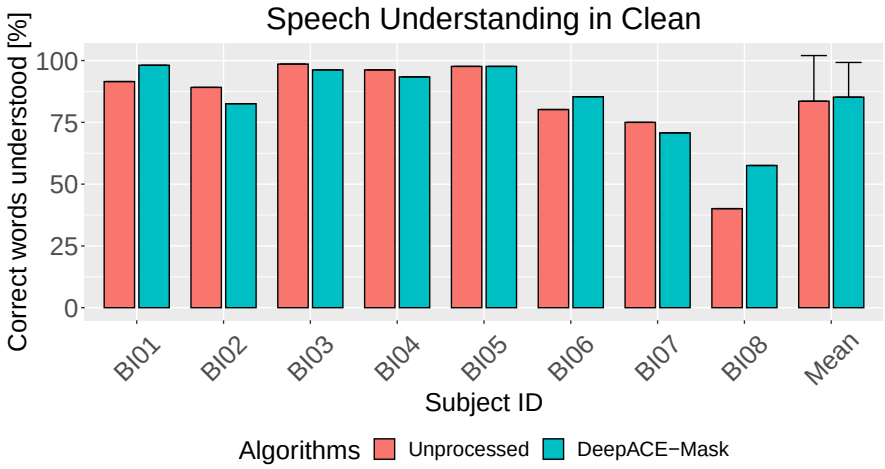

Figure 4.8: Speech intelligibility score in % of correct words understood for the HSM sentence test by subject and condition. The right-most bar group indicates the mean across subjects and per condition (error bars indicate standard deviations). Scores were obtained in quiet conditions.

Figure 4.9 shows the percentage of understood words of different algorithms for CCITT and ICRA7 background noises. Although the tested SNR values were different for each subject, our main concern here is the improvement brought by different algorithms with respect to ACE. From the mean value we found that all three algorithms are capable of improving substantially speech intelligibility in noise conditions then compared to the unprocessed ACE. Among them the DeepACE-Mask outperforms the other two and achieved an average improvement of about 20%. Moreover, in any case it is better than ACE. The subjective results once again demonstrate the effectiveness of our DeepACE model. Besides, the performance of TasNet+ACE model fluctuates greatly from subject to subject. Sometimes it can get very close to DeepACE's improvement, but sometimes it's only slightly better than ACE.

It's also worth noting that the Wiener+ACE also achieves a similar performance to TasNet+ACE, which was not expected according to the objective evaluation. The results in the previous section demonstrate that Wiener Filter can only bring very limited improvement in speech intelligibility, especially for non-stationary noise. But for Subjects BI02 and BI05, its performance is close to or even better than DeepACE-Mask. The reason may be that although wiener filter cannot remove noise to a large extent, it does not destroy the structure of speech too much. In contrast, the deep learning models (DeepACE-Mask and TasNet+ACE) remove most of the noise while also removing a small part of the speech information. It also cannot be ignored that BI02 and BI05 are the youngest of all subjects, they might be more sensitive to speech distortion than noise. With increasing age, patients become more and more sensitive to noise. For the eldest

of the subject BI07, Wiener+ACE produced barely any improvement. Despite this, the overall trends in listening test results are still consistent with objective assessments.

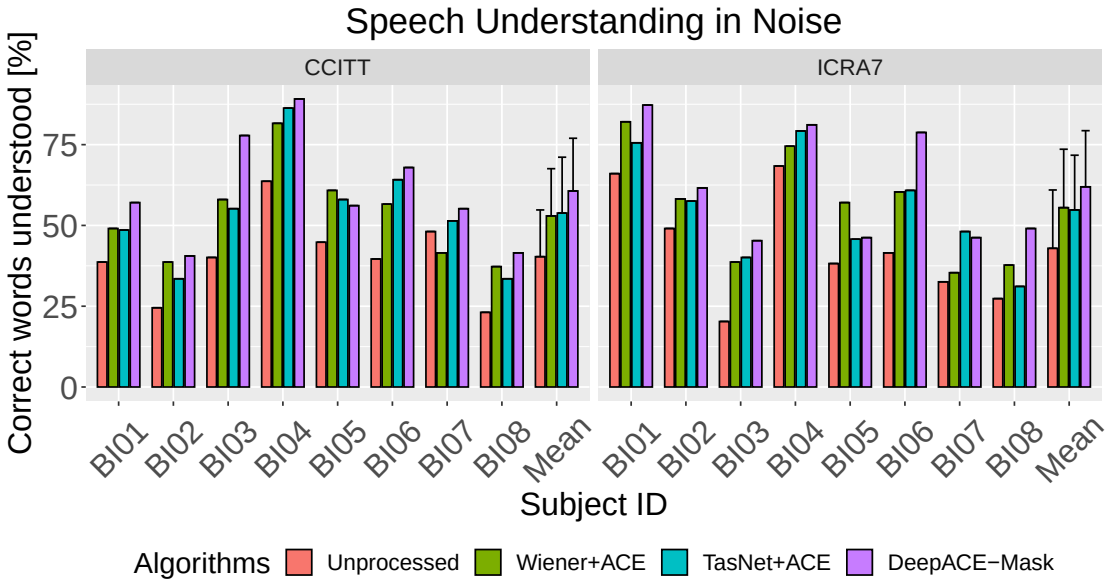

Figure 4.9: Speech intelligibility score in % of correct words understood for the HSM sentence test by subject and condition. The right-most bar group indicates the mean across subjects and per condition (error bars indicate standard deviations). Scores were obtained in noise conditions with 0, 5, 10 dB CCITT and ICRA7 noise.

4.3.2 Statistical Results

First of all, the speech understanding performance of all algorithms was evaluated. Figure 4.10 shows a violin plot overlapped with boxplot indicating the data distribution, mean and median values of the absolute performance of speech understanding processed by different algorithms. The black horizontal bars within each of the boxes represent the median for each condition, the diamond-shaped marks indicate the mean score, the top and bottom extremes of the boxes indicate the 75% and 25% quartiles, respectively, and whiskers indicate the variability outside the upper and lower quartiles. The stars at the top show the significance between the corresponding two groups.

According to the indicators and data distribution shown in the figure, it can be quantitatively seen that three algorithms were all better than the unprocessed ACE. To assess weather this is the case, a significance test was then performed. All the normality test results are shown in Table 4.7.

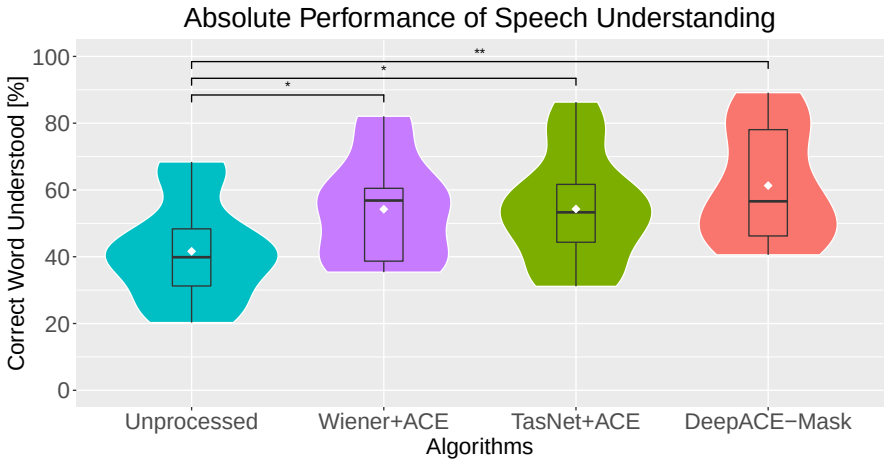

Figure 4.10: Violin plot overlapped with boxplot indicating the data distribution, mean and median values of the absolute performance of speech understanding processed by different algorithms. The black horizontal bars within each of the boxes represent the median for each condition, the diamond-shaped marks indicate the mean score, the top and bottom extremes of the boxes indicate the 75% and 25% quartiles, respectively, and whiskers indicate the variability outside the upper and lower quartiles. The stars at the top show the significance between the corresponding two groups.

The test revealed that all four sets of data are normally distributed and homogeneous of variance. Thus, a one way ANOVA was performed to compare the effect of algorithms on absolute performance. It showed that there was a statistically signifficant difference in absolute performance between at least two groups ($F(3) = [4.24]$, $p = 0.00215$). A subsequent pairwise multiple comparison using t-test gave the results shown in Table 4.8.

Table 4.7: Normality test analysis of result of absolute performance.

Algorithms	Samples	Mean	Standard Deviation	Skewness	Kurtosis	Homogenity of Variance	Shapro-Wilks Test		Normal Distribution?
							w-value	p-value	
Unprocessed	16	41.6	14.9	0.3884	2.0141	0.8874 (homogeneous)	0.9345	0.2874	Yes
Wiener+ACE	16	54.2	15.6	0.4179	1.8618		0.8943	0.0653	Yes
TasNet+ACE	16	54.3	16.4	0.3527	2.0564		0.9561	0.5921	Yes
DeepACE-Mask	16	61.3	16.8	0.3568	1.5025		0.9040	0.0933	Yes

Results showed that the performance of the three algorithms were significantly different from the unprocessed ACE, and the difference between DeepACE-Mask and ACE was the most significant. This illustrated that the three different speech enhancement algorithms all played a positive role in improving the speech understanding of CI users.

Furthermore, in order to compare the superiority among the three algorithms, a statistical

Table 4.8: Results of pairwise multiple comparisons using t-test.

Algorithms / Algorithms	DeepACE-Mask	TasNet+ACE	Unprocessed
TasNet+ACE	0.2403	-	-
Unprocessed	0.0015	0.0291	-
Wiener+ACE	0.2347	0.9924	0.0263

analysis was performed on the benefit they brought with respect to ACE. Figure 4.11 shows the benefit of speech understanding processed by different algorithms. According to the indicators and data distribution shown in the figure, it can be quantitatively seen that DeepACE-Mask is significantly better than other algorithms. For further quantitative investigation, normality analysis was performed again. The results are shown in Table 4.9.

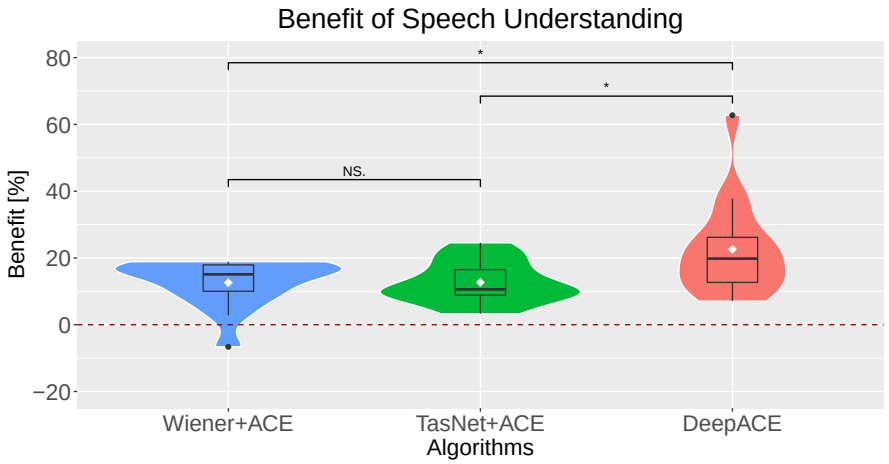

Figure 4.11: Violin plot overlapped with boxplot indicating the data distribution, mean and median values of the benefit of speech understanding processed by different algorithms compared to ACE. The black horizontal bars within each of the boxes represent the median for each condition, the diamond-shaped marks indicate the mean score, the top and bottom extremes of the boxes indicate the 75% and 25% quartiles, respectively, and whiskers indicate the variability outside the upper and lower quartiles. The stars at the top show the significance between the corresponding two groups.

Although all three sets of data were homogeneous of variance, only one set of data was normally distributed. Due to this, a non-parametric analysis of variance had to be performed, i.e. Kruskal-Wallis. A statistically signifficant difference ($Chisquare = 6.831$, $df = 2$, $p = 0.03287$) were found among the three algorithms. Table 4.10 shows the results of pairwise multiple comparisons by Wilcoxon test, which demonstrate that our DeepACE-Mask model has a significant improvement over the other two algorithms,

Table 4.9: Normality test analysis of result of benefit.

Algorithms	Samples	Mean	Standard Deviation	Skewness	Kurtosis	Homogenity of Variance	Shapro-Wilks Test		Normal Distribution?
							Statistic	P value	
Wiener+ACE	16	12.5885	7.0507	-1.2780	3.8753	0.09806 (homogeneous)	0.8308	0.0072	No
TasNet+ACE	16	12.6828	6.3274	0.3653	1.9345		0.9478	0.4553	Yes
DeepACE-Mask	16	22.6003	14.1382	1.3354	4.4437		0.8588	0.0184	No

while there is no significant difference between Wiener+ACE and TasNet+ACE. This further illustrates the practicality of the DeepACE model proposed in this work.

Table 4.10: Results of pairwise multiple comparisons using Wilcoxon test.

Algorithms / Algorithms	DeepACE-Mask	TasNet+ACE
TasNet+ACE	0.006145	-
Wiener+ACE	0.01055	0.6608

5 Discussion

In this work, two optimized DeepACE models, i.e. DeepACE-wMSE and DeepACE-Mask, were proposed. The results showed that our optimized DeepACE models outperformed the original DeepACE and two baseline models TasNet+ACE and Wiener+ACE in both objective instrumental evaluation and listening tests. The improvement was steady across different speech and noise datasets.

Early results [21] showed that TasNet+ACE was better than DeepACE when they had the same architecture in TCN. In this work, our optimized DeepACE models successfully surpassed the TasNet+ACE. TasNet used SI-SDR as loss function, which is specifically designed for speech tasks. On the contrary, original DeepACE only used a common loss function MSE which is not specific for electrodograms. We proposed two variants of DeepACE based on the hearing perception of CI users. Thus, DeepACE-wMSE and DeepACE-Mask were actually two implementations of a same idea, which aims to take into account the incorrect band selection in the loss function and the TCN architecture. Finally, the performance was greatly improved compared to TasNet+ACE and original DeepACE. At the same time, the results reconfirmed the consistent performance of DeepACE and ACE in a quiet.

It is noteworthy that for TasNet+ACE and Wiener+ACE, there was an inconsistency between the results in the objective instrumental evaluation and the listening tests. TasNet+ACE achieved far more better performance in SNR improvement and STOI scores but performed poor in listening tests and vice versa. This is different from the results in [21], which showed that the TasNet+ACE surpassed the Wiener+ACE in both objective instrumental evaluation and listening test. However, [21] added some speech and noise data with similar features as the testing dataset into the training dataset, which obviously improved the performance of TasNet+ACE. A reason might be that compared to Wiener+ACE, TasNet+ACE can remove more noise but disrupt speech structure at the same time. The discrupted speech structure caused distortion and affected the speech understanding of CI users. Unfortunately, as far as to the author's knowledge, there is no one else that has tested the effect of either TasNet+ACE on speech understanding of CI users or TasNet on speech understanding of normal hearing listeners. Therefore, our conjecture cannot be further verified.

It was also observed that the results with DEMAND noise dataset fluctuate greatly in each condition compared to CCITT and ICRA7 noise. This is because the DEMAND dataset is far more realistic and challenging. After checking the performance of each single sentence, we found that our models still had difficulties to remove the speech

in background noise. For instance, in the noise scenarios 'Meeting' and 'Cafeteria' of the DEMAND dataset, when there was relatively clear speech in background noise, our model treated it as speech and remained. For the other scenarios like 'Park' and 'Bus Station', the denoising performance is much better. But in general, our models still provided positive improvement in most cases, which showed the generalization ability and practicality of the DeepACE based models.

So far, the vast majority of speech enhancement models for CI users are still front-end. The enhanced speech will then be processed by the sound coding strategies. However, our end-to-end DeepACE models directly let the network to learn the calculation of sound coding strategies and therefore, latency was reduced. For instance, TasNet+ACE has 4 ms latency, while our DeepACE models only have 2 ms. This is crucial for devices such as CIs that need to transmit signals with minimal latency.

Despite the encouraging results, the two optimized DeepACE models still cannot do the best in both noise reduction and speech preservation. The improvement methods proposed for the problems mentioned above and the possible directions for further research will be discussed in the next chapter.

6 Conclusions and Future Work

6.1 Conclusions

This thesis presented an end-to-end deep learning speech denoising and coding strategy to enhance the speech intelligibility of CI users. The inspiration is the recently proposed DeepACE, which effectively reduces the noise. In this context, two new implementations were proposed to optimize the DeepACE for better hearing performance of CI users.

To begin with, we built the model based on the original DeepACE framework. The hyper-parameters were first optimized to determine the best performing TCN and encoder/decoder architectures. Afterwards, based on the experimental results showing that CI users are more sensitive to band selection correction than to the magnitude distortion, a custom loss function (DeepACE-wMSE) and architecture (DeepACE-Mask) were implemented respectively to suppress the incorrect band selection. Besides, in addition to the commonly used CCITT and ICRA7 datasets, a DEMAND dataset with more realistic and challenging background noises was introduced to investigate the generalization ability of the models. A rigorous dataset setup was used to ensure that all test speech and noise samples were unseen in training.

These two models were then evaluated through an objective instrumental evaluation and a listening test with CI users. The SNR improvement and the LCC between target and predicted electrodograms were calculated in the electrodogram domain while STOI scores was implemented in the vocoded audio domain. In the listening test, the percentage of understood words from 8 CI patients was recorded. The performance was compared to the ACE and two baseline models TasNet+ACE and Wiener+ACE.

Results showed that the two proposed models outperformed the other baseline models, brought great and steady improvement objectively. The largest average SNR improvement reached 11 dB with LibriVox speech and the largest average STOI improvement reached 0.15 with HSM speech, under 0 dB CCITT noise condition. Even if the complex DEMAND noises were utilized, a positive improvement was still reached. In the listening test, the selected DeepACE-Mask model also achieved the best performance and improved by more than 20% speech intelligibility with respect to ACE. Results of statistical analysis showed that our DeepACE-Mask model and the two baseline models all performed significantly better than ACE in speech intelligibility. Among them, DeepACE-Mask brought the most significant benefit of speech understanding with respect to ACE. Further research found that two different implementations of the same idea actually led to different preferences. DeepACE-wMSE was much better at noise

reduction but disrupted the speech structure. On the other hand, DeepACE-Mask kept most of the speech information but was relatively weak in denoising.

In summary, all evaluation results demonstrated that the two proposed DeepACE models can get better performance with close parameter quantities and less latency compared to the front-end TasNet model. In quiet condition, the DeepACE models performed almost identically to ACE. Meanwhile, the two proposed models were no worse than the original ACE most of the time. Still, lots of challenges need to be clarified and overcome, like how to make the model retain as much speech information as possible while removing more noise, and how to further improve the performance for realistic noisy environments. But in general, the present study indicated that the optimized DeepACE models had great potential for application to CIs to ameliorate speech understanding they provide.

6.2 Future Work

In this last section, the typical factors that can be considered in future investigation are discussed. First of all, the architecture of DeepACE still has a lot of room for improvement on the basis of this work. Deep encoder/decoder and large TCNs can significantly improve performance when used individually, but how to combine them for better performance remains a question. Kadiouglu et al. [70] proposed three variants to improve the encoder/decoder structure. Due to time constraints, this work only implemented the first Deep encoder/decoder structure. It would be valuable to investigate whether the remaining two variants are suitable for large TCNs. Because improving the encoder/decoder structure can get a huge improvement by only adding a small amount of parameters.

In our DeepACE-Mask model, we used some 1D convolutional layers to reduce the dimensionality of the input feature. The advantage of the convolution layer is that it can learn certain properties. Another commonly used method for dimension reduction is pooling. Some works [81] also showed that sometimes pooling is better than convolution. Pooling is a cheaper operation than convolution, both in terms of the amount of computation and number of parameters, which shows shows great potential to be applied to wearable devices such as CIs that require low power consumption.

Besides, this work showed the inconsistency of the results of TasNet+ACE and Wiener+ACE in objective instrumental evaluation and listening tests. One guess is that different patients have different levels of sensitivity to speech distortion and noise. Our DeepACE models achieved better denoising performance, but removed speech information to a certain extent at the same time. A possible improvement measure is to construct two decoders to output the separated speech and noise respectively, and then mix a part of the noise into the final output to compensate for speech distortion. Kang et al. [82] investigated the perceptual sensitivities of the CI recipients to noise and distortion and implemented a new loss function. By adjusting the weights for trading off the speech

distortion and the noise residue, their contributions to speech intelligibility of CI recipients were investigated. An LSTM trained with preference-biased-loss was developed. This also shows great potential to improve speech understanding of CI users. In this case, two decoders were constructed to output the speech and noise respectively, and then mix a part of the noise into the final output to compensate for speech distortion.

Finally, to further improve the hearing perception of CI users, a psychoacoustic model could be considered. In this work, we tried to implement the idea in PACE [2] into the loss function, but due to its complex structure, the model could not converge. Zheng et al. [63] introduced a perceptually weighted error into the loss function, using a perceptual weight function which is widely adopted in speech codecs. The results showed that the temporal fine structure could be better extracted. So it will be interesting to implement something like a LPC filter for perceptual weighting in ths loss function.

Bibliography

[1] Cochlear Implants. https://mackayhearing.com.au/cochlear-implants/. Accessed Jun. 10, 2022.

[2] Waldo Nogueira, Andreas Büchner, Thomas Lenarz, and Bernd Edler. A psychoacoustic" nofm"-type speech coding strategy for cochlear implants. *EURASIP Journal on Advances in Signal Processing*, 2005(18):1–16, 2005.

[3] Michael A Nielsen. *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, USA, 2015.

[4] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Han Li, Kean Chen, Lei Wang, Jianben Liu, Baoquan Wan, and Bing Zhou. Sound source separation mechanisms of different deep networks explained from the perspective of auditory perception. *Applied Sciences*, 12(2):832, 2022.

[7] Waldo Nogueira, Thilo Rode, and Andreas Büchner. Spectral contrast enhancement improves speech intelligibility in noise for cochlear implants. *The Journal of the Acoustical Society of America*, 139(2):728–739, 2016.

[8] Thomas Baer, Brian CJ Moore, and Stuart Gatehouse. Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times. *Journal of rehabilitation research and development*, 30:49–49, 1993.

[9] Li-Ping Yang and Qian-Jie Fu. Spectral subtraction-based speech enhancement for cochlear implant patients in background noise. *The journal of the Acoustical Society of America*, 117(3):1001–1004, 2005.

[10] Nicolas Guevara, Alexis Bozorg-Grayeli, Jean-Pierre Bebear, Marine Ardoint, Sonia Saaï, Dan Gnansia, Michel Hoen, Philippe Romanet, and Jean-Pierre Lavieille. The voice track multiband single-channel modified wiener-filter noise reduction system for cochlear implants: patients' outcomes and subjective appraisal. *International Journal of Audiology*, 55(8):431–438, 2016.

[11] Raphael Koning, Nilesh Madhu, and Jan Wouters. Ideal time–frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners. *IEEE Transactions on Biomedical Engineering*, 62(1):331–341, 2014.

[12] Tom Gajecki and Waldo Nogueira. A synchronized binaural n-of-m sound coding strategy for bilateral cochlear implant users. In *Speech Communication; 13th ITG-Symposium*, pages 1–5. VDE, 2018.

[13] Enrique A Lopez-Poveda, Almudena Eustaquio-Martín, Joshua S Stohl, Robert D Wolford, Reinhold Schatzer, José M Gorospe, Santiago Santa Cruz Ruiz, Fernando Benito, and Blake S Wilson. Intelligibility in speech maskers with a binaural cochlear implant sound coding strategy inspired by the contralateral medial olivocochlear reflex. *Hearing Research*, 348:134–137, 2017.

[14] Ying-Hui Lai, Fei Chen, Syu-Siang Wang, Xugang Lu, Yu Tsao, and Chin-Hui Lee. A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation. *IEEE Transactions on Biomedical Engineering*, 64(7):1568–1578, 2016.

[15] Ying-Hui Lai, Yu Tsao, Xugang Lu, Fei Chen, Yu-Ting Su, Kuang-Chao Chen, Yu-Hsuan Chen, Li-Ching Chen, Lieber Po-Hung Li, and Chin-Hui Lee. Deep learning–based noise reduction approach to improve speech intelligibility for cochlear implant recipients. *Ear and hearing*, 39(4):795–809, 2018.

[16] Yi Hu and Philipos C Loizou. Environment-specific noise suppression for improved speech intelligibility by cochlear implant users. *The Journal of the Acoustical Society of America*, 127(6):3689–3695, 2010.

[17] Nursadul Mamun, Soheil Khorram, and John HL Hansen. Convolutional neural network-based speech enhancement for cochlear implant recipients. In *Interspeech*, volume 2019, page 4265. NIH Public Access, 2019.

[18] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *International conference on latent variable analysis and signal separation*, pages 91–99. Springer, 2015.

[19] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

[20] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266, 2019.

[21] Tom Gajecki and Waldo Nogueira. An end-to-end deep learning speech coding and denoising strategy for cochlear implants. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3109–3113. IEEE, 2022.

[22] Blake S Wilson and Michael F Dorman. Cochlear implants: current designs and future possibilities. *J Rehabil Res Dev*, 45(5):695–730, 2008.

[23] Philipos C Loizou and Gibak Kim. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE transactions on audio, speech, and language processing*, 19(1):47–56, 2010.

[24] Dongmei Wang and John HL Hansen. Speech enhancement for cochlear implant recipients. *The Journal of the Acoustical Society of America*, 143(4):2244–2254, 2018.

[25] Abigail A Kressner, Tobias May, and Torsten Dau. Effect of noise reduction gain errors on simulated cochlear implant speech intelligibility. *Trends in Hearing*, 23:2331216519825930, 2019.

[26] Bas van Dijk, Marc Moonen, Jan Wouters, et al. Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility. *Hearing Research*, 299:79–87, 2013.

[27] Waldo Nogueira, Marta Lopez, Thilo Rode, Simon Doclo, and Andreas Buechner. Individualizing a monaural beamformer for cochlear implant users. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5738–5742. IEEE, 2015.

[28] Christopher J Plack. *The sense of hearing*. Routledge, 2018.

[29] AJ Hudspeth. Snapshot: auditory transduction. *Neuron*, 80(2):536–e1, 2013.

[30] Tohru Yoshioka and Manabu Sakakibara. Physical aspects of sensory transduction on seeing, hearing and smelling. *Biophysics*, 9:183–191, 2013.

[31] World Health Organization. *World report on hearing*, 2021.

[32] Joseph Sataloff and Robert Thayer Sataloff. Hearing loss. 2005.

[33] Jan Wouters, Hugh Joseph McDermott, and Tom Francart. Sound coding in cochlear implants: From electric pulses to hearing. *IEEE Signal Processing Magazine*, 32(2):67–80, 2015.

[34] Philipos C Loizou. Speech processing in vocoder-centric cochlear implants. *Cochlear and brainstem implants*, 64:109–143, 2006.

[35] J-F Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

[36] Florian Mayer, Donald S Williamson, Pejman Mowlaee, and DeLiang Wang. Impact of phase estimation on single-channel speech separation based on time-frequency masking. *The Journal of the Acoustical Society of America*, 141(6):4668–4679, 2017.

[37] Jonathan Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3):235–238, 1977.

[38] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[39] Stephen Grossberg, Krishna K Govindarajan, Lonce L Wyse, and Michael A Cohen. Artstream: a neural network model of auditory scene analysis and source segregation. *Neural networks*, 17(4):511–536, 2004.

[40] Guoning Hu and DeLiang Wang. Auditory segmentation based on onset and offset analysis. *IEEE transactions on audio, speech, and language processing*, 15(2):396–405, 2007.

[41] Guoning Hu and DeLiang Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on neural networks*, 15(5):1135–1150, 2004.

[42] Pascal Scalart et al. Speech enhancement based on a priori signal to noise estimation. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 629–632. IEEE, 1996.

[43] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

[44] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[45] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649. IEEE, 2012.

[46] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[47] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep neural networks for object detection. *Advances in neural information processing systems*, 26, 2013.

[48] Yoshua Bengio. *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[49] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[50] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[51] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[52] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

[53] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[54] Ajay Shrestha and Ausif Mahmood. Review of deep learning algorithms and architectures. *IEEE access*, 7:53040–53065, 2019.

[55] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[57] Dario Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5073. IEEE, 2018.

[58] Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

[59] Martin Strauss and Bernd Edler. A flow-based neural network for time domain speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758. IEEE, 2021.

[60] ETSI 2013. "electromagnetic compatibility and radio spectrum matters (erm); system reference document; short range devices (srd); technical characteristics of wireless aids for hearing impaired people operating in the vhf and uhf frequency range," technical report etsi tr 102 791 v1.2.1. 2013.08.

[61] Federico Bolner, Tobias Goehring, Jessica Monaghan, Bas Van Dijk, Jan Wouters, and Stefan Bleeck. Speech enhancement based on neural networks applied to cochlear implant coding strategies. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6520–6524. IEEE, 2016.

[62] Nursadul Mamun, Soheil Khorram, and John HL Hansen. Convolutional neural network-based speech enhancement for cochlear implant recipients. In *Interspeech*, volume 2019, page 4265. NIH Public Access, 2019.

[63] Nengheng Zheng, Yupeng Shi, Yuyong Kang, and Qinglin Meng. A noise-robust signal processing strategy for cochlear implants using neural networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8343–8347. IEEE, 2021.

[64] Benjamin Beilharz, Xin Sun, Sariya Karimova, and Stefan Riezler. Librivoxdeen: A corpus for german-to-english speech translation and speech recognition. In *Proceedings of LREC*, 2020.

[65] I Hochmair-Desoyer, E Schulz, L Moser, and M Schmidt. The hsm sentence test as a tool for evaluating the speech understanding in noise of cochlear implant users. *The American journal of otology*, 18(6 Suppl):S83–S83, 1997.

[66] Victor Zue, Stephanie Seneff, and James Glass. Speech database development at mit: Timit and beyond. *Speech communication*, 9(4):351–356, 1990.

[67] Hugo Fastl and Eberhard Zwicker. Psychoacoustics - facts and models. *Springer*, 3rd edition, 2007.

[68] Wouter A Dreschler, Hans Verschuure, Carl Ludvigsen, and Søren Westermann. Icra noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment: Ruidos icra: Señates de ruido artificial con espectro similar al habla y propiedades temporales para pruebas de instrumentos auditivos. *Audiology*, 40(3):148–157, 2001.

[69] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. Demand: a collection of multi-channel recordings of acoustic noise in diverse environments. In *Proc. Meetings Acoust*, pages 1–6, 2013.

[70] Berkan Kadıoğlu, Michael Horgan, Xiaoyu Liu, Jordi Pons, Dan Darcy, and Vivek Kumar. An empirical study of conv-tasnet. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7264–7268. IEEE, 2020.

[71] Yuichiro Koyama, Tyler Vuong, Stefan Uhlich, and Bhiksha Raj. Exploring the best loss function for dnn-based low-latency speech enhancement with temporal convolutional networks. *arXiv preprint arXiv:2005.11611*, 2020.

[72] Stefan J Mauger, Komal Arora, and Pam W Dawson. Cochlear implant optimized noise reduction. *Journal of Neural Engineering*, 9(6):065007, 2012.

[73] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE, 2010.

[74] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech.

IEEE Transactions on Audio, Speech, and Language Processing, 19(7):2125–2136, 2011.

[75] Lars St, Svante Wold, et al. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272, 1989.

[76] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[77] Howard Levene. Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, pages 279–292, 1961.

[78] FR Helmert. Die genauigkeit der formel von peters zur berechnung des wahrscheinlichen beobachtungsfehlers director beobachtungen gleicher genauigkeit. *Astronomische Nachrichten*, 88:113, 1876.

[79] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[80] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[81] Shuyang Sun, Jiangmiao Pang, Jianping Shi, Shuai Yi, and Wanli Ouyang. Fishnet: A versatile backbone for image, region, and pixel level prediction. *Advances in neural information processing systems*, 31, 2018.

[82] Yuyong Kang, Nengheng Zheng, and Qinglin Meng. Deep learning-based speech enhancement with a loss trading off the speech distortion and the noise residue for cochlear implants. *Frontiers in Medicine*, 8, 2021.