Cluster Computing (2019) 22:S8887–S8897
https://doi.org/10.1007/s10586-018-2013-6

Research on speech separation technology based on deep learning

Yan Zhou[1,2] · Heming Zhao[2] · Jie Chen[1,2] · Xinyu Pan[3]

Received: 25 December 2017 / Revised: 28 January 2018 / Accepted: 2 February 2018 / Published online: 14 February 2018

Abstract
In order to solve the problem of instability of the traditional speech separation algorithm, a kind of reverberation speech separation model based on deep learning is proposed. The problem of speech separation in reverberation environment has been studied. The auditory scene analysis is used to simulate the human auditory perception ability. According to the ideal two value mode principle, the target speech signal can be extracted. Moreover, the deep neural network (DNN) shows great learning ability in speech recognition and artificial intelligence. In this paper, a DNN model is proposed to learn the inverse reverberation and denoising by learning the spectrum mapping between "contaminated" speech and pure speech. By extracting a series of spectrum features, the time dynamic information of adjacent frames is fused. The DNN is used to transform the coded spectrum, and restore the pure voice frequency spectrum. Finally, the time domain signal is reconstructed. In addition, the feature classification ability of DNN is also proposed to complete the separation of double sound reverberation speech. The binaural features ITD and ILD and the mono features GFCC are fused to form a long eigenvector. The DNN is pre-trained by RBM to complete the classification task. The results show that the proposed model improves the quality and intelligibility of the speech separation, and enhances the stability of the system significantly.

Keywords Auditory scene analysis · Speech separation · Spectrum feature · Deep learning

1 Introduction

The most direct way of human communication is language expression. Reverb exists in every aspect of our daily life. However, reverberation is also a problem that cannot be ignored. Due to the diversity of the environment, especially in relatively closed indoor spaces such as concert halls, classrooms and conference halls, when the distance between the microphone and the speaker is far, the voice signal received by the microphone contains more obvious reverb [1]. The reverberation reduces the clarity of the speaker's voice, and even severely affects the clarity and intelligibility of the voice quality.

In recent years, digital signal processing technology has been rapidly developing. It is widely used in people's study, work and life. With the increasing demand for the quality of voice communication, the demand for communication equipment has been gradually improved [2–4]. Reverberation not only affects people's hearing, especially in patients with hearing impairment, but also cannot neglect the performance of communication equipment such as communication and automatic speech recognition system. In severe cases, the system may be paralyzed. Therefore, to suppress or eliminate the influence of reverberation, it is necessary to separate the target speech [5].

Reverb Speech separation is the extraction of the target signal from the room reverb, which is similar to the Cocktail-Party problem, suppressing, reducing the noise interference, selecting and extracting the sound signal of interest. With the development of DSP technology, voice separation technology is widely used in various fields such as speech recognition, speaker recognition, wireless videoconferencing, cochlear implant, hearing aid research and sound source localization [6]. The research of speech separation technology is a hot topic and a difficult topic,

✉ Yan Zhou
zhyan_1221@hotmail.com

[1] College of Electronic and Information Engineering, Suzhou Vocational University, Suzhou, China

[2] School of Electronic and Information Engineering, Soochow University, Suzhou, China

[3] College of Electronics and Information Engineering, Suzhou Science and Technology University, Suzhou, China

which has strong theoretical significance and broad prospects for development.

This study is based on the speech depth learning of auditory scene analysis technology, and focuses on the problem of speech separation under reverberation conditions. According to human ear auditory perception and neural network learning ability, a reverberant speech separation model based on computational auditory scene analysis and deep neural network is proposed. Simulation results show that the proposed algorithms are better than similar algorithms.

2 Methods

2.1 The concept of auditory scene analysis techniques

Computational auditory scene analysis (CASA) is based on simulating human's auditory perception mechanism [7]. From the perspective of physiological and psychological hearing, the process of human perception and processing of sound is modeled. The modeling system can be used as a human ear to understand the separation of various physical sound sources from a mixed sound source signal and artificial intelligence.

The analysis of auditory scene analysis is mainly to study the process of human ear processing sound. Through the use of a computer, the effects of various organs of the human ear on sound processing are simulated, and the human ear can be used to process sound signals. Before Bregman put forward the theory of auditory scene analysis, researchers also did some work on speech separation. However, they all ignore the temporal and spatial characteristics of speech signals, and only study the characteristics of their frequency domain. In 1985, Weintraub established the first auditory system to simulate the human ear. It separates the sound signals from two sound sources of a mixed boy and a female voice, which is the embryonic form of the field of speech separation. Although there are obvious defects in previous studies, they still lay a solid theoretical foundation for the in-depth study of the follow-up CASA. In 1990, Bregman published "Auditory Scene Analysis." The time-domain characteristics and spatial characteristics of the speech signals ignored by researchers are supplemented. Furthermore, during the process of separating the target sound source, the two become important clues to the separation. Later, the researchers conducted a more in-depth study on this basis. In 1993, Cooke proposed a system of computational auditory scene analysis for speech separation in a noisy environment [8–10]. The input sound signal must first go through the auditory edge analysis. After that, it is decomposed into a certain number of time–frequency units. Based on the similarities and continuity of spectrum structure, they are merged into synchronization strings and then recombined to form the desired target auditory flow. In 2004, Guoning Hu and Deliang Wang analyzed the process of amplitude modulation (AM) of audio signals. Based on the existing mono separation, a CASA system is proposed by "slicing" and "combining" operations. It can improve the audio signal separation effect in high frequency area, namely Hu-Wang system. In 2010, Hu-Wang proposed a tandem algorithm that can effectively separate voiced speech from harmonic features. He translates the problem of voiced speech separation into a way of separating the noisy target speech. Two related steps have been added to estimate the pitch and binary mask. The Tandem algorithm can well separate the pitch and the turbid speech. With the continuous development of science and technology, the voice separation technology based on CASA has been progressed continuously. However, there are still many problems due to the constraints of technical level and other problems. This also provides a greater space for further scientific research [11]. At the same time, CASA is also closely related to more disciplinary fields (artificial intelligence, machine perception), and continues to promote its rapid development.

Bregman believes that the process of human perception of sound is a free combination of sound components. Based on this combination process, the same sound source information from the mixed sound signal is reassembled into a sound flow.

2.2 Auditory scene analysis techniques

According to Bregman's point of view, people's ability to combine sound is inherent. The ability to combine sequences of sounds is traced through their own learning and adaptability in acquired growth [12]. Through a large number of simulation experiments on human auditory system, some criteria for auditory scene analysis are obtained. It is applied to many mathematical analysis models. Two of the guidelines are introduced.

Exclusive distribution criteria: exclusive distribution of meaning is just a sensory element. Tones and timbres can only belong to one of multiple sound streams at a given moment. In short, a particular sensory element must be uniquely assigned to a sound stream.

Closure or continuous criteria: Most phenomena or things that originate in life are considered continuous. However, after entering the sensory organs of the human brain, these will be cut off due to other information effects and lose continuity. The human brain, however, is able to recombine these discrete pieces of information without affecting our overall perception.

2.3 Classification of auditory scenes

Information in the system has different flow patterns. Therefore, CASA can be divided into schema-driven and data-driven [13]. In schema-driven CASA, information is collected and processed by lower-level systems to deliver the results of the processing to higher-level systems. While receiving data, high-level systems also predict the direction of flow of sound information based on prior knowledge already stored in the brain. Therefore, as a whole, the flow of information is bi-directional. This flow is very close to the human ear hearing system processing, but this treatment is very complicated. So, it is difficult to implement it. In contrast, in data driven CASA, the flow of information is one-way. The data are processed step by step, and the source information is decomposed. This is very similar to the traditional process, and the engineering is relatively easy to implement.

2.3.1 Data driven CASA

According to the principle of synchronicity of auditory perception, in 1992, the M.P.Cooke of Sheffield University in England proposed a "synchronous series" time frequency representation of speech signals. He used a set of Gammatone filter sets to simulate the characteristics of human cochlear implants. In the system, the received original speech signal is frequency-divided by the Gammatone filter, and the output result is the representation of time–frequency. The corresponding calculation is carried out. In the case of meeting the corresponding threshold, the neighboring sound sources are considered to come from the same independent synchronization string, so that it is classified into one category. At the same time, it is tracked in time. In 1994, G.J.Brown and M.P.Cooke studied how to separate sound (voiced speech) signals from interfered audio signals. The time frequency regions with the same amplitude modulation characteristics are combined with the gene frequency. The data driven CASA model is shown in Fig. 1.

Data driven CASA system consists of four parts. The auditory peripheral system model consists of the outer ear model, the middle ear model, the cochlear model, and the Meddis model. After the release of the cochlear filter and inner hair cells, the sound is converted from mechanical vibration to the probability of firing of the nerve impulses. After the previous analog processing, the results will be processed according to the voice clues. Hearing calculation chart is formed. According to the second part of the information obtained, the auditory scene is analyzed. Through the resynthesis of various sound sources, it can subjectively and qualitatively analyze the processing effect of the entire CASA system.

2.3.2 Schematic-driven CASA

Differing from the one-way flow of data-driven CASA information, researchers found that the flow of information is bi-directional in the process of handling sound. It has adaptive tuning and some advanced logical reasoning capabilities. The brain will predict the information that will be received in advance. At present, many speech recognition technologies are based on linguistic knowledge to assist in word or word search. Varga and Moore use two hidden Markov models (HMMs) to separate speech and noise signals. Cooke et al. used Kohonen neuron and HMM model to identify the incomplete voice signals. However, many people also use the blackboard model to construct CASA system. The above information has bidirectional fluidity. It is classified as a schema-driven CASA. Figure 2 is a schematic driven CASA system based on a blackboard model.

There are four typical modules in a schematic-driven CASA system. The blackboard is a complete hypothetical database. All the states of the system are made up of these assumptions. They each have a clear definition and pattern. All execution modules within the system share them.

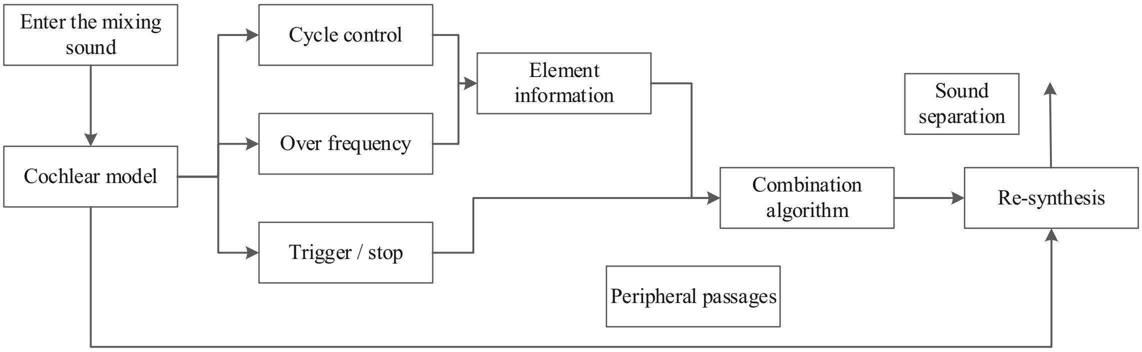


Fig. 1 Data driven CASA system

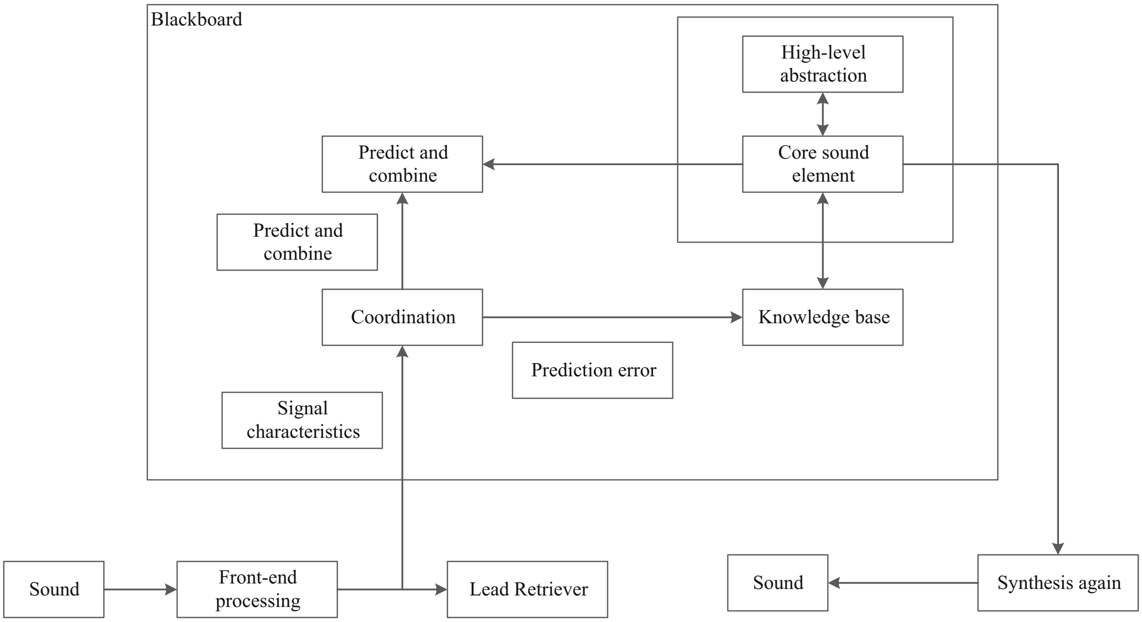


Fig. 2 Schematic-driven CASA system

Hypothetical hierarchies usually divide a blackboard into several well-defined layers. The assumptions of each layer are connected by "explanation" and "support". The knowledge base can not only produce assumptions, but also modify assumptions. Because of the different levels, the hypothesis is divided into two kinds: "explanation" and "prediction". A control system, in the course of a system execution at a certain time, can choose which knowledge base to be used.

3 Speech separation of speech deep neural networks based on auditory analysis

3.1 Deep neural network

DNN is essentially an artificial neural network with a multi-layer network structure [14]. Generally, the number of hidden layers is more than 1 of artificial neural networks. The extension and development of artificial neural networks form a deep neural network structure. Professor Hinton at the University of Toronto published "Reducing the dimensionality of data with neural networks" in 2006, which introduces a new neural network model. Deep learning has aroused the enthusiasm and interest of relevant researchers. The basic principle is to establish a cognitive model of the human brain. By integrating the features of the low layer, the high-level information expressed in the abstract is generated. Then, top-level information is used to express attributes, goals, or characteristics. The distribution of the characteristic expressions of the data is mined. It can gather, classify and process information on large scale data, so that the computer can interpret the data in the imitation of the cognitive ability of the human brain. In recent years, the research of deep learning theory has made rapid development in various fields, such as speech, image and artificial intelligence.

By using the mathematical model, the deep neural network is set up a certain incentive function to carry out supervised or unsupervised training. The network parameters of the data are extracted to be used in the task of feature extraction. The hierarchical topology network structure of a large number of neurons simulates a human neural network system. Through different weights, neurons stimulate each other to fit the complex relationship between the data, dig deeper information between the data and affect the output of the entire neural network. As its depth deepens, the advantages of deep neural networks become more and more prominent. The increase of hidden layers greatly improves the adaptability and self-organizing ability of its network.

3.1.1 Structure of deep neural networks

Deep neural network is developed on the basis of artificial neural network. DNN is one of the most classic model structures of the depth learning algorithm. Compared with the early MLP, the structural difference lies in the inclusion of multiple hidden layers between its input and output layers. As shown in Fig. 3, a typical DNN model structure contains 4 hidden layers. Compared with the shallow layer model, there are many hidden layers in the deep neural

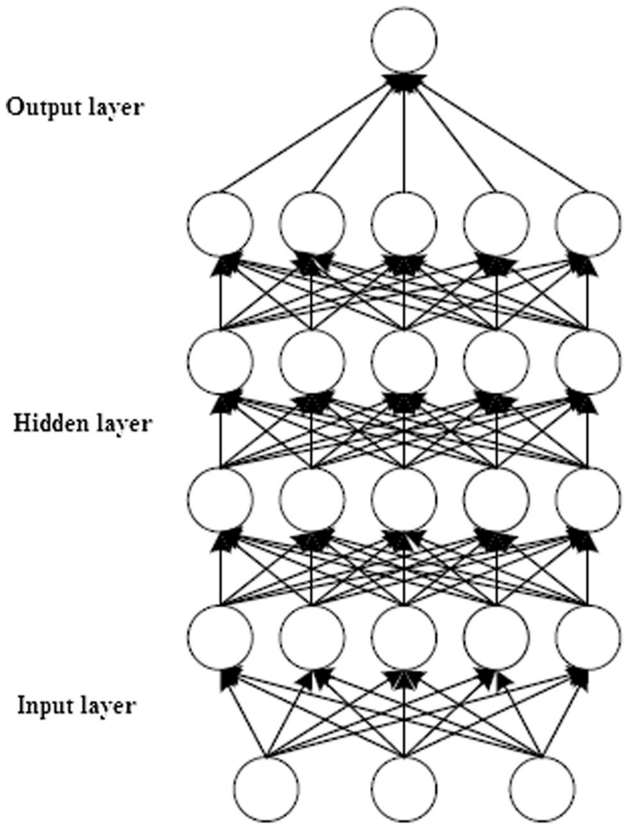


Fig. 3 DNN structure diagram with four hidden layers

network. The network structure of the multi hidden layer makes it closer to the hierarchical process of human brain processing information. It has excellent nonlinear modeling and characteristic learning ability. The characteristics of this learning have a more essential characterization of the data, which is beneficial to the classification. The essence of deep learning is to build more hidden layers and use a lot of training data to learn more advanced and valuable features, so as to improve the accuracy of prediction. In short, the deep neural network model is only a means, and the feature learning is the goal. By learning the deep nonlinear network structure, the depth model can implement the complex function approximation to express the distributed expression of the input data.

3.1.2 Deep neural network training

In the deep learning process, if all the layers are trained at the same time, the computational complexity will be too high. However, if training is carried out layer by layer, the accumulated deviation of the generated deviation will increase, resulting in a serious problem of under-fitting input and output. Deep Learning uses bottom-up unsupervised learning and top-down supervised learning. The main steps are as follows:

3.1.2.1 Pre-training Starting from the lowest level, unsupervised methods are often used to learn the network parameters layer by layer from the bottom up. The output of the upper layer is used as the input of the next layer, until the top level is trained, to discover the abstract features of the high-level information in the input data.

3.1.2.2 Fine adjustment The network parameters determined by the pre-training are used as initial values. Top-down supervised learning is adopted. The data with the label is used for training. The error is transmitted from the top down and the network is fine-tuned.

Deep neural networks have strong learning ability and universal applicability in speech signal and image processing technology. Deep learning is a non-linear feature extraction of low-level inputs, which reveals the mapping of low-level to high-level. A large number of experimental data show that using deep neural network to select and classify features, it can obtain effective feature abstract expression. In recent years, it has been widely used in the fields of Internet, image processing and voice enhancement.

This chapter studies the reverberation speech separation of deep neural networks based on auditory scene analysis. The speech separation algorithm of DNN is proposed, to get a more robust speech separation feature.

3.2 Mono reverberation speech separation based on DNN learning

Reverb corresponds to the convolution of direct sound and room impulse response. In the time and frequency domain, the voice's soundtracks are distorted. Therefore, anti-reverberation can be seen as inverse filtering. The magnitude relationship between the echo-free signal and its reverb version is relatively consistent under different reverberation conditions, especially in the same room. Even when the reverb speech and the background noise are mixed, it is possible to restore the speech to some extent from the mixed speech. Voice is highly structured. From the above, the reverberation process is modeled using supervised learning methods.

Based on the DNN monophonic reverberant speech separation algorithm, a mapping learning from reverberant speech to pure speech spectrogram is proposed. The input is the spectral representation of the reverberant speech and the desired output is its pure speech. In this process, the mapper completes the training and then extends the spectrum mapping method to perform anti-reverberation and denoising. DNN-based mono reverberation speech separation algorithm consists of three parts: feature extraction, model training and post-processing.

3.2.1 Feature extraction

First, spectrum mapping extracts features. The time domain input signal $S(t)$ is given. Short time Fourier transform (STFT) is used to extract features. The input signal is decomposed into a series of units. The frame length is 20 ms and the frame shift is 10 ms. Then the logarithmic spectrum of each time frame is calculated by fast Fourier transform. For a 16 kHz signal, the 320—point fast Fourier transform is used, and the number of frequency points is 161. The logarithmic order of the m frame and the k frequency point is expressed as $X(m,k)$. Therefore, in the spectrum domain, each time frame can be expressed as a vector $X(m)$:

$$X(m) = [X(m,1), X(m,2), \cdots, X(m,161)]^T \tag{1}$$

Since adjacent time frames contain useful information for feature extraction, merging adjacent time frames into eigenvectors will improve learning accuracy. In order to combine the temporal dynamics, the spectral features of adjacent time frames are merged into the eigenvectors. Therefore, the input feature vector $\tilde{X}(m)$ of the deep neural network feature map is expressed as:

$$\tilde{X}(m) = [X(m-d), \cdots, X(m), \cdots X(m+d)]^T \tag{2}$$

Among them, d denotes the number of adjacent frames on each side of the m time frame. In this study, d is set to 5. The input dimension is $16 \times 11 = 1771$.

The expected output of the neural network is the spectrogram of the pure speech at frame m at the current time. It is expressed as a 161-dimensional feature vector $Y(m)$. The element corresponds to the logarithmic magnitude of each frequency at time frame m.

3.2.2 DNN model training

This section trains deep neural networks to learn spectrum maps from reverberant speech or reverberation plus noise signals to pure speech signals. DNN includes three hidden layers in this paper. The optimization objective function is based on the mean square error function, and the cost function for each training sample is as follows:

$$L(Y, X; \Theta) = \sum_{c=1}^{C} (y_c - f_c(X))^2 \tag{3}$$

$c = 161$. It is equivalent to the highest frequency point index. $Y = (y_1, \ldots, y_C)^T$ represents the desired output vector. $f_c(\cdot)$ refers to the actual output of the cth neuron in the output layer.Θ indicates learning parameters. In order to train the neural network, the training center inputs the normalized zero mean and outputs the normalized in $[0, 1]$.

DNN weights are initialized at random. In this paper, 512 samples of each batch are used to train the DNN model using back propagation algorithm and stochastic gradient descent method. The actual cost of each mini-batch is calculated by Eq. (3).

3.2.3 Post-processing

The post processing is the target speech spectrum of the DNN output, and the target time domain signal is reconstructed. The most direct way is to use reverse fast Fourier transform. In order to minimize the incoherence between the phase of the reconstructed signal and the magnitude of the magnitude, the time domain signal is reconstructed by multiple iterations. The phase ϕ is repeatedly updated by the STFT and reverse STFT of each step, and the target Y^0 is always fixed as the output of DNN. In the condition of giving the order of magnitude spectrum, the purpose of the iteration is to find the nearest order of magnitude spectrum. The time domain signal is reconstructed as the output of the waveform of the system.

3.3 Speech separation of double channel reverberation based on DNN classification

This study solves the problem of the separation of double sound reverberation. The difference between the single channel separation and the double channel separation is the feature extraction. The classification based approach is a prominent advantage. The separation of double sound and mono channel can be easily solved by the simple fusion of the characteristics of the double channel and the single channel. Under a variety of unfavorable conditions, the classification method based on the characteristics of single channel and double channel is used to simulate the auditory characteristics of the human ear. The proposed dual channel classification method includes single track cues in classification. The single-track clues are crucial when the interference voice and target speech are in the same or near distance. This paper does not solve the problem of sound source positioning.

3.3.1 The principle of the algorithm

Figure 4 shows a flow chart of the model of a dual channel speech separation system based on DNN classification. Two identical auditory filter banks convert the input signals of the left and right channels into the time frequency domain. The output of each frequency channel is divided into a time frequency T-F unit with a frame length of 20 ms. It is represented by $u[c,m]$. Among them, c represents the channel. m is a time frame. The peripheral

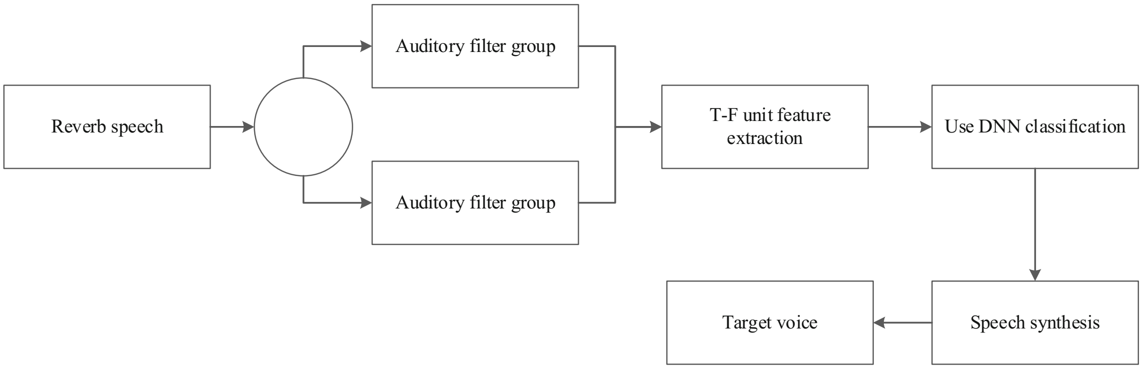


Fig. 4 Flow chart of dual channel speech separation system based on DNN classification

auditory analysis produces the time frequency representation of the reverberation speech.

The feature extraction of both of them is carried out in the time frequency domain. The characteristics of the dual channel are calculated by each pair of T-F units in the left and right channels. The single channel feature GFCC is extracted from the single frame T-F unit in the left channel signal. Each T-F unit is used as input by processing the signal. The traditional feature extraction of time frame is realized by calculating the eigenvalues of each T-F unit.

Then, DNN is trained by the discrimination of the whole feature set in the noise reverberation environment. Since the characteristics of the dual channel and mono channel vary with the frequency, a DNN classifier is trained for each frequency channel. The training label is provided by IBM. In the test, the output of the DNN is interpreted as the posterior probability that the T-F unit is dominated by the target voice. The marking criterion is used to evaluate the IBM, and all the T-F units marked as the target constitute the separated target stream.

3.3.2 Feature extraction

Human ear auditory system is an important physiological structure for receiving speech information. Its auditory mechanism is very complex. It has very important research significance. The human ear auditory system is more reliable and convenient than other speech separation systems, especially in the reverberation environment. Therefore, the accurate establishment of the auditory model in the speech separation system is of great significance for improving the performance of the system.

Auditory peripheral processing. Front end processing is extracted before feature, including pre—weighting, framing and adding window.

Pre-emphasis. Pre-emphasis generally uses a first-order FIR high-pass filter with a transfer function $H(z) = 1 - \alpha z^{-1}$. The coefficient is set to 0.97. The filter function is calculated as follows:

$$x(n) = y(n) - 0.97y(n-1) \tag{4}$$

Among them, $y(n)$ is the input voice, and the $x(n)$ is the preaccentuated output speech.

3.3.2.1 Framing The frame processing is overlapped and segmented. The overlapped time of adjacent frames is half of the frame length to achieve a smooth transition. Here, the frame takes 20 ms, and the frame shift is 10 ms.

3.3.2.2 Window The signal $x(n))$ and the window function $y(n)$ are multiplied to realize the frame. There are two common window functions in speech signal processing. One is a rectangular window, and the expression is as follows:

$$w(n) = \begin{cases} 1 & 0 \le n \le (N-1) \\ 0, & otherwise \end{cases} \tag{5}$$

The auditory peripheral processing is to use the Gammatone filter bank to filter the input signal. The Gammatone filter simulates the cochlear model to realize the frequency division of the basement membrane. The central frequency of the filter is distributed in the way of equal distance bandwidth. The impact response of the filter is expressed as:

$$g(c,t) = \left\{ \begin{matrix} t^{d-1}e^{-2\pi b(f_c)t}\cos(2\pi f_c t),\ if\ t \ge 0 \\ 0\ , otherwis \end{matrix} \right\} \tag{6}$$

c represents the filter channel. There are 64 passages in all. The central frequency f_c of the filter varies from 80 to 5000 Hz. $b(f_c)$ represents bandwidth. The filter order d is 4.

Single channel feature extraction. In order to extract the feature of the single channel, the unit response $x_{cm,1}$ of the left channel is used as an ordinary signal. The extraction of GFCC feature parameters requires 5 steps, as shown in Fig. 5.

3.3.2.3 Extraction of dual channel characteristics With the signal input of the double channel, two basic dual steps, as shown in Figure 5.

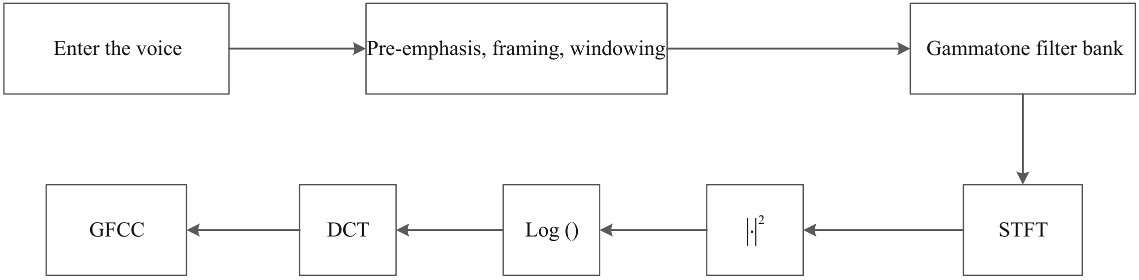


Fig. 5 Block diagram of GFCC calculation step

channel features were extracted: the time difference between the ears was ITD and the difference between the ear level was ILD. The ITD is obtained by calculating the normalized cross correlation function (CCF) between the double channel signals.l represents the left channel signal.r represents the right channel signal. The index delay is τ. The CCF of a set of T-F units is expressed in the following form:

$$CCF(c,m,\tau)=\frac{\sum_{k}(x_{cm,l}(k)-\tilde{x}_{cm,1})(x_{cm,r}(k-\tau)-\tilde{x}_{cm,r})}{\sqrt{\sum_{k}(x_{cm,l}(k)-\tilde{x}_{cm,1})^2}\sqrt{\sum_{k}(x_{cm,r}(k-\tau)-\tilde{x}_{cm,r})^2}} \tag{7}$$

In the above equation, the range of τ values is $[-1,1]$ms. k indexes a single T-F unit signal, and the cross line on it represents the average. For a signal with a sampling rate of 16 kHz, there are 33 CCF values. The CCF value of -1 ms is ignored, and the CCF features of the 32 dimension of each group of T-F units are obtained.

3.4 Experimental results and analysis

The room dimension is set to (10 m*7 m*3 m). The reflection coefficient changes with the reverberation time T60. For each T60 condition, sound source signals and receivers in the room are uniformly and randomly located. The distance between the two is more than 0.5 m, which can effectively avoid the close range speech scene. In order to train the system, this paper uses three kinds of reverberation time T60, which are 0.3 s, 0.6 s and 0.9 s, respectively. Two different room impact responses RIRs are generated for each T60. From the TIMIT voice library, 50 sentences are selected as the training speech, and 20 sentences are selected as the test speech. There is no overlap between the training statement and the test statement. The training set consists of 50 speech convolutions and 3 RIRs, and the 3 noise is mixed with the 0 dB signal to noise ratio. The three kind of noise comes from the Noisex-92 speech database. They are speech-shaped noise, babble noise and factory noise respectively. 20 sentences of test speech are mixed with 3 kinds of training noise and 3 kinds of new noise. The new noise is white noise, cocktail noise and crowded noise in play ground. As the first 3 kinds are used for training at the same time, the test data obtained from the first 3 noise reverberations are matched test conditions. The 3 new noises are not used for training. The 3 kinds of noise are new to the model, and the generated data are non-matching test conditions. The results obtained under non-matched test conditions can indirectly reflect the general application ability of the model to the non-matched noise.

Figure 6 is the Seg-SNR value obtained from two models for mixed matching noise. DNN learning represents a single channel reverberation speech separation model based on DNN learning spectrum mapping. DNN classifications represent a two—channel reverberation speech separation model based on DNN classification. In the category of matching noise, the mean signal to noise ratio (SNR) of the single channel DNN is improved by about 8.5 dB. The method of double channel classification is more 0.5 dB. As a benchmark, this paper also uses the ideal two value mode (IBM) for the visible noise to reverberate and denoise. The target in IBM is the non-echo pure voice, and the masking is the residual signal. Because IBM uses ideal signal information, the signal to noise ratio of the DNN model is only 0.5 dB lower than the average signal to noise ratio of IBM.

Figure 7 is the Seg-SNR value obtained from two models for mixed mismatched noise. In the category of non-matched noise, the two proposed algorithms have also improved significantly. The average output signal to noise ratio (SNR) has an increase of about 7.5 dB, which shows good generalization performance. The DNN-classified two-channel separation system has a higher signal-to-noise gain for white noise and cocktail noise. For noisy people, the DNN learning spectrum mapping method is slightly better. The reason is that the two-channel separation system based on the DNN classification is aimed at a noisy crowd. The process of extracting features is numerous and complex. The computational complexity is very high, and the learning rate decreases seriously. Therefore, the effect of

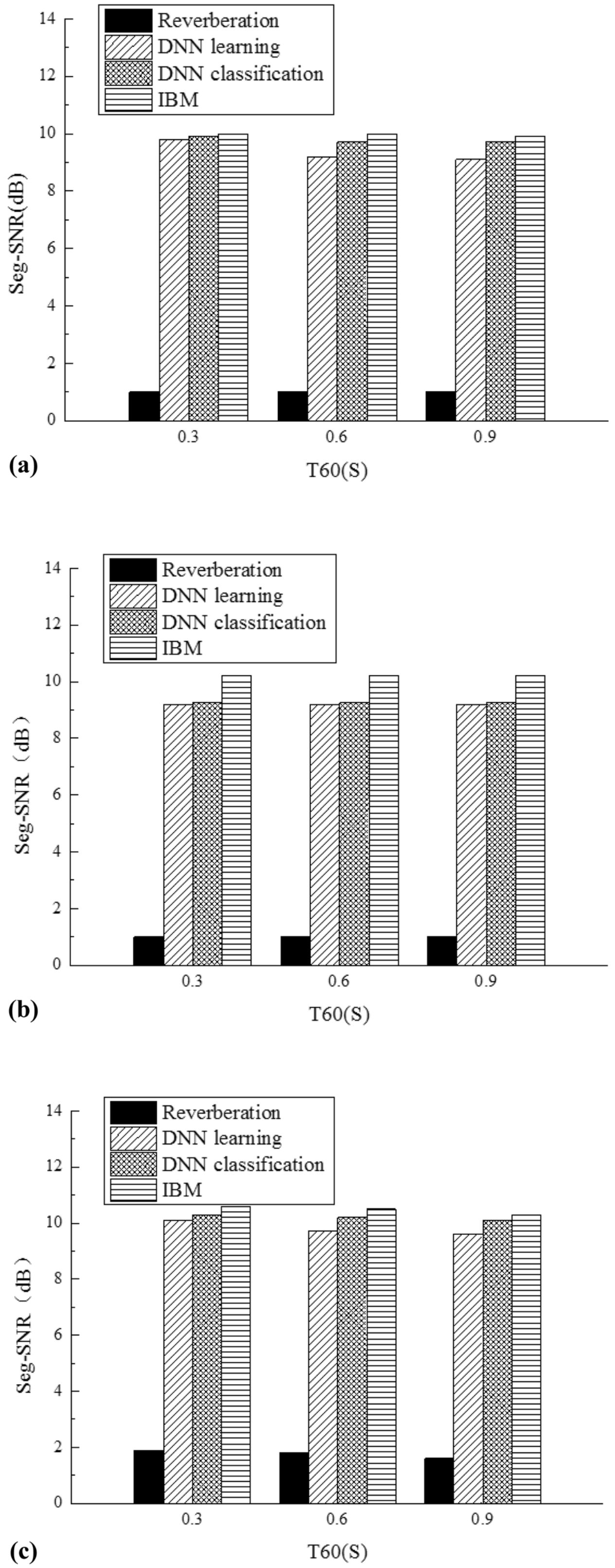


Fig. 6 Comparison of Seg-SNR results with different models for matching noise. **a** Speech-shaped noise, **b** Babble noise, **c** factory noise

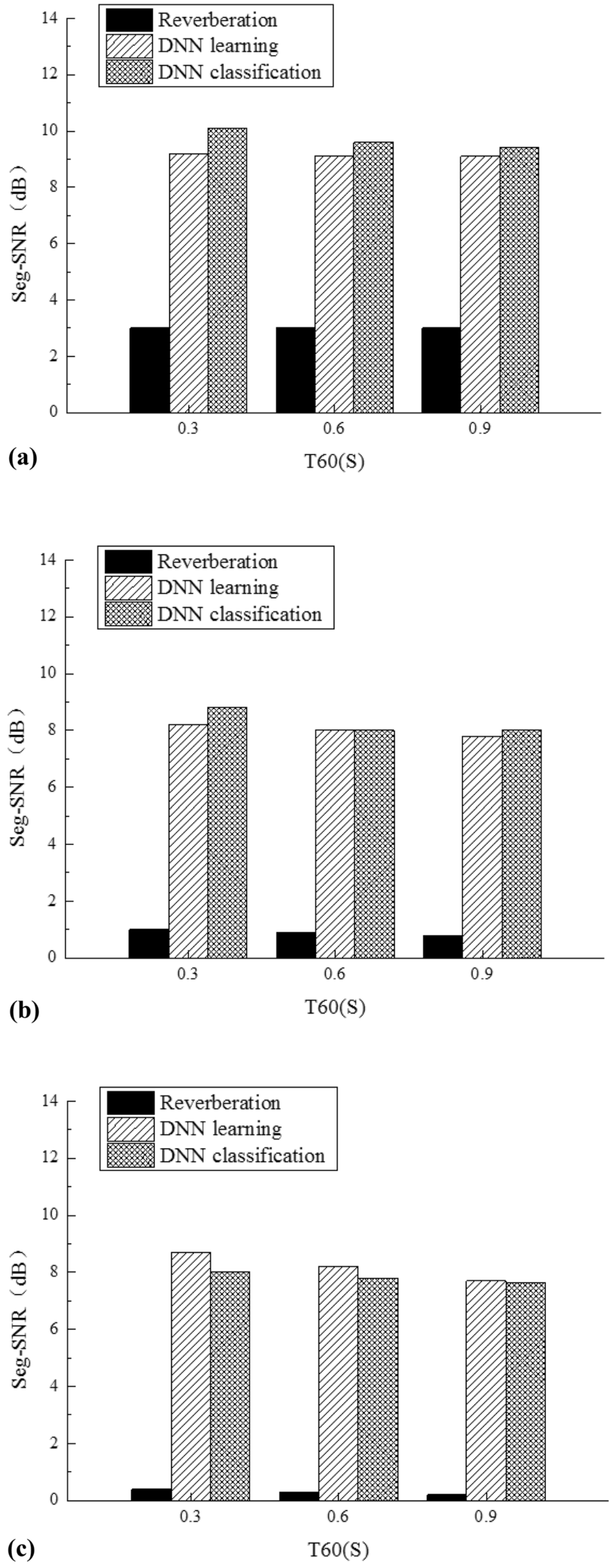


Fig. 7 Comparison of Seg-SNR results with different models for unmatched noise. **a** White noise, **b** cocktail party noise, **c** crowd noisy noise

recovering the target voice is reduced, resulting in a relatively lower gain of the signal-to-noise ratio. In general, the two-channel classification method for improving the output signal to noise ratio has some advantages.

Two-channel anti-reverberation and denoising based on DNN supervised learning mono and DNN classification is proposed. The trained DNN solves the problem of anti-reverb and de-noising by studying the spectral mapping between contaminated speech and pure speech. Although cochlear mapping slightly outperformed spectral mapping on anti-reverb effects, the ideal separation was not possible with noisy reverb. In this study, because of the time-varying information of the speech signal, the spectral characteristics of adjacent time frames in the frequency window are provided, in order to achieve better extraction results. The mapping cochlear map produced by the Gammatone filter was used to train DNN. In this case, a small portion of the input vector corresponds to the logarithmic energy of each time–frequency element of the contaminated speech. The output vector corresponds to the pure logarithmic energy of each time–frequency unit of pure speech. Oversight of learning methods significantly improves anti-reverb and noise removal issues. The dual channel DNN classification reverberation speech separation model is processed, and the multi-dimensional features are extracted. It includes cross-correlation function CCF, mono channel feature GFCC, dual channel feature ITD and ILD. The restricted Boltzmann pre-training DNN can effectively complete the classification task and extract the target speech from the reverberation speech. The experimental results show that the model can improve speech comprehension and speech quality. Compared with the early single channel field research, the same features are extracted, and the DNN classifier is superior to the MLP classifier. By combining mono channel and dual channel features, the output signal-to-noise ratio can be greatly improved. Even at very low input signal to noise ratio and strong reverberation conditions, the proposed model can also get excellent separation performance.

4 Conclusions

With the increasing demand for human–machine interaction, it has practical applications to extract desired signals from complex observation signals with noise and interference in communication and signal processing fields. In a complex environment such as a room reverberation, the characteristics of the harmonics are destroyed. The observation signal is also very complex. It is very difficult to extract the target signal. Reverberation speech separation has always been a hot and difficult point in the research of speech signal processing.

Aiming at the significant reduction of speech intelligibility and sound quality under reverberation and noise conditions, two DNN-based reverberation speech separation algorithms are proposed. The DNN-based monophonic reverberant speech separation algorithm directly conducts training on the deep neural network by learning the mapping from the polluted speech to the pure frequency speech spectrum. The distortion caused by the reverberation environment and the background noise is substantially attenuated. The DNN based speech separation algorithm for double channel reverberation makes use of the powerful learning ability of the deep neural network. By realizing the classification task of time frequency unit, the extraction of target speech is completed. The binaural features of binocular time difference and strong difference, as well as the mono features of Gammatone frequency cepstral coefficients, were used as the major auditory features of the classification. The evaluation and comparison show that the DNN-based reverberant speech separation algorithm produces good separation results under various multi-source reverberation conditions.

Acknowledgement The authors acknowledge the National Natural Science Foundation of China (Grant: 61372146, 61373098), the Youth Natural Science Foundation of Jiangsu Province of China (Grant: BK20160361), the Qinglan Project Young and Middle-aged Academic Leader Foundation of Jiangsu Province, the Professional Leader Advanced Research Project Foundation of Higher Vocational College of Jiangsu Province (Grant: 2017GRFX046).

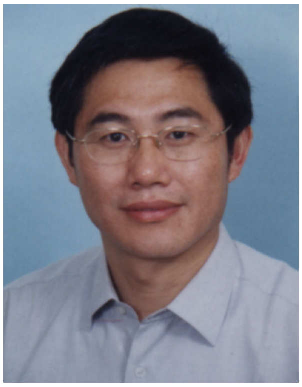
References

1. Barker, J.P.: Evaluation of scene analysis using real and simulated acoustic mixtures: lessons learnt from the chime speech recognition challenges. J. Acoust. Soc. Am. **141**(5), 3693–3693 (2017)
2. Asaei, A., Taghizadeh, M. J., Cevher, V.: Computational methods for underdetermined convolutive speech localization and separation via model-based sparse component analysis. Speech Commun. **76**(C), 201–217 (2016)
3. Josupeit, A., Kopčo, N., Hohmann, V.: Modeling of speech localization in a multi-talker mixture using periodicity and energy-based auditory features. J. Acoust. Soc. Am. **139**(5), 2911 (2016)
4. Scholes, C., Palmer, A.R., Sumner, C.J.: Stream segregation in the anesthetized auditory cortex. Hear. Res. **328**(2), 48–58 (2015)
5. Denham, S., Coath, M.: The role of form in modeling auditory scene analysis. J. Acoust. Soc. Am. **137**(4), 2249–2249 (2015)
6. Vander, G.M., Bourguignon, M., de Beeck, M., Wens, V., Marty, B., Hassid, S., et al.: Left superior temporal gyrus is coupled to attended speech in a cocktail-party auditory scene. J. Neurosci. **36**(5), 1596–1606 (2016)
7. Rogalsky, C., Poppa, T., Chen, K.H., Anderson, S.W., Damasio, H., Love, T., et al.: Speech repetition as a window on the neurobiology of auditory-motor integration for speech: a voxel-based lesion symptom mapping study. Neuropsychologia **71**(01), 18 (2015)
8. White-Schwoch, T., Davies, E.C., Thompson, E.C., Carr, K.W., Nicol, T., Bradlow, A.R., et al.: Auditory-neurophysiological

responses to speech during early childhood: effects of background noise. Hear. Res. **328**, 34–47 (2015)

9. Moossavi, A., Mehrkian, S., Lotfi, Y., Faghih Zadeh, S., Adjedi, H.: The effect of working memory training on auditory stream segregation in auditory processing disorders children. Optics Commun **281**(9), 2491–2497 (2015)
10. Kenway, B., Tam, Y.C., Vanat, Z., Harris, F., Gray, R., Birchall, J., et al.: Pitch discrimination: an independent factor in cochlear implant performance outcomes. Otol. Neurotol. **36**(9), 1472–1479 (2015)
11. Mathon, B., Ulvin, L.B., Adam, C., Baulac, M., Dupont, S., Navarro, V., et al.: Surgical treatment for mesial temporal lobe epilepsy associated with hippocampal sclerosis. Revue Neurol. **171**(3), 315–325 (2015)
12. Leclère, T., Lavandier, M., Culling, J.F.: Speech intelligibility prediction in reverberation: towards an integrated model of speech transmission, spatial unmasking, and binaural de-reverberation. J. Acoust. Soc. Am. **137**(6), 3335–3345 (2015)
13. Léger, A.C., Reed, C.M., Desloge, J.G., Swaminathan, J., Braida, L.D.: Consonant identification in noise using hilbert-transform temporal fine-structure speech and recovered-envelope speech for listeners with normal and impaired hearing. J. Acoust. Soc. Am. **138**(1), 389–403 (2015)
14. Koralus, P.: Can visual cognitive neuroscience learn anything from the philosophy of language? ambiguity and the topology of neural network models of multistable perception. Synthese **193**(5), 1409–1432 (2016)

Yan Zhou received the B.Sc. degree and M.Sc. degree in China University of Geosciences in June 2003 and June 2006, respectively. And in June 2017, she received the Doctor's degree in Signal and Information Processing in University of Suzhou, Jiangsu, China. From July 2006 to now, she works at the Department of Communication Technology, Electronic Information Engineering College, Suzhou Vocational University. At present, her research interests include speech signal processing, Deep Learning and Intelligent Computing.



Heming Zhao received the M.Sc. degree in Soochow University in 1982. From 1984 to 1985, he studied in Tsinghua University, and from 1988 to 1990, he was a visiting scientist in Technical University of Munich, German. Since 1982, he works at Soochow University, during1995–1997, he was a director at the Department of Electrical Engineering, and now he is a doctoral tutor and a dean at the College of Electronic Information. He is a member of IEEE, and senior member of Chinese Electronic Institute, and he is a member of the editorial boards of Signal Processing. At present, his research interests include speech signal processing, and intelligent computing.

Jie Chen received the Master's degree from Zhejiang University, Hangzhou, China. He is currently pursuing the Ph.D. degree at the School of Rail Transportation of Soochow University, in the area of video processing and machine vision. From April 2006 to now, he -works at the Institute of Electronic Information Engineering, Suzhou Vocational University, Suzhou, Jiangsu, China. At present, His research interests include image processing, video coding, and pattern recognition.

Xinyu Pan received the B.Sc. degree in University of Science and Technology of Suzhou in June 2004, and the M.Sc. degree in Soochow University in June 2007, respectively. And in December 2016, he received the Doctor's degree in Signal and Information Processing in University of Suzhou, Jiangsu, China. From July 2007 to now, he works at the College of Electronics & Information Engineering, University of Science and Technology of Suzhou. His research interests include speech signal processing and Soft Computing.