applied sciences

MDPI

Article

An Electroglottograph Auxiliary Neural Network for Target Speaker Extraction

Lijiang Chen, Zhendong Mo, Jie Ren, Chunfeng Cui and Qi Zhao *

School of Electronic and Information Engineering, Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing 100191, China

* Correspondence: zhaoqi@buaa.edu.cn; Tel.: +86-010-8231-6739

Abstract: The extraction of a target speaker from mixtures of different speakers has attracted extensive amounts of attention and research. Previous studies have proposed several methods, such as SpeakerBeam, to tackle this speech extraction problem using clean speech from the target speaker to provide information. However, clean speech cannot be obtained immediately in most cases. In this study, we addressed this problem by extracting features from the electroglottographs (EGGs) of target speakers. An EGG is a laryngeal function detection technology that can detect the impedance and condition of vocal cords. Since EGGs have excellent anti-noise performance due to the collection method, they can be obtained in rather noisy environments. In order to obtain clean speech from target speakers out of the mixtures of different speakers, we utilized deep learning methods and used EGG signals as additional information to extract target speaker. In this way, we could extract target speaker from mixtures of different speakers without needing clean speech from the target speakers. According to the characteristics of the EGG signals, we developed an EGG_auxiliary network to train a speaker extraction model under the assumption that EGG signals carry information about speech signals. Additionally, we took the correlations between EGGs and speech signals in silent and unvoiced segments into consideration to develop a new network involving EGG preprocessing. We achieved improvements in the scale invariant signal-to-distortion ratio improvement (SISDRi) of 0.89dB on the Chinese Dual-Mode Emotional Speech Database (CDESD) and 1.41dB on the EMO-DB dataset. In addition, our methods solved the problem of poor performance with target speakers of the same gender and the different between the same gender situation and the problem of greatly reduced precision under the low SNR circumstances.

Keywords: speech extraction; SpeakerBeam; electroglottograph; pre-processing

Citation: Chen, L.; Mo, Z.; Ren, J.; Cui, C.; Zhao, Q. An Electroglottograph Auxiliary Neural Network for Target Speaker Extraction. *Appl. Sci.* **2023**, *13*, 469. https://doi.org/10.3390/app13010469

Academic Editors: Douglas O'Shaughnessy and Javier Hernando

Received: 19 October 2022
Revised: 5 December 2022
Accepted: 26 December 2022
Published: 29 December 2022

1. Introduction

Speech extraction refers to the extraction of individual signals from mixed signals, which was first proposed to address the cocktail party problem [1,2]. Interference in speech can decrease the quality of information being communicated. In addition, interference in the speech can severely affect other related tasks, such as automatic speech recognition (ASR). Current speech recognition technology can accurately recognize an individual speaker, but when there are two or more speakers, the accuracy of speech recognition is greatly reduced. Thus, speech extraction has become an important factor in obtaining speech with better quality and intelligibility.

Many studies have been carried out on the speech extraction problem. The initial attempt was based on traditional methods. For example, methods based on signal processing estimate the power spectrogram of noise or ideal Wiener filters from the perspective of signal processing, e.g., spectral subtraction [3] and Wiener filter [4,5]. Additionally, another algorithm based on decomposition is represented as follows:

$$X = WH \tag{1}$$

where X is the spectrogram of a signal and is decomposed into the matrix product of a base matrix W and activation matrix H. Non-negative matrix factorization (NMF) lets W and H be non-negative to obtain non-negative matrix factorization[6,7], which can be used to obtain the basic spectral patterns of non-negative data. Another method is called computational auditory scene analysis (CASA) [8], which uses auditory grouping cues. The CASA method [9] models auditory signals and utilizes the similarities between the fundamental frequencies of speech signals. A method based on Bayesian inference rules was proposed in a study by Barniv [10]. This method describes the formation process of pure tone sequencing as "auditory streaming" and estimates auditory sequencing by processing prior probability using the Bayesian criterion. Another method based on neural computation represents auditory flows in terms of units of neurons and competition between auditory flows is realized by inhibitory connections between neurons. Wang [11] developed a method that utilizes local and global inhibitory mechanisms to separate auditory flows. Another method by Mill [12] is based on time coherence, which uses a prediction mechanism to promote competition among different groups. The methods mentioned above extract speaker by establishing mathematical models, so their performances are not applicable in more complex situations. Additionally, the accuracy of the extracted speech using traditional methods is far below that of the deep learning methods.

In recent years, deep learning methods have achieved good results in speech extraction under challenging conditions such as non-stationary interference. There have been many studies on speech extraction using deep learning methods. However, most of the previous studies have relied on the use of clean speech from the target speaker and they have achieved their targets by training networks on data from the target speaker, thus generating models to extract particular targets. The models are trained on fixed speaker pairs or target speakers. These models also rely on the assumptions that the amount of the data being put into the network is sufficient and substantial, and that speakers without substantial data cannot be extracted. However, clean speech from particular target speakers cannot always be obtained, and sometimes only a few of the utterances from the target speaker can be recorded in a conversation. So, we tried to find a method that could solve the problem of needing substantial clean speech from the target speaker.

Some methods have been already proposed to solve this problem. The SpeakerBeam network [13] was proposed to solve the speech extraction problem by training speaker-independent models that are informed by additional speaker information rather than creating particular models for target speakers. However, this method relies on data provided by the additional speaker information and the additional information can only be obtained from conversations without any speech overlaps or the personal devices, which is inadequate for training.

To solve the problem of needing additional speech records, we explored some new approaches to speech extraction and investigated EGG signals. These kind of signals come from the vibrations in human throats, and can be recorded without interference from other noises. In addition, the EGG signals of particular speakers can be recorded during conversations in any situation, which increases the amount of additional target speaker information. By utilizing EGG signals from a target speaker, features can be extracted from the signals and applied to deep learning methods for speech extraction. Because EGG signals provide information about particular speakers, designated speech can be extracted from a conversation. Since EGG signals can be obtained in any situation during the process of speaking, they can also be used for real-time speech extraction.

This paper is organized as follows: In Section 2, we examine studies related to speech extraction and present our work on the topic. Section 3 introduces the materials and the methods used in our study and illustrates our model in detail. In Section 4, we compare the results from our network to those from previous studies using different datasets and under different signal-to-noise ratios (SNRs). In Section 5, we discuss our work and findings. Finally, Section 6 presents our conclusion and introduces the future directions of our study.

2. Related Works

Speech extraction algorithms can be divided into single-channel speech extraction algorithms and multi-channel speech extraction algorithms [14,15] according to the number of microphones that are used to record the speakers. Single-channel speech extraction is usually solved with time-domain or frequency-domain methods while multi-channel speech extraction is solved using the methods for extracting coherent signals from different speakers. In our work, we mainly focus on the problems of single-channel speech extraction.

With the rapid development of machine learning and artificial intelligence, the performance of deep-learning-based audio signal processing algorithms has been further improved. Speech extraction technology has also advanced thanks to the development of deep learning [16–24]. In most cases, target speech is extracted in the frequency domain. In these methods, networks obtain spectrograms of the target speech using short-time Fourier transforms(STFTs) and then generate spectrograms of the estimated speech. Other methods are based on the time domain, which mainly work by extracting the time domain features of target speech and can solve the problem of phase mismatch by using adaptive front-end and direct regression instead of STFTs.

Deep clustering (DC) is a frequency-domain method that was proposed by Hershey [25] in 2016. In this method, the amplitude spectrogram features of (T, F) dimension mixed speech are mapped into a higher dimension (T, F, D) deep embedded feature space, i.e., each time-frequency unit (T, F) is mapped into a D-dimensional feature vector, which makes the mixed input features more distinguishable. The target of this method is to generate binary masks, which allocate the areas belonging to the target speech with a mask of 1 and the areas belonging to the other speech with a mask of 0. By multiplying the binary masks by the spectrograms of mixed speech, the network can cover the areas of noise on the spectrograms of mixed speech and obtain target speech from the mixed speech.

Another technique is permutation invariant training(PIT). In 2017, Yu [26] proposed the PIT method and applied it to the speech extraction task. This method selects the smallest mean square error (MSE) as the optimization target and effectively solves the problem of the permutation of the target and interference to find the best match for the desired target compared to DC. In 2021, Yousefi [27] combined the traditional PIT method with long short-term memory (LSTM) and managed to improve the algorithm efficiency. The algorithm has a probabilistic optimization framework and solves the problem of the low efficiency of PIT by finding the best output label allocation. This method is significantly superior to traditional speech extraction methods that use the signal-to-distortion ratio(SDR) and source-to-interference ratio (SIR). In conclusion, the PIT algorithm provides a good training criterion for speaker-independent speech extraction to deal with the permutation and combination problems.

In summary, the frequency-domain methods depend on the consistency of the outputs. These methods have multiple outputs, which make it difficult to define the target outputs of the network. Moreover, inverse short-time Fourier transforms (ISTFTs) with enhanced amplitude spectrogram and original mixed-phase spectrograms have certain impacts on speech extraction performance.

To solve the problems with the frequency-domain methods, the time-domain methods are used. Conv-TasNet is one of the most common methods. In 2019, Luo [28] proposed the convolution time-domain audio separation network (Conv-TasNet), which is superior to several time-frequency amplitude masks for dual-speaker speech extraction. This method build learnable front ends instead of STFTs, thereby generating features that are similar to those of a spectrogram.

In 2021, Li [29] proposed the dual-path recurrent neural network (DPRNN), which breaks long audio clips into smaller chunks to optimize the recurrent neural network (RNN) in a deep model. The DPRNN significantly minimizes the model size compared to the time-domain audio separation network (TasNet) and enhances speech extraction performance.

Although the studies mentioned above have obtained excellent results in dealing with speech extraction problems, the networks tend to be complicated and lack portability.

Time-domain speech extraction methods can achieve good extraction effects, but they are calculated point by point, and the time-domain models tend to be complex and require expensive computation costs.

SpeakerBeam and VoiceFilter [30] are examples of target speech extraction models that can be used in both the frequency domain and time domain using extra information from target speakers. SpeakerBeam uses a sequence summary network to generate spectrograms containing the features of the target speaker's speech while VoiceFilter concatenates spectrogram features and d-vector features, which are extracted from the last hidden layer of the deep neural network, to estimate the clean speech of target speakers. In 2018, Žmolíková [31] optimized this method using ASR technology and utilized predicted hidden Markov model (HMM)-state posteriors to improve the masks. In 2019, Žmolíková [32] also refined SpeakerBeam to train models using extra information about target speaker instead of training particular models for target speakers. This network utilizes information about target speakers from adapted speech [33], both for single-channel and multi-channel speech extraction and achieves better performance than former networks. SpeakerBeam has also been modified with an attention mechanism [34] to extract features from the additional information, which effectively improves the performance of the multimodal SpeakerBeam network [35]. Additionally, Delcroix [36] proposed an implementation of SpeakerBeam in the time-domain, with auxiliary speaker information added to the network. However, all of the speech extraction algorithms mentioned above are based on the premise that all mixed speech can be separated to obtain clean speech from a target speaker, but it is difficult to achieve this in practical application.

Therefore, we focused on finding other signals that could provide information about the target speech and eventually identified EGGs [37–40], which were invented by Adrian Fourcin [41]. EGGs are a kind of skin electrical signal that measure vocal cord vibrations during laryngeal vocalization. The acquisition of EGG signals is not susceptible to other noise or vibrations. Thanks to method of collecting EGG signals directly from people's throat, they can be obtained effectively in extremely noisy environments. In 2020, Bous [42] utilized a deep neural network and EGGs to estimate glottal glosure instants(GCI), and optimized the method by using the analysis synthesis settings of real speech signals, which improved the final performance for glottal closure instants. In 2020, Cangi [43] proposed a measurement to test the reliability of EGGs, which shows the differences in the values of vowels between different genders or individuals. This method illustrates the possibility of EGGs being used in speech processing. The features of the EGG signal extraction module were proposed based on LSTM units [44] to replace SpeakerBeam's feature extraction network. This method works through voiced segment extraction, feature extraction, and F0 smoothing and achieves a 91.2% accuracy in the classification of EGGs. In 2022, Chen [45] proposed a cross-modal emotion distillation model that uses fundamental frequencies from EGG signals to improve the emotion recognition accuracy of emotional databases. Since previous studies have proved that EGGs can improve performance for other acoustic tasks, we applied EGGs to speech extraction tasks to verify whether EGGs can improve speech extraction performance.

In our work, considering that EGG signals are not susceptible to other noises, we proposed a network based on EGGs to extract target speakers from mixed speech. This method differs from previous speech extraction methods that require clean speech from the target speaker as it only needs to collect EGG signals when recording speech, which simplifies the speech extraction procedure. As seen in the waveforms of EGG signals and speech signals, sound segments and silent segments are both relevant. In addition, to utilize the time domain features of EGG signals, we proposed a method to process mixed signals using information provided by the EGG signals.

3. Materials and Methods

3.1. Materials

The CDESD that was mainly used in our work was collected by Jing [46]. Each record in this database consists of one channel of speech signals and one channel of EGG signals. Moreover, there is a temporal correlation between the EGG waveforms and the speech waveforms, and the voiced (V), silent (S), and unvoiced (U) segments [47] of speech correspond to different parts of the EGG signals, which can easily be discriminated. The speech samples in the database were recorded using high-standard equipment and contain different genders and different speech and EGG signals. The database contains 20 individual speakers, including 7 females and 13 males, expressing different sentences and emotions. The female speakers are labeled as F11, F12, F15, F16, F17, F18 and F20, and the male speakers are labeled as M01, M02, M03, M04, M05, M06, M07, M08, M09, M10, M13, M14 and M19. The annotations for the speech samples reaches a high level with 11,363 documents in all. Table 1 illustrates the gender and length distributions of the records in the CDESD.

Table 1. The gender and length distributions of the records in the CDESD.

Gender Distribution		**Length Distribution**		
Male	Female	<1 s	≥1 s, <2 s	≥2 s
743	3926	1550	7839	1974

Since EGG signals and speech signals are preserved in the same file, before our EGG model could be trained, the speech and EGG signals needed to be separated from the channels. The whole experiment was split into four parts and in each part, we chose different combinations of genders as the target and interference speakers and compared the SDR, signal-to-distortion ratio improvement (SDRi), scale invariant signal-to-distortion ratio (SISDR) and SISDRi to evaluate the speech extraction performance in each group. SISDR is usually taken as the objective function, which can further improve speech extraction performance. SISDR and SDR are defined as follows:

$$\begin{cases} X_T = \frac{X^* \cdot \hat{X}}{\|\hat{X}\|^2} \\ X_E = X^* - X_T \\ SDR = 10log_{10}\frac{\|\hat{X}\|}{\|\hat{X}-X^*\|} \\ SISDR = 10log_{10}\frac{\|kX_T\|^2}{\|kX_E\|^2} \end{cases} \tag{2}$$

where X^* is the output speech of the network and $\hat{X}$ is the original signal. SISDR illustrates whether the extraction results meet expectations. The SISDRi and SDRi calculation is as follows:

$$\begin{cases} SDR_1 = 10log_{10}\frac{\|\hat{X}_1\|}{\|\hat{X}_1-X_1^*\|} \\ SDR_2 = 110log_{10}\frac{\|\hat{X}_2\|}{\|\hat{X}_2-X_2^*\|} \\ SDR_i = SDR_2 - SDR_1 \\ SISDR_1 = 10log_{10}\frac{\|kX_{T1}\|^2}{\|kX_{E1}\|^2} \\ SISDR_2 = 10log_{10}\frac{\|kX_{T2}\|^2}{\|kX_{E2}\|^2} \\ SISDR_i = SISDR_2 - SISDR_1 \end{cases} \tag{3}$$

where $SISDR_1$ represents the $SISDR$ between mixed speech and true speech, and $SISDR_2$ represents the $SISDR$ between generated speech and true speech.

As is shown in Figure 1, Equations (2) and (3), using SDR as a measurement can cause problems. If the volume of X^* is increased properly, X_E decreases (as shown by

the dotted line), and the SDR increases, which means that the level of volume affects the SDR. Using SISDR as a measurement can avoid these problems, for X_E and X_T increase simultaneously when X^* increase. If SDR is chosen as the single indicator, numerical indicators can improve when the actual effects deteriorate. The SISDRi measurement uses the square of the projection modulus of the estimated value vector in the direction of the truth vector, instead of the square of the projection modulus of the estimated value vector in the perpendicular direction to the truth vector, in order to avoid the SDR problem. SISDRi is defined as the difference between the estimated mixed signal SISDR and the estimated true value signal SISDR. Compared to the common SISDR measurement, SISDRi can better measure optimization effects.

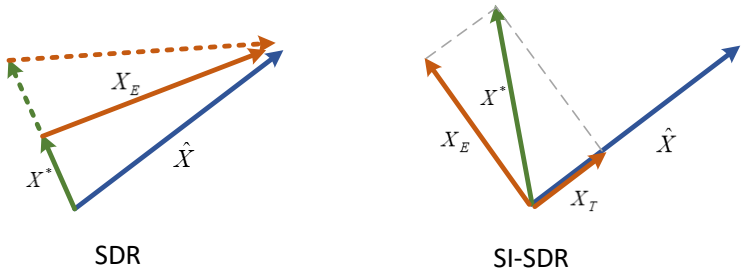


Figure 1. Diagrammatic sketches of SDR and SISDR, in which speech signals are indicated by vectors.

3.2. Methods

The methods utilized in our study relied on a basic SpeakerBeam network. Considering the difference between EGG signals and speech signals, we proposed a network that was suitable for EGG feature extraction. In addition, we took the time-domain characteristics of EGG signals into consideration, according to correlaitions between EGG signals and speech signals in silent and unvoiced segments, by performing pre-processing before mixed speech was input into the network.

3.2.1. The Configuration of SpeakerBeam

Our proposed models to deal with speech extraction problems were inspired by the SpeakerBeam network. The SpeakerBeam network is a combined network with multiple spectrogram amplitude inputs and spectrogram amplitude outputs. According to previous studies, the network can be divided into two parts: an auxiliary network that inputs the speech spectrograms of the target speaker's additional information and extracts the target speaker's features; and a main network that inputs mixed speech spectrograms and the target speaker's features and outputs estimated time-frequency masks. Finally, ISTFTs are performed using the enhanced amplitude spectrograms and the original phase spectrograms. Figure 2 shows the specific configuration of the SpeakerBeam network.

The method used to extract features from the frequency domain of the target speaker's speech has worked well in previous studies. In this method, speech signals are transformed from the time domain into the two-dimensional frequency domain via STFTs.

In our study, the vertical coordinate of a spectrogram was the feature dimension, which represented the feature vector of one frame. The horizontal coordinate was the frame dimension, and the time-domain sequence length divided by the window interval was rounded up to 173 columns. The length of the different speech samples was set to 1 s. The spectrograms were imported into the network after taking the modulus value. The estimated amplitude spectrograms of target speech signals were obtained by multiplying the amplitude spectrograms of mixed speech by the estimated mask values. The process was defined as follows:

$$M = g(|Y|, |A|) \tag{4}$$

$$\hat{S} = M \odot Y \tag{5}$$

where g stands for neural network's estimation of the mask of the STFT signal M, Y and A represent the mixed speech and the extra information from the target speaker, respectively. $|\cdot|$ is the magnitude of the signals, and $\odot$ represents matrix multiplication of the mixed speech and binary masks. The horizontal coordinates of the spectrograms represented time, the vertical coordinates represented frequency, and the values of the coordinate points represented voice energy. In Figure 3, the deeper the color, the stronger the voice energy of the coordinate point.

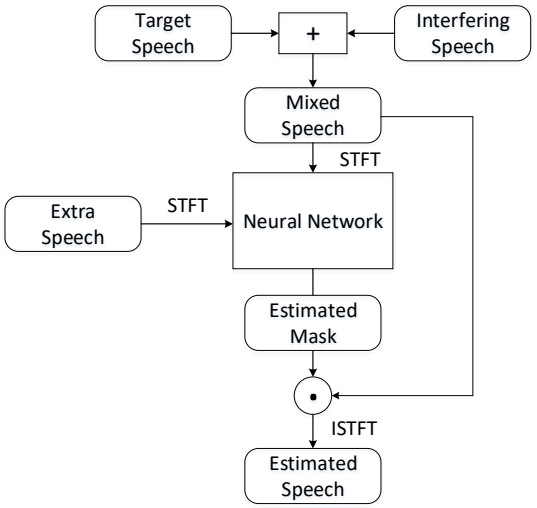


Figure 2. The overall scheme of the SpeakerBeam network used to separate target speech from mixed speech.

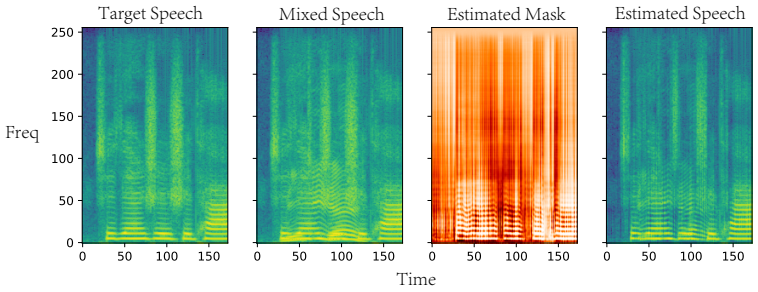


Figure 3. The spectrograms of the target speech, the mixed speech, mask and the estimated speech. The energy value is represented by the color. The deeper the color, the stronger the voice energy.

After the STFT transformation, the amplitudes of the different frequencies of speech were converted into logarithmic scales and the color sizes (amplitudes) were converted into decibels to form spectrograms. Then, the estimated speech was obtained by multiplying the mixed speech by the estimated masks.

Neural networks that use additional information from target speakers are common modeling methods for solving acoustic problems. These networks can be expressed as follows:

$$X_{k+1} = \sigma_k(L_k(X_k, W, b)) \tag{6}$$

where X_k is the input of the layer k, σ_k is the activation function, and $L_k(X_k, W, b) = WX + b$, where W is the weight and b is the bias vector of the layer. These networks can extract features from target speakers but may lack specificity for different features. The process of these methods is as follows:

$$X_{k+1} = \sigma_k(\lambda \odot L_k(X_k, W, b)) \tag{7}$$

where λ is the weight vector of the features, which is determined by the target to be extracted by the network. SpeakerBeam [48] utilizes a sequence summarizing network with an attention mechanism. The attention mechanism can learn from different frames of speech and summarize the frequency information of a target speaker's speech. In a previous study, an attention mechanism was used to obtain target areas that need to be focused on and quickly screen out high-value information from large amounts of data. The value of attention is obtained as follows:

$$attention(X, q) = \sum_{i=1}^{N} \alpha_i x_i \tag{8}$$

where α is the outcome of the Softmax layer, which contributes to the weight of each feature, and x_i represents the frames of the spectrogram matrix. Attention mechanism [49,50] can estimate binary mask for each frame in a spectrogram and improve the SI-SNR of mixed speech extraction.

3.2.2. The Processing of the Datasets

Before generating our model for target speaker extraction, all signal amplitudes had to be normalized to the [−0.5,0.5] interval on a linear scale and the length of each signal was extended or shortened to 1 s at a sample rate of 22,050 Hz. The target and interference choice is to be made either according to gender or randomly. In this study, we created a subset of the database that was based on combinations of the genders of the target speaker and the interference speaker to generate mixed signals. Figure 4 demonstrates our procedure for generating the subsets.

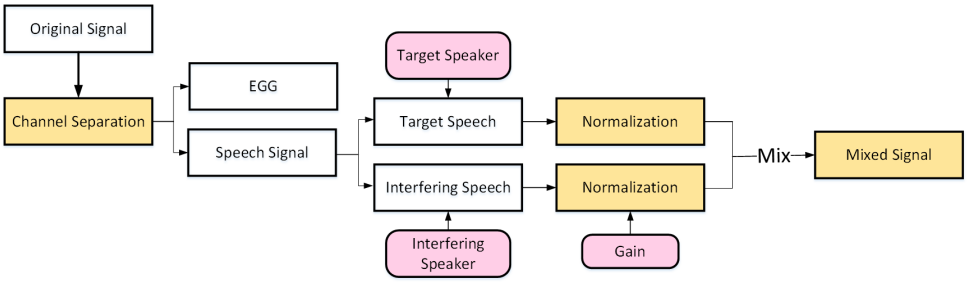


Figure 4. The process of EGG and speech signal extraction on the CDESD including the separation and the mixing procedures. The original signals consist of both speech signals and EGG signals.

In the training stage of the experiment, we selected target speakers and interference speakers randomly from the male and female samples in the CDESD. The mixing model was defined as follows:

$$Y^{(m)}[n] = s_0^{(m)}[n] + G * v^{(m)}[n] \tag{9}$$

where Y is the mixed signal, s_0 is the target speech, G is the gain of the interference speech, v is the interference speech, and m is the index of discrete time. In our study, the following combinations of target-interference speakers were generated: female-female (FF), female-male (FM), male-female (MF), and male-male (MM).

In the FF and MM groups, we selected seven speakers as the target speaker and subsequent samples in the database were selected as the interference speakers. In the FM and MF groups, we selected one female or male target speaker and an interference speaker of the other gender. In each group, 100 target speakers and 100 interference speakers were randomly chosen from the CDESD to generate seven sets of mixed speech, which contained 70,000 ($100 \times 100 \times 7$) pieces of data in total. Out of all the samples in the database, 90% (63,000) were chosen to be training data while 10% were chosen to be validation data. In terms of the extra speech for feature extraction, considering every speaker in the database has 20 sentences that were recorded with different emotions, we chose the first three sentences to ensure that the content of the validation set was different from the content of

the test set. When mixing the target speech and interference speech, we set the ratio of the target speaker to interference speaker to obtain different SNRs. In most of our experiments, the SNR was set to 2.5 dB.

3.2.3. Model For Electroglottograph Speech Extraction

In our experiments, the extra speech was replaced by EGG signals. As a result, the auxiliary network for extracting the features of target speakers needed to be improved. As EGGs come from vibrations in the throat, only the fundamental frequencies of the EGG waveforms carry information about speech. However, the inputs for our auxiliary network were speech spectrograms, which were two-dimensional data. So, the EGGs had to be transformed into two-dimensional data. In our study, we generated spectrograms of the EGG signals, which encoded information about the glottal pulses.

As Figure 5 indicates, the spectrograms of the recorded EGG signals contained extra information such as signals from the throat, which can provide more information about the identity of the speaker. Based on previous studies on EGGs and speech extraction, we eventually developed a speech extraction model that was fit for EGG signals.

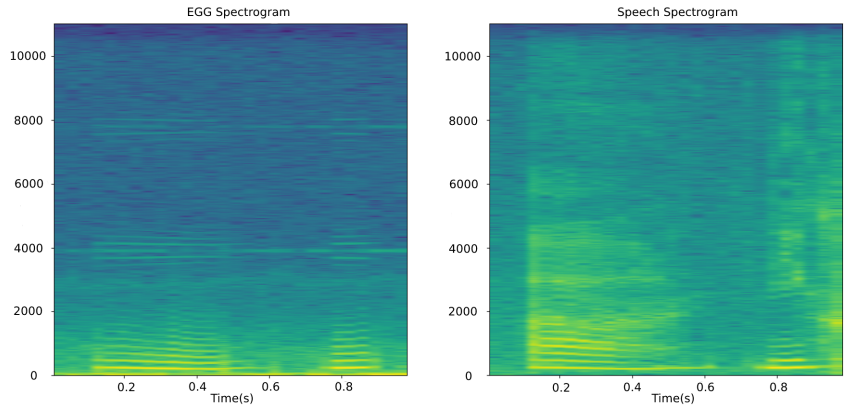


Figure 5. A comparison between the spectrograms of EGG signals and speech signals from the same utterance.

Based on the features of the EGG signals, we designed a speech extraction neural network, which is shown in Figure 6. The main network accepted the spectrograms of mixed speech as inputs while the auxiliary network accepted the spectrograms of EGG signals as inputs. In the main network, Tanh layers were used after each linear layer for nonlinearity. The spectrograms of target speakers were obtained by multiplying the binary masks and the spectrograms of mixed speech. The MSE loss between the estimated target speaker speech spectrogram and the original target speaker speech spectrogram was taken as the cost function. The LSTM layers in the bottom half of the network were removed because of the gradient disappearance that occurs in the training process of multi-layer convolutional networks. Our model used a logistic sigmoid to activate the output layer. The auxiliary network for extracting the characteristics of target speaker EGGs consisted of two convolution layers, one ReLU activation layer, and a Softmax layer. This new network reduced computational complexity and produced a better speech extraction performance by using EGGs.

3.2.4. The Electroglottograph Pre_Processing Algorithm

The original EGG signals had baseline drift interference that was caused by irrelevant vibrations. To obtain EGG signals with less interference, the EGG signals were filtered in the time domain and frequency domain. High-pass filtering was performed first. A high-pass filter was used to remove interference from the power supply.

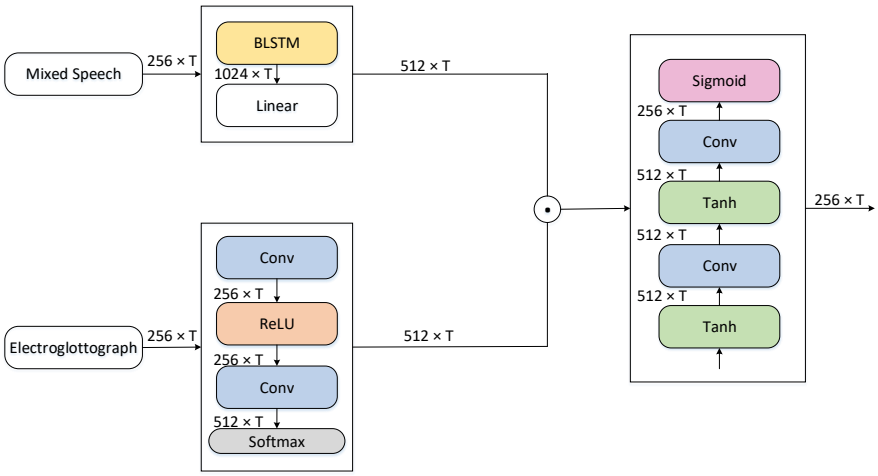


Figure 6. The configuration of our EGG_auxiliary network, which utilized a Softmax layer to extract the features of a target speaker via an attention mechanism. The mixed speech and EGG signals were processed simultaneously.

The frequency-domain features of EGG signal were also utilized in the neural network. However, the EGG signals were input after undergoing STFTs, which could negelect time-domain features. As a result, considering the relevance of EGG signals and speech signals in the time domain, we used the signal envelopes [51] from the Hilbert transforms to identify the speech of the target speaker and extract it from the EGG signals. The process is shown in Figure 7.

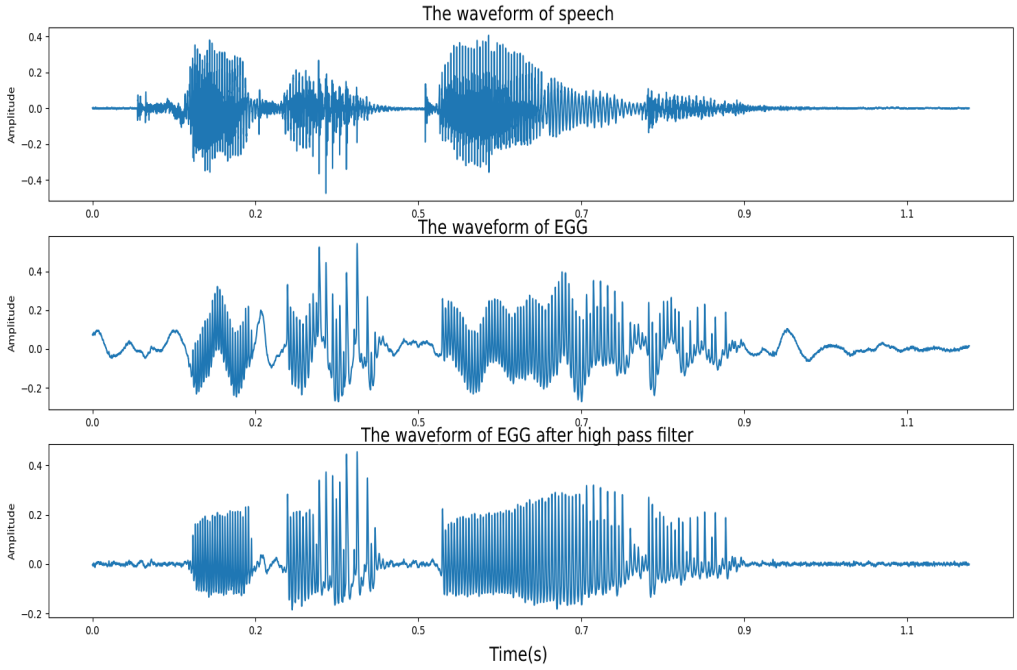


Figure 7. An example of a waveform in CDESD. The top panel shows the speech of the speaker, the middle panel shows the EGG signals, and the bottom panel shows the EGG signals after high-pass filtering.

Since the EGG signals were recorded at the same time as the speech signals, they were similar in some parts of the time domain. In silent segments of speech, there were no vibrations in the throat, so the EGGs showed silent segments. The same occurred in other segments as well and the trend of the waveforms was consistent. As a result, there

were correlations between the EGG signals and the speech signals in some of the speech segments, which meant the time-domain features of EGG signals could be used to solve the speech extraction problems. So, we also developed a pre_processing algorithm to deal with the correlations between the EGG and speech signals and applied it to the speech extraction task. The method filtered the silent segments and processed other relevant parts of speech. The average value of the filtered envelope signals was taken as the threshold, with values greater than the threshold considered to be 1 and those less than the threshold were considered to be 0.5, according to the window function method. Finally, we obtained the target speaker's speech. The division method was accurate for voiced segments involving vowels but could lose information from the original speech waveforms of unvoiced and silent sounds.

As shown in Figure 8, the envelopes of the EGG signals were calculated to process the mixed speech and the new mixed speech was then preliminarily processed (the silent segments from the target speech were filtered out and some of the unvoiced segments were processed before being input into the network).

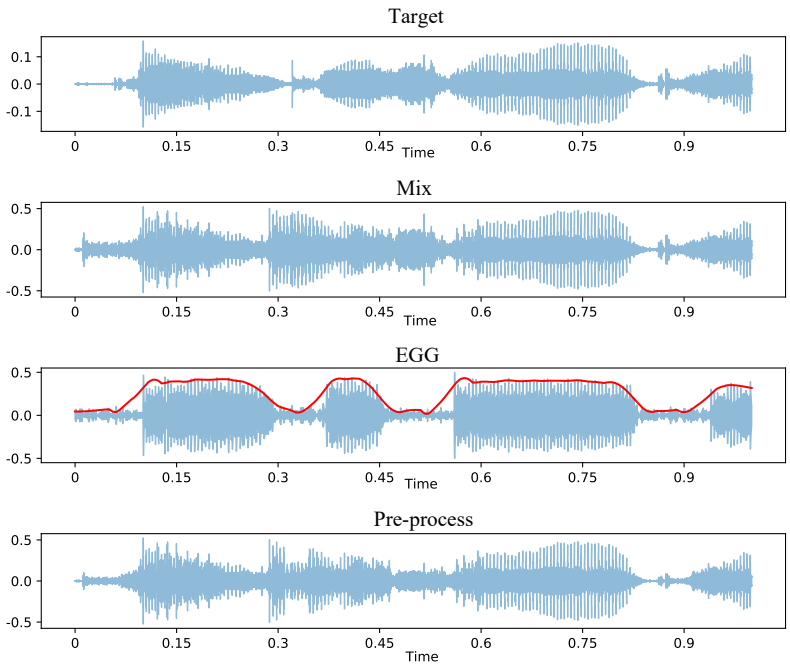


Figure 8. The waveform of one of the signals in the pre_processing algorithm.

4. Experiments and Results

This section details our experiments,in which we compared the performances of different inputs for our neural network (i.e., extra target speech and EGG signals). The experiments were evaluated using the SDR and SISDR metrics, as defined in Section 3. In addition, the Adam [52] optimizer was used with a learning rate of 1×10^{-4} . The training batch size was set as 256 with 100 epochs.

4.1. Experiments

4.1.1. Datasets

The experiments were performed on two datasets: the CDESD and the EMO-DB, which was recorded in German. Table 2 shows our comparison results and the details of each dataset, including the number of male and female speakers and the number of speech data points in the training and test sets.

Table 2. The total number of speakers and the numbers of target speaker and mixed speech samples.

	Training Sets			Test Sets		
	Speakers	**Mixtures**	**Target**	**Speakers**	**Mixtures**	**Target**
CDESD	20	63,000	9000	20	7000	1000
EMO-DB	10	16,200	3240	10	1800	360

The first dataset (CDESD) contains recordings in two channels. In total, 100 samples from each speaker were selected as target speech and interference speech. The EGG signals in the datasets and three sentences from each speaker were then used for feature extraction. The datasets have 7 males and 13 female speakers, which we combined into seven groups for training and testing. The second dataset (EMO-DB) consists from 10 speakers, including the speech signals and EGG signals of each speaker saying the same sentences in German. We mixed the speech samples from these datasets to generate SNRs between -5 and 5 dB. For both datasets, we selected the samples randomly to form mixtures of target speech and interference speech. We used a 22,050 Hz sampling frequency for all of our experiments.

4.1.2. Experimental Settings

Based on the SpeakerBeam network, we proposed EGG_Aux and Pre_EGG_Aux networks. We set the batch size to 256, with a learning rate of 1×10^{-4}. We used sequence summarization with an attention mechanism for the extraction of target speaker features to inform the networks about the target speaker. Due to the attention mechanism, the networks could focus on the more useful information provided by the EGG signals and extract target speech more accurately.

The architecture of the networks consisted of bidirectional long short-term memory (BLSTM) layers and convolution layers. The loss function was MSE loss between the amplitudes of target speech and estimated signals. Additionally, although SDRi has been selected as the metric to compare performance in previous studies, in our work, we considered SISDRi metric in order to compare the outcomes more reasonably.

4.2. Results

4.2.1. Comparison to SpeakerBeam

To evaluate the effectiveness of using EGG signals to solve speech extraction tasks, we compared the performances of FD-SpeakerBeam and SpeakerBeam based on EGG signals. For the SpeakerBeam experiments, we provided the network with extra information about the target speaker, including the extra sentences from the datasets. For the EGG_Aux network experiments, we input EGG signals into the network. In addition, we utilized the time-domain characteristics of the EGG signals to process the silent segments from the mixed speech before the Pre_EGG_Aux experiments. Therefore, both frequency domain and time domain features were utilized in this work. An example of the estimated waveforms is shown in Figure 9.

For a fair comparison between the different speech extraction methods, the same amount of extra information was used from both datasets and we used the SDRi and SISDRi metric to compare the results. Table 3 shows the SDRi and SISDRi values from the different experiments.

Table 3. A comparison between the SDRi and SISDRi on the CDESD for SpeakerBeam and the two methods proposed in our paper. The parameters are the average of all test results.

Model	**Training Dataset**	**SDRi (dB)**	**SISDRi (dB)**
SpeakerBeam	CDESD	3.43	4.39
EGG_Aux	CDESD	4.55	5.25
Pre_EGG_Aux	CDESD	4.58	5.28

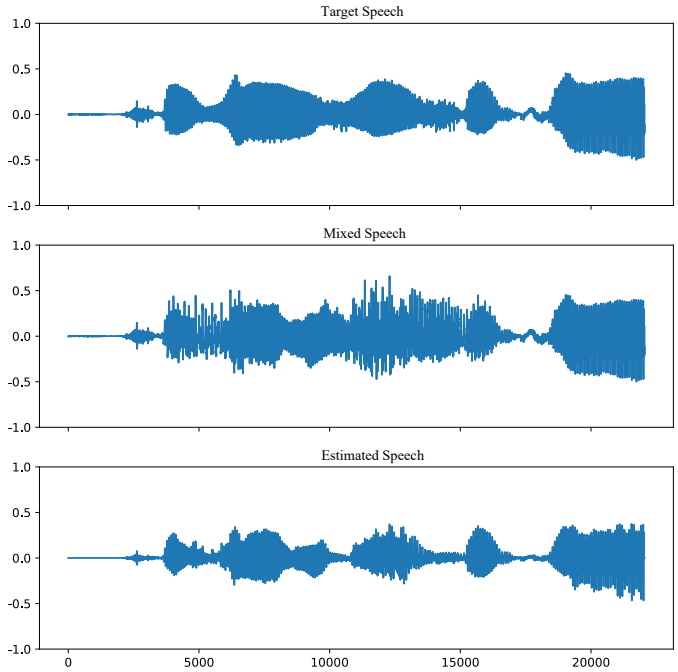


Figure 9. An example of estimated target speech.

The results showed that in the field of deep learning, EGG signals could provide more information and a better target speech extraction performance than common speech signals, EGG signals could also be obtained in noisy environments, which offers extensive prospects. The first set of experiments compared the performances of the considered methods on the CDESD and EMO-DB datasets, which were originally created for emotion recognition. The first part of Table 3 shows the results for the SpeakerBeam network, which achieved an SISDRi of 4.39dB for mixtures of two voices on CDESD. The second part of the table shows that the EGG signals achieved a better performance than normal speech signals for the SpeakerBeam network. The SDRi was 4.58 dB and the SISDRi was 5.28 dB, which were 33% and 20% higher in the validation sets, respectively, showing that the extracted frequency-domain features of the EGG signals provided more information for the network to solve the speech extraction problem. The experiment also utilized the time-domain features of the EGG signals, because previous experiments have only focused on the frequency domain and have overlooked the time-domain features of EGG signals, which are consistent with normal speech signals in silent and unvoiced segments. The results showed that the time-domain features of the EGG signals improved the performance of the networks at a low level. The reason for the this could be that time-domain features were taken into consideration in the STFTs. Moreover, the algorithm that filtered out the silent segments during prepocessing could have filtered out some of the unvoiced or voiced segments by mistake. Despite the slight improvement in the SDRi and SISDRi metrics, the results showed that the pre_processing of EGG signals could help to solve the target speech extraction problem.

To verify this improvement in solving speech extraction task using EGG signals, we carried out further experiments on EMO-DB dataset, which was recorded in German and has 2 s speech samples compared to the 1 s speech samples in the CDESD. Table 4 shows the SDRi and SISDRi results for the two considerd methods on the EMO-DB dataset.

Table 4 shows that the EGG signals improved the speech extraction performance on both the CDESD and EMO-DB dataset, meaning that EGGs could help in the extraction of speech in different languages.

Table 4. A comparison between the SDRi and SISDRi on the EMO-DB dataset. The parameters are the average of all test results.

Model	Training Dataset	SDRi (dB)	SISDRi (dB)
SpeakerBeam	EMODB	2.28	0.99
Pre_EGG_Aux	EMODB	3.69	2.71

4.2.2. Comparison between Genders

Considering the shortcomings in the same-gender speech extraction performance of the SpeakerBeam network, we compared speech extraction performance when the target and interference were of different genders. Table 5 shows the results which were calculated using the average values for different speakers of the same gender.

Table 5. The SISDRi (dB) on the CDESD. The results are divided into four groups according to the gender of the target speaker and the gender of the interference speaker. The left letter represents the gender of the target and the right letter represents the gender of the interference speaker.

Model	FF	MM	FM	MF
SpeakerBeam	3.12	2.46	3.55	4.58
Pre_EGG_Aux	5.12	3.46	5.37	4.37

Table 5 shows the SDRi when the target speakers and interference speakers were of different genders. For the SpeakerBeam network, the speaker extraction was better when the speakers were of different genders but the same-gender speech extraction performance was poor. In our experiment, the SpeakerBeam network with EGG signals performed better when dealing with same-gender speech extraction. The SDRi for female-female(FF) speech extraction was 5.12 dB and the SDRi of male-male(MM) speech extraction was 3.46 dB improvement, which meant that the differences between the EGG signals were more significant than the differences between speech signals when the speakers were of the same gender. When the speakers were of different genders, the female speech extraction performance was better. In conclusion, EGG signals could better identify the features of individual speakers in most situations.

In Figure 10, it can be seen that the speech extraction performance was better when dealing with female speakers and that the speech extraction performance tended to be poor when dealing with male speakers. The different throat construction of males and females could be the main cause of this phenomenon. Additionally, the EGGs of the female speakers were more suitable for speech extraction.

Figure 11 shows the results of the speech extraction experiments involving different target speakers. The SDRi was better for female speakers than male speakers. The results showed that EGG signals were more useful for female speech extraction.

Figure 12 shows the SDRi and SISDRi under different SNRs using SpeakerBeam and Pre_EGG_Aux network.

4.2.3. Comparisons between Different SNRs

In our previous experiments, we set the SNR a 2.5 dB, which meant that the amplitude spectrogram of interference speech was 0.75 times that of the target speech. To evaluate the speech extraction performance under different circumstances, we set SNRs from -5 dB to 5 dB to test the validation sets and compare speech extraction performance.

From Table 6, it can be seen that the improvements increased when the noise increased, meaning that the EGG signals performed better for speech extraction in noisy environments. The SDRi and SISDRi values decreased by about 1.5 dB while the SNR value increased by 2.5 dB, which showed that speech extraction performance was better in environments with low SNRs.

Table 6. The SDRi and SISDRi under different SNRs on the CDESD using our Pre_EGG_Aux network.

SNR (dB)	−5	−2.5	0	2.5	5
SDR (Mixed Speech)	−3.99	−1.41	1.07	3.10	6.11
SDR (Prediction)	5.98	6.74	6.95	8.87	11.35
SISDR (Mixed Speech)	−4.00	−1.41	1.07	4.08	6.11
SISDR (Prediction)	4.70	6.74	7.35	9.45	11.08
SDRi	9.97	8.15	6.95	5.77	5.25
SISDRi	8.70	7.15	6.29	5.37	4.97

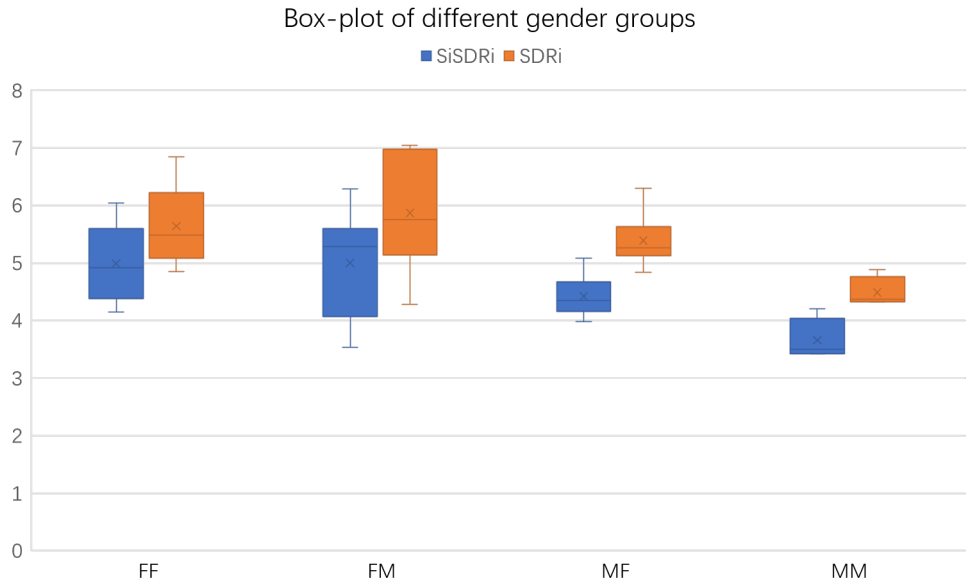


Figure 10. A box-plot of the SDRi and SISDRi results for different gender groups using the pre_EGG_Aux method.

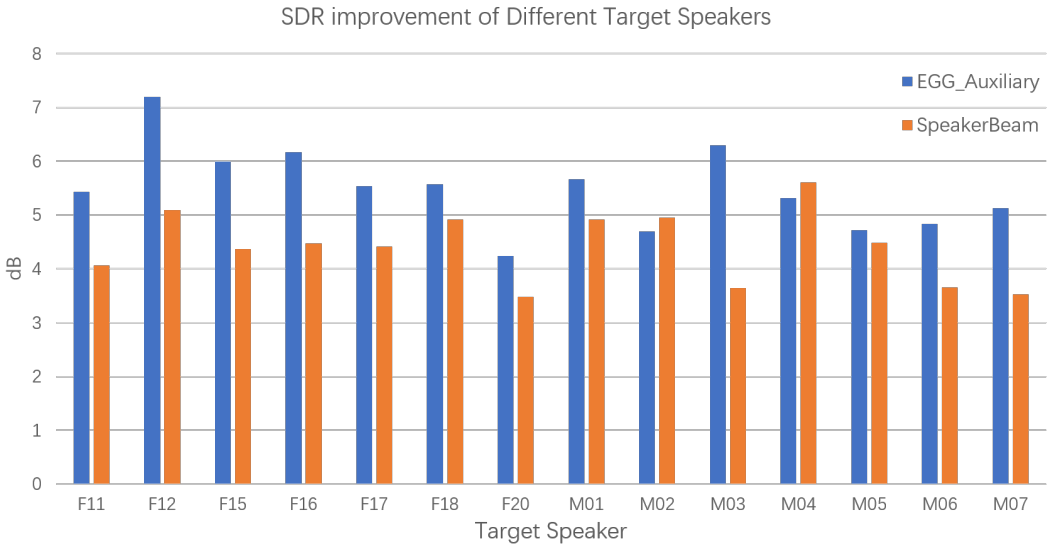
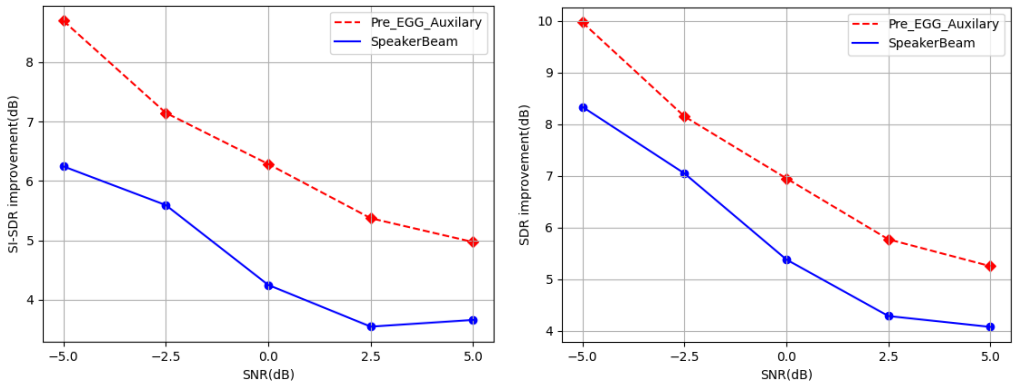


Figure 11. A column chart of the SDRi for different target speakers using the two methods.



SDRi SISDRi

Figure 12. The SDRi and SISDRi under different SNRs on the CDESD using the SpeakerBeam and our Pre_EGG_Aux network.

The figures above show the speech extraction results under different SNRs for target speakers and interference speakers of different genders. It can be seen that the Pre_EGG_Aux algorithm performed better under the different SNR circumstances, meaning that our speech extraction method for performed better in different environments.

5. Discussion

In Section 4, we detailed the series of speech extraction experiments that we conducted to compare speech extraction performance and identify the best method. To explore the effects of information extraction using different signals, we compared the SpeakerBeam network and our EGG_Aux and Pre_EGG_Aux networks. The results showed that on the CDESD, the SDRi increased by 1.12 dB while the SISDRi increased by 0.86 dB with the EGG_Aux network. When using the Pre_EGG_Aux network, the SDRi increased by 1.15 dB while the SISDRi increased by 0.89 dB. From these results, we could infer that the EGG signals provide more information than speech signals and could extract more information from time-domain features. In addition, we tested our networks on the EMO-DB. These results showed that using the Pre_EGG_Aux network increased SDRi by 1.41 dB and the SISDRi by 1.72 dB, meaning that our proposed method had a better speech extraction performance for different languages.

As for different speech extraction circumstances, we conducted experiments involving different genders. The EGG method achieved a better performance in most situations, especially when the target speaker and interference speaker were of the same gender. In the female-female situation, the network using EGGs achieved a 2 dB increase, while a 1 dB increase was achieved in the male-male situation. In terms of speech extraction performance involving speakers of the same gender, the Pre_EGG_Aux network achieved a similar level to that involving different genders. As shown above, when dealing with female speech extraction, the network using EGG signals achieved better results, which meant that the EGG signals from female speaker were easier to recognize than those of male speakers.

To verify the performance of our model under different SNRs, we mixed samples from the datasets using different amplitude spectrograms and compared the results for SNRs ranging from -5 dB to 5 dB. As shown in the Results Section, the SISDRi was 8.70 dB when the SNR was set as -5 dB, while the SISDRi was 4.97 dB when the SNR was set as 5 dB. These results suggested that our model had a better speech extraction performance in noisier environments.

6. Conclusions

In this paper, we proposed a new method for target speech extraction. We input the EGG signals of target speakers and extracted target speech from noisy environments.

As EGGs are not susceptible to noise, we could solve the speech extraction problem in situations with significant noise. When using the EGG_Aux network, the speech extraction performance increased by 32.7% in terms of SDRi, while the SISDRi value improved by 19.6%. When using the Pre_EGG_Aux network, the speech extraction performance increased by 33.5% in terms of SDRi, while the SISDRi value improved by 20.3%. Moreover, the gap between speech extraction performance when dealing with problems involving different genders was reduced by using EGGs. The gap between the speech extraction performance for the same gender and different genders was initially 1.28 dB but decreased to 0.58 dB by using the Pre_EGG_Aux network, showing that EGGs helped the network distinguish target speakers more accurately.

In future work, we aim to design new speech extraction models that can solve problems in more complex situations. Additionally, we hope to also propose other methods to integrate more information from target speakers and achieve better results.

Author Contributions: Conceptualization, L.C.; methodology, Z.M.; investigation, Z.M. and J.R.; resources, L.C.; writing—original draft preparation, Z.M.; project administration, L.C.; funding acquisition, Q.Z. and C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 62072021), the National Science Foundation for Young Scholars of China (Grant No. 61603013), and the Fundamental Research Funds for the Central Universities (Grant No. YWF-22-L-532 and No. YWF-22-T-204).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We gratefully acknowledge all subjects involved in building the dataset for our study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Haykin, S.; Chen, Z. The Cocktail Party Problem. *Neural Comput.* **2005**, *17*, 1875–1902.https://doi.org/10.1162/0899766054322964.
2. Brown, J.A.; Bidelman, G.M. Familiarity of Background Music Modulates the Cortical Tracking of Target Speech at the & ldquo; Cocktail Party & rdquo;. *Brain Sci.* **2022**, *12*, 1320. https://doi.org/10.3390/brainsci12101320.
3. Christian, Y.; Darmawan, I. Rindik rod sound separation with spectral subtraction method. *J. Phys. Conf. Ser.* **2021**, *1810*, 012018.
4. Amarjouf, M.; Bahja, F.; Martino, J.D.; Chami, M.; Elhaj, E.H.I. Denoising Esophageal Speech using Combination of Complex and Discrete Wavelet Transform with Wiener filter and Time Dilated Fourier Cepstra. In Proceedings of the 4th International Conference on Computing and Wireless Communication Systems (ICCWCS 2022), Tangier, Morocco, 21–23 June 2022.
5. Luo, Y. A Time-domain Generalized Wiener Filter for Multi-channel Speech Separation. *arXiv* **2021**, arXiv:2112.03533.
6. Roux, J.L.; Hershey, J.R.; Weninger, F. Deep NMF for speech separation. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015.
7. Wisdom, S.; Powers, T.; Pitton, J.; Atlas, L. Deep recurrent NMF for speech separation by unfolding iterative thresholding. In Proceedings of the 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 15–18 October 2017.
8. Kłosowski, P. A Rule-Based Grapheme-to-Phoneme Conversion System. *Appl. Sci.* **2022**, *12*, 2758. https://doi.org/10.3390/app12052758.
9. Brown, G.J.; Cooke, M. Computational auditory scene analysis. *Comput. Speech Lang.* **1994**, *8*, 297–336. https://doi.org/10.1006/csla.1994.1016.
10. Dana, B.; Israel, N.; Joel, S. Auditory Streaming as an Online Classification Process with Evidence Accumulation. *PLoS ONE* **2015**, *10*, e0144788.
11. Wang, D.; Brown, G. Computational Auditory Scene Analysis: Principles, Algorithms and Applications. *IEEE Trans. Neural Netw.* **2008**, *19*, 199–199.
12. Mill, R.W.; B?Hm, T.M.; Bendixen, A.; Winkler, I.; Denham, S.L.; Sporns, O. Modelling the Emergence and Dynamics of Perceptual Organisation in Auditory Streaming. *PLoS Comput. Biol.* **2013**, *9*, e1002925.
13. Cheng, S.; Shen, Y.; Wang, D. Target Speaker Extraction by Fusing Voiceprint Features. *Appl. Sci.* **2022**, *12*, 8152. https://doi.org/10.3390/app12168152.

14. Higuchi, T.; Ito, N.; Yoshioka, T.; Nakatani, T. Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016.
15. Buchner.; H..; Aichner.; R..; Kellermann.; W.. A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. *IEEE Trans. Speech Audio Process.* **2005**, *13*, 120–134.
16. Vincent, E.; Barker, J.; Watanabe, S.; Roux, J.L.; Nesta, F.; Matassoni, M. The second 'CHiME' Speech Separation and Recognition Challenge: Datasets, tasks and baselines. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013.
17. Al-Barhan, H.A.; Elyass, S.M.; Saeed, T.R.; Hatem, G.M.; Ziboon, H.T. Modified Speech Separation Deep Learning Network Based on Hamming window. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1076*, 012059.
18. Nandal, P. Speech Separation Using Deep Learning; Sustainable Communication Networks and Application. In Proceedings of the International Conference on Security and Communication Networks (ICSCN), Erode, India, 6–7 August 2020.
19. Liu, C.; Inoue, N.; Shinoda, K. Joint training of speaker separation and speech recognit ion based on deep learning. In Proceedings of the ASJ 2017 Autumn Meeting, Tokyo, Japan, 25 September 2017.
20. Elminshawi, M.; Mack, W.; Chakrabarty, S.; Habets, E. New Insights on Target Speaker Extraction. *arXiv* **2022**, arXiv:2202.00733.
21. Ji, X.; Yu, M.; Zhang, C.; Su, D.; Yu, D. Speaker-Aware Target Speaker Enhancement by Jointly Learning with Speaker Embedding Extraction. In Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020.
22. Zhang, C.; Yu, M.; Weng, C.; Yu, D. Towards Robust Speaker Verification with Target Speaker Enhancement. *arXiv* **2021**, arXiv:2103.08781.
23. Pan, Z.; Ge, M.; Li, H. A Hybrid Continuity Loss to Reduce Over-Suppression for Time-domain Target Speaker Extraction. *arXiv* **2022**, arXiv:2203.16843.
24. Wang, F.L.; Lee, H.S.; Tsao, Y.; Wang, H.M. Disentangling the Impacts of Language and Channel Variability on Speech Separation Networks. *arXiv* **2022**, arXiv:2203.16040.
25. Hershey, J.R.; Chen, Z.; Roux, J.L.; Watanabe, S. Deep clustering: Discriminative embeddings for segmentation and separation. *arXiv* **2016**, arXiv:1508.04306.
26. Yu, D.; Kolbk, M.; Tan, Z.H.; Jensen, J. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017.
27. Yousefi, M.; Hansen, J. Single-channel speech separation using Soft-minimum Permutation Invariant Training. *arXiv* **2021**, arXiv:2111.08635.
28. Luo, Y.; Mesgarani, N. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1256–1266. https://doi.org/10.1109/TASLP.2019.2915167.
29. Li, C.; Luo, Y.; Han, C.; Li, J.; Yoshioka, T.; Zhou, T.; Delcroix, M.; Kinoshita, K.; Boeddeker, C.; Qian, Y. Dual-Path RNN for Long Recording Speech Separation. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021.
30. Wang, Q.; Sridhar, P.; Moreno, I.L.; Muckenhirn, H. Targeted voice separation by speaker conditioned on spectrogram masking. *arXiv* **2020**, arXiv:1810.04826.
31. Zmolikova, K.; Delcroix, M.; Kinoshita, K.; Higuchi, T.; Cernocky, J. Optimization of Speaker-Aware Multichannel Speech Extraction with ASR Criterion. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
32. Žmolíková, K.; Delcroix, M.; Kinoshita, K.; Ochiai, T.; Nakatani, T.; Burget, L.; Černocký, J. SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 800–814. https://doi.org/10.1109/JSTSP.2019.2922820.
33. Delcroix, M.; Zmolikova, K.; Ochiai, T.; Kinoshita, K.; Araki, S.; Nakatani, T. Compact Network for Speakerbeam Target Speaker Extraction. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6965–6969. https://doi.org/10.1109/ICASSP.2019.8683087.
34. Xiao, T.; Mo, J.; Hu, W.; Chen, D.; Wu, Q. Single-channel speech separation method based on attention mechanism. *J. Phys. Conf. Ser.* **2022**, *2216*, 012049.
35. Ochiai, T.; Delcroix, M.; Kinoshita, K.; Ogawa, A.; Nakatani, T. Multimodal SpeakerBeam: Single Channel Target Speech Extraction with Audio-Visual Speaker Clues. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019.
36. Delcroix, M.; Ochiai, T.; Zmolikova, K.; Kinoshita, K.; Tawara, N.; Nakatani, T.; Araki, S. Improving Speaker Discrimination of Target Speech Extraction With Time-Domain Speakerbeam. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 691–695. https://doi.org/10.1109/ICASSP40776.2020.9054683.
37. Baken, R.J. Electroglottography. *J. Voice* **1992**, *6*, 98–110. https://doi.org/10.1016/S0892-1997(05)80123-7.
38. Herbst, C.T. Electroglottography—An Update. *J. Voice* **2020**, *34*, 503–526. https://doi.org/10.1016/j.jvoice.2018.12.014.
39. Childers, D.; Krishnamurthy, A. A critical review of electroglottography. *Crit. Rev. Biomed. Eng.* **1985**, *12*, 131–161.

40. Chen, L.; Ren, J.; Chen, P.; Mao, X.; Zhao, Q. Limited text speech synthesis with electroglottograph based on Bi-LSTM and modified Tacotron-2. *Appl. Intell.* **2022**, *52*, 15193–15209.
41. Fourcin, A.; Abberton, E.; Miller, D.; Howells, D. Laryngograph: Speech pattern element tools for therapy, training and assessment. *Int. J. Lang. Commun. Disord.* **1995**, *30*, 101–115.
42. Bous, F.; Ardaillon, L.; Roebel, A. Semi-supervised learning of glottal pulse positions in a neural analysis-synthesis framework. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands, 18–21 January 2020.
43. Cangi, M.E.; Ylmaz, G. Test-Retest Reliability of Electroglottography Measurement. *J. Acad. Res. Med.* **2021**, *11*, 126–136.
44. Chen, P.; Chen, L.; Mao, X. Content Classification With Electroglottograph. *J. Phys. Conf. Ser.* **2020**, *1544*, 012191. https://doi.org/10.1088/1742-6596/1544/1/012191.
45. Chen, L.; Ren, J.; Mao, X.; Zhao, Q. Electroglottograph-Based Speech Emotion Recognition via Cross-Modal Distillation. *Appl. Sci.* **2022**, *12*, 4338. https://doi.org/10.3390/app12094338.
46. Jing, S.; Mao, X.; Chen, L.; Zhang, N. Annotations and consistency detection for Chinese dual-mode emotional speech database. *J. Bjing Univ. Aeronaut. Astronaut.* **2015**, *41*, 1925.
47. Atal, B. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Trans. Acoust. Speech Signal Process.* **2003**, *24*, 201–212.
48. molíková, K.; Delcroix, M.; Kinoshita, K.; Higuchi, T.; Nakatani, T. Speaker-Aware Neural Network Based Beamformer for Speaker Extraction in Speech Mixtures. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017.
49. Li, C.; Qian, Y. Deep Audio-Visual Speech Separation with Attention Mechanism. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020.
50. Fan, C.; Tao, J.; Liu, B.; Yi, J.; Wen, Z.; Liu, X. Deep Attention Fusion Feature for Speech Separation with End-to-End Post-filter Method. *arXiv* **2020**, arXiv:2003.07544.
51. Chen, L.; Mao, X.; Yan, H. Text-Independent Phoneme Segmentation Combining EGG and Speech Data. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 1029–1037.
52. Attrapadung, N.; Hamada, K.; Ikarashi, D.; Kikuchi, R.; Matsuda, T.; Mishina, I.; Morita, H.; Schuldt, J. Adam in Private: Secure and Fast Training of Deep Neural Networks with Adaptive Moment Estimation. *arXiv* **2021**, arXiv:2106.02203.