Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn



Latent source-specific generative factor learning for monaural speech separation using weighted-factor autoencoder*

Jing-jing CHEN[1], Qi-rong MAO‡[1,2], You-cai QIN[1], Shuang-qing QIAN[1], Zhi-shen ZHENG[1]

[1]*School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China*

[2]*Jiangsu Key Laboratory of Security Technology for Industrial Cyberspace, Zhenjiang 212013, China*

E-mail: 2221808071@stmail.ujs.edu.cn; mao_qr@ujs.edu.cn; 2211908026@stmail.ujs.edu.cn; 2211908025@stmail.ujs.edu.cn; 3160602062@stmail.ujs.edu.cn

Received Jan. 13, 2020; Revision accepted June 21, 2020; Crosschecked Sept. 8, 2020

Abstract: Much recent progress in monaural speech separation (MSS) has been achieved through a series of deep learning architectures based on autoencoders, which use an encoder to condense the input signal into compressed features and then feed these features into a decoder to construct a specific audio source of interest. However, these approaches can neither learn generative factors of the original input for MSS nor construct each audio source in mixed speech. In this study, we propose a novel weighted-factor autoencoder (WFAE) model for MSS, which introduces a regularization loss in the objective function to isolate one source without containing other sources. By incorporating a latent attention mechanism and a supervised source constructor in the separation layer, WFAE can learn source-specific generative factors and a set of discriminative features for each source, leading to MSS performance improvement. Experiments on benchmark datasets show that our approach outperforms the existing methods. In terms of three important metrics, WFAE has great success on a relatively challenging MSS case, i.e., speaker-independent MSS.

Key words: Speech separation; Generative factors; Autoencoder; Deep learning
https://doi.org/10.1631/FITEE.2000019 **CLC number:** TN912.3

1 Introduction

Speech separation, also known as audio source separation, is a significant task in signal processing. The aim of speech separation is to separate target speech from a mixed audio signal, and this work is important for some real-world applications. For example, it can separate clean speech from a noisy speech signal to improve the accuracy of automatic speech and speaker recognition. A machine solution to this problem is critical to enable resolvability of many scenarios, such as meeting transcription, hearing prosthesis, and mobile telecommunication. Currently, speech separation tasks are divided mainly into monaural speech separation (MSS) and multi-microphone speech separation. MSS deals with the task of separating audio signals of multiple simultaneous speakers from a single-channel recording, while multi-microphone speech separation tackles the situation where multiple monaural recordings are available.

An array of microphones can furnish multiple

‡ Corresponding author

* Project supported by the Key Project of the National Natural Science Foundation of China (No. U1836220), the National Natural Science Foundation of China (No. 61672267), the Qing Lan Talent Program of Jiangsu Province, China, and the Key Innovation Project of Undergraduate Students in Jiangsu Province, China (No. 201810299045Z)

ORCID: Jing-jing CHEN, https://orcid.org/0000-0003-2968-0313; Qi-rong MAO, https://orcid.org/0000-0002-0616-4431

© Zhejiang University and Springer-Verlag GmbH Germany, part of Springer Nature 2020

monaural recordings, which may include information indicating the spatial locations of the audio sources. When there are multiple mixed audio signals, we can make full use of these spatial locations to achieve the goal of speech separation (Qian et al., 2018). The main solutions are independent component analysis (Hyvärinen and Oja, 2000; Araki et al., 2007), beamforming (Benesty et al., 2008), and so on. However, it is more common to address the problem of MSS in daily life, which is typically more difficult than its multi-microphone counterpart. In general, MSS is more common and highly challenging.

Despite the challenging nature of this task, a number of approaches for MSS have been proposed in previous work. These approaches can be divided roughly into two categories, namely traditional methods and deep learning approaches. Before deep learning, many popular techniques have been used to solve the problem of MSS, such as computational auditory scene analysis (CASA) (Bregman, 1990; Brown and Cooke, 1994; Wang DL and Brown, 2006; Hu and Wang, 2013), non-negative matrix factorization (NMF) (Schmidt and Olsson, 2006; Smaragdis, 2007), and Bayesian methods (Nadas et al., 1989; Ghahramani and Jordan, 1997; Roweis, 2001). In recent years, many approaches based on deep learning have been proposed to deal with various difficult problems (Gou et al., 2018; Xia et al., 2019). Similarly, many techniques (Huang et al., 2014; Hershey et al., 2016; Wang YN et al., 2016; Chen et al., 2017; Yu et al., 2017; Luo and Mesgarani, 2019; Luo et al., 2019) based on neural networks have been used to solve the task of MSS. For example, Huang et al. (2014) proposed joint optimization of the deep learning models (deep neural networks and recurrent networks) with an extra masking layer, which enforces a reconstruction constraint. Some researchers applied autoencoder or related methods to the task of speech separation (Grais and Plumbley, 2017; Osako et al., 2017; Pandey et al., 2018; Williamson, 2018; Karamatlı et al., 2019), and their results showed that the autoencoder and variational autoencoder (VAE) can learn the inherent latent representation of a source by encoding the non-linear dependencies. However, these approaches directly use a decoder to construct a specific audio source of interest, ignoring the process of reconstructing the original mixed signal. This signal cannot learn the generative factors of the original input, leading to negative effect on separation performance.

In this study, we propose a novel weighted-factor autoencoder (WFAE), which maintains all parts of the original autoencoder and adds a new separation layer constrained by a regularization loss. By incorporating a latent attention mechanism and a supervised source constructor, our WFAE can learn source-specific generative factors and a set of discriminative features for each source, thus achieving better separation performance. The contributions of this study are as follows:

1. To the best of our knowledge, this is the first work that introduces generative factors when applying autoencoders or related architectures to MSS.

2. We use a latent attention mechanism to weigh the generative factors of the mixed speech signal, to learn source-specific generative factors and a set of discriminative features for each acoustic source. Extensive experiments on benchmark datasets show that our proposed approach significantly outperforms the current methods, and that our model has great success on a relatively challenging MSS case, i.e., speaker-independent MSS.

2 Related work

In this section, we provide the background on MSS. Two main research directions have been studied in the field of MSS: (1) traditional methods; (2) deep learning approaches. We briefly review traditional methods, and then focus on deep learning approaches as they are more related to this work.

2.1 Traditional methods for monaural speech separation

Traditional methods for MSS consist of three techniques: (1) CASA; (2) NMF; (3) Bayesian methods. Inspired by the perceptual ability of human beings to extract signals of interest from background sounds, CASA (Brown and Cooke, 1994; Wang DL and Brown, 2006; Hu and Wang, 2013) has been used to deal with MSS. Generally, there are two main stages in CASA methods: segmentation and grouping. Segmentation decomposes a mixed signal into time-frequency (T-F) segments and the simultaneous and sequential grouping gathers the T-F segments to extract a particular speaker from the mixed signal. Various segmentation and grouping regulations have been used to assemble T-F segments

that are believed to pertain to the same speaker. On the other hand, NMF (Schmidt and Olsson, 2006; Smaragdis, 2007) decomposes the spectrogram of the mixed signal into specific activations relating to each speaker with non-negative dictionaries, and an individual target signal from these activations can be approximated using these dictionaries. Apart from the above popular techniques, Bayesian methods such as factorial hidden Markov model (HMM) (Ghahramani and Jordan, 1997; Roweis, 2001) have been explored. However, these methods are not always applicable in practical scenarios because of the lack of prior knowledge of speakers. In addition, limited to modeling capabilities, these methods introduce only moderate performance gain and lag far behind human ability.

2.2 Deep learning approaches for monaural speech separation

In general, deep learning approaches can be grouped into three categories: (1) conventional regression algorithms; (2) clustering-related methods; (3) autoencoder-related approaches. Using a conventional regression algorithm, Wang YN et al. (2016) extended a deep neural network (DNN) approach to unsupervised speech separation of two speakers who are both unknown, and measured the dependence of this approach on the distance between target speaker and a competing speaker. Huang et al. (2014) explored learning the optimal hidden representations to reconstruct the target spectra by a method of deep learning, and proposed a discriminative training criterion for the neural networks to further enhance separation performance.

Clustering-related methods have been proposed to address the label permutation problem in speech separation. For example, to predict the segmentation labels of the target speech from the mixture, Hershey et al. (2016) trained a deep clustering network to allot contrastive embedding vectors to each T-F region of the spectrogram. A novel neural network framework called a deep attractor network (Chen et al., 2017), which creates attractor points in a high-dimensional embedding space, has been proposed to solve general speech separation problems. Compared with deep clustering, the main advantage of the deep attractor network is efficiency in performing end-to-end mapping, and the main drawback of the deep attractor network is the added complexity associated with estimating attractor points during inference.

In recent years, there have been some approaches that attempt to separate monaural speech using autoencoders or related frameworks. For example, in Grais and Plumbley (2017), an autoencoder was changed to a supervised mode, keeping the model structure unchanged excepting for the output. The output of the model was no longer devoted to reconstructing the given input, but rather to the construction of a specific audio signal of interest. This method uses the fundamental frame of the autoencoder to achieve speech separation, and trains an autoencoder for each type of source while treating other sources as background noise. A similar structure has been adopted (Pandey et al., 2018) with the difference of turning the autoencoder into a variational autoencoder. It uses an encoder to learn robust compressed features of sources corrupted with noise, and the compressed features are then fed to a decoder to yield the separated source. This type of approach is intended to learn the compressed features of the input data and then use these compressed features to construct the target audio signal of interest. In Karamatlı et al. (2019), a weak supervision method was used for learning to separate short utterance mixtures. This applies only class information rather than complete knowledge of the source signals. In addition, Osako et al. (2017) employed the autoencoder mechanism for dictionary training such that the non-linear network can encode the target source with higher expressiveness than a generative NMF model. In fact, it uses deep autoencoders as dictionaries to represent sources. Williamson (2018) presented an approach that uses a deep denoising autoencoder to estimate phase-aware training targets from phase-aware input features. The pivotal difference between Williamson (2018)'s approach and all the other approaches is that deep learning is used to map phase-aware features to phase-aware training targets. The above studies proved the feasibility of the autoencoder method through experiments.

While these autoencoder-related methods can learn the inherent latent representation of a source by encoding the non-linear dependencies, the existing autoencoder-based methods neither ensure that the compressed features capture the generative factors of the original input, nor construct each audio source in the mixed speech. To deal with this

problem, we propose a novel WFAE model inspired by the autoencoder-based MSS methods. Our WFAE method introduces a latent attention mechanism and a supervised source constructor in a separation layer to learn source-specific generative factors and a set of discriminative features for each source, while maintaining all parts of the original autoencoder to ensure that the compressed features capture the generative factors of the input signal.

3 Weighted-factor autoencoder

3.1 Extracted features and training target

In this study, short-time Fourier transform (STFT) features are extracted for MSS, comprising the STFT magnitude spectra and STFT phase spectra. Specifically, as described in Yu et al. (2017), we denote the set of clean source signals in the time domain as $x_s(t)$ $(s = 1, 2, \ldots, S)$ and the mixed signal sequence as $y(t) = \sum_{s=1}^{S} x_s(t)$, where S is the number of speech sources. The STFT features of clean source signals $x_s(t)$ and the mixed signal sequence $y(t)$ are denoted as $\boldsymbol{X}_s(t, f)$ and $\boldsymbol{Y}(t, f) = \sum_{s=1}^{S} \boldsymbol{X}_s(t, f)$, respectively, for time t and frequency f. In general, it is assumed that only the STFT magnitude spectra of the speech signal are provided during separation, and that phase information is used only when recovering the time-domain waveforms, even though some researchers studied speech separation with phase spectra (Erdogan et al., 2015; Williamson, 2018). So, given the magnitude spectra $|\boldsymbol{Y}(t, f)|$ of the mixture, the goal of MSS is to recover the magnitude spectra $|\boldsymbol{X}_s(t, f)|$ of each audio source.

In the field of MSS, it is important to define a suitable training target for learning. There are two main training objectives: masking targets and mapping targets. The first describes the T-F relationship between clean speeches and mixed speech, and the other maps the spectra representation of clean speeches directly. In our proposed method, we use one of the most common masking targets, the spectral magnitude mask (Wang YX et al., 2014). The spectral magnitude mask is defined based on the STFT features of clean speeches and mixed speech:

$$M_s(t, f) = \frac{|\boldsymbol{X}_s(t, f)|}{|\boldsymbol{Y}(t, f)|}. \tag{1}$$

Here, take Eq. (1) as the training target of our model.

3.2 Architecture of weighted-factor autoencoder

In this subsection, we propose a novel autoencoder-based model called WFAE to learn source-specific generative factors and discriminative features for MSS. The architecture of WFAE is shown in Fig. 1, which consists of an encoder, a decoder, and a separation layer. The function of the encoder is to condense the input mixed voice signal into compressed features $\boldsymbol{z}$, and the decoder is used to make sure that the compressed features $\boldsymbol{z}$ capture the generative factors of the original signal by reconstructing the given input. The separation layer is designed to separate the mixed speech and obtain the clean speech of each source. Next, we will describe our model in detail.

3.2.1 Encoder

We use the convolutional neural network (CNN) as an encoder to model a sequence of frames. This is consistent with the adoption of CNN in most current autoencoder or related work (Grais and Plumbley, 2017; Hsu et al., 2017; Zhang and Zhang, 2018; Karamatlı et al., 2019). CNN can not only deal with a segment of speech signal at one time, but also ensure that the parameters of the model are limited to a certain extent. The structure of the proposed encoder is based on Hsu et al. (2017), where 1-by-F filters are applied at the first convolutional layer and w-by-1 filters are applied at the remaining layers. F represents the number of frequency bins, i.e., 129 in this study. As recommended in Radford et al. (2016), instead of using a pooling layer, we make the stride size >1 for down-sampling along the time axis. In the (inChannel, outChannel)–kernel size–(strideW, strideH) format, the encoder model has one (1, 64)–(1, 129)–(1, 1) convolutional layer, one (64, 128)–(3, 1)–(2, 1) convolutional layer, one (128, 256)–(3, 1)–(2, 1) convolutional layer, and a flattened layer. Then, the layer of generative factors with 256 units is connected to the flattened layer of the encoder.

We first obtain the waveform of the mixed speech signal, and then extract STFT features from the waveform. The magnitude spectra of STFT are compressed as low-dimensional compressed features $\boldsymbol{z}$ by the encoder. Note that $\boldsymbol{z}$ will capture the generative factors of the input speech signal as long as we use a decoder to reconstruct the given input. $\boldsymbol{z}$

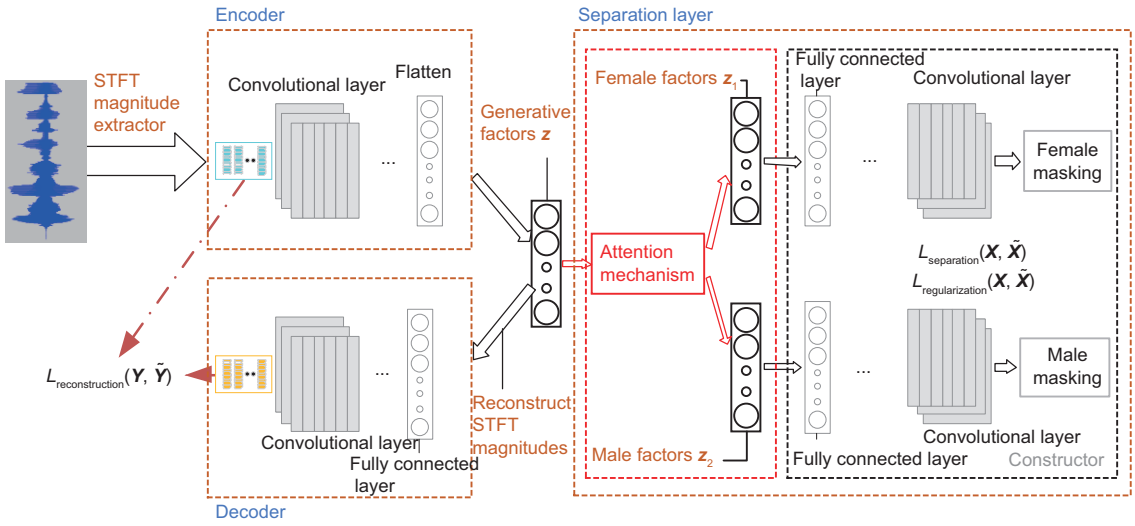


Fig. 1 Architecture of our weighted-factor autoencoder

References to color refer to the online version of this figure

can be denoted as

$$\boldsymbol{z} = \mathrm{ReLu}(W^{\mathrm{e}}\mathrm{encoder}(|\boldsymbol{Y}(t,f)|) + b^{\mathrm{e}}), \quad (2)$$

where $|\boldsymbol{Y}(t,f)|$ represents the STFT magnitude spectra of the input speech at time t with frequency bins f. W^{e} and b^{e} represent the weight and bias of the layer of generative factors, respectively.

3.2.2 Decoder

To ensure that the compressed features $\boldsymbol{z}$ in the encoder grasp the generative factors of the original speech signal, it is necessary to train a decoder to reconstruct the given input. In the decoder, we use a network architecture that is symmetrical to the encoder. This is a common technique for autoencoders. Here, we define the reconstruction loss as

$$L_{\mathrm{reconstruction}} = \frac{1}{TF}\||\tilde{\boldsymbol{Y}}(t,f)| - |\boldsymbol{Y}(t,f)|\|^2, \quad (3)$$

where T and F denote the numbers of time frames and frequency bins, respectively. $|\tilde{\boldsymbol{Y}}(t,f)|$ denotes the reconstructed signal corresponding to homologous $|\boldsymbol{Y}(t,f)|$. $|\tilde{\boldsymbol{Y}}(t,f)|$ can be formulated by $|\tilde{\boldsymbol{Y}}(t,f)| = \mathrm{ReLu}(\mathrm{decoder}(\boldsymbol{z}))$.

3.2.3 Separation layer

To learn specific-source generative factors for each source, an attention mechanism and a constructor are used in the separation layer (Fig. 1). As can be seen from the red dashed frame in Fig. 1, the attention mechanism is used to weigh generative factors $\boldsymbol{z}$, and there are two outputs because our experiment is aimed at learning source-specific generative factors $\boldsymbol{z}_s$ ($s \in \{1,2\}$) for two audio sources, female speech and male speech. The detailed form of the attention mechanism is shown in Fig. 2, and the generative factors $\boldsymbol{z}$ are connected to two fully connected layers separately. After softmax activation, we can obtain the attention weights of the generative factors $\boldsymbol{z}$ for each audio source:

$$\boldsymbol{a}_s = \mathrm{softmax}(\boldsymbol{W}^s\boldsymbol{z} + b^s), \quad (4)$$

where $s = 1, 2$, and $\boldsymbol{W}^s$ and b^s represent the weights and biases of the fully connected layers, respectively. Then, the source-specific generative factors $\boldsymbol{z}_s$ for each speaker are given by

$$\boldsymbol{z}_s = \boldsymbol{a}_s\boldsymbol{z}. \quad (5)$$

The constructor is shown in the grey dashed line in Fig. 1, and its architecture is symmetrical to that of the encoder. It is used to construct a spectral magnitude mask of each source according to the corresponding generative factors.

As mentioned in Section 3.1, the goal of MSS is to recover the magnitude spectra $|\boldsymbol{X}_s(t,f)|$ of each audio source. Therefore, we take the spectral magnitude mask $M_s(t,f)$ as the training target of our model. From these setups, we construct the loss function of the separation layer in the following way: We multiply the $\tilde{M}_s(t,f)$ obtained by our model with

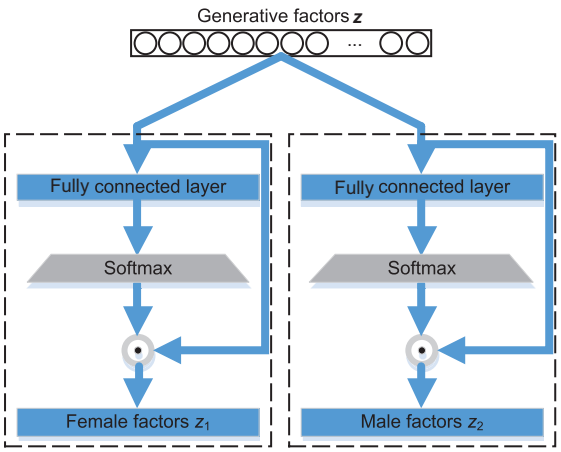


Fig. 2 Latent attention mechanism in the separation layer

the magnitude spectra $|\boldsymbol{Y}(t,f)|$ of the mixed speech to obtain the separated magnitude $|\tilde{\boldsymbol{X}}_s(t,f)|$, and then make a mean squared error with $|\boldsymbol{X}_s(t,f)|$ to construct the separation loss as

$$L_{\text{separation}} = \frac{1}{TF}\left[\left\||\tilde{\boldsymbol{X}}_1(t,f)| - |\boldsymbol{X}_1(t,f)|\right\|^2 + \left\||\tilde{\boldsymbol{X}}_2(t,f)| - |\boldsymbol{X}_2(t,f)|\right\|^2\right], \quad (6)$$

where $|\tilde{\boldsymbol{X}}_s(t,f)| = \tilde{M}_s(t,f)\,|\boldsymbol{Y}(t,f)|$ and $\tilde{M}_s(t,f) = \text{sigmoid}(\text{constructor}(\boldsymbol{z}_s))$ (Here, s=1, 2). With the attention mechanism, this loss function can constrain $\boldsymbol{z}_s$ to obtain respective source-specific generative factors from $\boldsymbol{z}$ for each speech source.

For MSS, one of the goals is to obtain a higher source-to-interference ratio (SIR). In other words, we would like to isolate one signal without containing other signals. Usually, we add a regularization term to the loss function to achieve this goal. Therefore, a discriminative training target (Huang et al., 2014) is introduced to maximize the difference between one speaker and the estimated version of the other. During the training process, we use this discriminative training target as a regularization term:

$$L_{\text{regularization}} = -\frac{\alpha}{TF}\left[\left\||\tilde{\boldsymbol{X}}_1(t,f)| - |\boldsymbol{X}_2(t,f)|\right\|^2 + \left\||\tilde{\boldsymbol{X}}_2(t,f)| - |\boldsymbol{X}_1(t,f)|\right\|^2\right], \quad (7)$$

where α is an empirical constant in the range of 0.05–0.20.

To achieve the best separation effect, we combine the separation loss in Eq. (6), the regularization term in Eq. (7), and the reconstruction loss in Eq. (3). The final loss function of WFAE is given as

$$L_{\text{WFAE}} = \lambda \cdot (L_{\text{separation}} + L_{\text{regularization}}) + L_{\text{reconstruction}}, \quad (8)$$

where the hyper-parameter λ weighs the contribution of speech separation and reconstruction.

4 Experiments

4.1 Datasets

Since there are no public datasets containing both mixed speech and the corresponding clean speech, we choose three public universal databases to construct two-talker mixtures for experimental evaluations. These three public universal databases are TIMIT (DARPA TIMIT acoustic-phonetic continuous speech dataset) (Garofolo et al., 1993), Mini LibriSpeech (a public speech corpus harvested from LibriSpeech) (Panayotov et al., 2015), and Common Voice CN (a Mandarin Chinese database).

1. TIMIT

The TIMIT dataset is an acoustic-phonetic continuous speech corpus, consisting of 6300 sentences from 630 speakers in eight major dialects of the United States. About 70% of speakers are male, and the remaining are female. This popular benchmark dataset for MSS has been widely used (Wang YN et al., 2016; Grais and Plumbley, 2017; Pandey et al., 2018). We construct mixtures (each containing one female speaker and one male speaker) from the TIMIT training set at various signal-to-noise ratios (SNRs) uniformly chosen between 0 and 5 dB. In other words, each utterance in the constructed training set is a mixture of female and male voices, and a total of 9520 mixtures are constructed. In the same way, the test set of TIMIT is used to construct the mixed sounds, and the mixtures are divided into a speaker-independent verification set (unseen speakers) and a speaker-independent test set (unseen speakers), each of which has about 560 mixed waveforms. In addition, we generate about 560 mixed waveforms in the speaker-dependent test set (seen speakers) for subsequent experimental comparison.

2. Mini LibriSpeech

The LibriSpeech dataset is a new corpus of read English speech, suitable for training and evaluating speech separation and recognition systems. The Mini LibriSpeech corpus with 54 speakers speaking

approximately 2600 utterances is derived from the LibriSpeech corpus. The fact that the Mini LibriSpeech is a small subset of the LibriSpeech corpus leads to an inconsistent data distribution between the Mini LibriSpeech training set and test set. So, we put together the data from the training set and test set, and then construct speech mixtures with SNRs selected uniformly between 0 and 5 dB. The mixtures are divided into 4000 mixtures for the training set, 500 mixtures for the validation set, and 500 mixtures for the test set.

3. Common Voice CN

Common Voice is a project to help make voice recognition open to everyone. It is Mozilla's initiative to help teach machines how real people speak. This dataset consists of acoustic waveforms in various languages such as English, German, French, and Chinese. We perform experiments on the Chinese dataset (Common Voice CN) to evaluate the proposed method. Mixtures are created in the same way as in the Mini LibriSpeech corpus.

4.2 Experimental setup

In our experiments, each sample is the stack (over 19 frames) of the 129-dimensional STFT spectral magnitude of the speech signal, computed with a sampling frequency of 8 kHz, a frame size of 32 ms, and a frame shift of 16 ms.

On the TIMIT database, we evaluate our method based on the source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifact ratio (SAR). These were proposed in Vincent et al. (2006). We compare our method with the benchmark models in Pandey et al. (2018): no-mask, binary-mask, soft-mask, autoencoder, VAE, and deep-VAE. We also compare the performances of our system in different cases: (1) with and without generative factors of the original input (deep-VAE and zAE-MLP); (2) with multilayer perception (MLP) or CNN structure in the encoder/decoder/constructor (zAE-MLP and zAE-CNN); (3) with and without a latent attention mechanism in the separation layer (zAE-CNN and WFAE-no_reg); (4) with and without the regularization term (WFAE-no_reg and WFAE). After carrying out the above experiments in the case of speaker-dependent sets, we show the corresponding results on a relatively challenging MSS case, speaker-independent MSS. On Mini LibriSpeech and Common Voice CN, we compare our method with deep-VAE (Pandey et al., 2018) since it is the best baseline model.

Specifically, no-mask, binary-mask, and soft-mask represent the type of masks used to train MSS as a conventional regression task. Autoencoder refers to compressing the mixed speech signal into low-dimensional information and constructing the speech of interest. VAE and deep-VAE replace the basic structure of the autoencoder with a variational autoencoder and a deep variational autoencoder, respectively, while keeping the operation of compression and construction unchanged. WFAE denotes the model proposed in this study. WFAE-no_reg optimizes the model directly without using the regularization term. zAE-CNN, instead of using the attention mechanism in the separation layer, uses the simple fully connected layer directly without the regularization term. The difference between zAE-MLP and zAE-CNN is the replacement of the CNN structure with the MLP structure in the encoder/decoder/constructor.

All the experiments are carried out with Ubuntu 16.04.9 LST, Python 3.5.5 with TensorFlow 1.8.0, and Keras 2.2.0 on an NVIDIA Tesla P100 GPU. The value of α in Eq. (7) is set to 0.05 to achieve SIR improvement and maintain SAR. This is derived from Huang et al. (2014). We perform a simple incremental search to choose the optimal number of units in generative factors (integers in [64, 1024]) and the hyper-parameter λ in Eq. (8) (an integer in [1, 5]). The pair of parameters (number of units = 256 and $\lambda = 3$) is chosen. This gives the best results on the validation set. We use a training principle similar to that in Kolbæk et al. (2017), which means that the maximum number of training epochs is set to 100 and that the learning rate is set to 0.0005 with the decay equaling 1e-6. We use the loss function in Eq. (8) on the validation set for early stopping, and the criterion for early stopping does no decrease in the loss on the validation set for 10 epochs.

4.3 Performance evaluation

4.3.1 Performance comparison on the TIMIT corpus

In this subsection, we first explain how to determine the hyper-parameter λ and the number of units in generative factors $\boldsymbol{z}$, and then evaluate the separation performance based on three important

metrics by comparing the methods of our system in four different cases with the baselines. We compare and analyze the results in the speaker-dependent and speaker-independent cases. In Table 1, the first half of the speaker-dependent results (no-mask, binary-mask, soft-mask, autoencoder, VAE, and deep-VAE) comes directly from Pandey et al. (2018). For fair comparison, we reproduce the two best baselines (VAE and deep-VAE) and conduct experiments on them with our created mixtures (seen speakers). Those results are listed in the second half of the speaker-dependent results. Results of our methods with our created mixtures (seen speakers and unseen speakers) are also shown in Table 1.

1. The number of units in generative factors and the hyper-parameter λ

As can been seen from Fig. 3a, the best results are obtained when the number of units is equal to 256, and too large or small values will lead to degradation of performance. It is easy to see from Fig. 3b that 3 is slightly more optimal than other values for λ. Thus, we choose this setting (units=256 and λ=3) in our experiments.

2. Speaker-dependent case

(1) Efficiency of the generative factors $\boldsymbol{z}$

The goal of this subsection is to analyze whether the generative factors obtained by reconstructing the given input can bring better separation performance. Autoencoder, VAE, and deep-VAE compress the input signal into compressed features, and then use these compressed features to construct the target source directly. In the method of zAE-MLP, the original input is reconstructed by the decoder to ensure that the compressed features capture the generative factors, and then the mixed speech is separated by the generative factors. From the results in Table 1, we can see that zAE-MLP outperforms the best baseline model deep-VAE with an improvement of about 2.21, 1.63, and 2.72 dB for SDR, SIR, and SAR, respectively. This shows that our proposed method can learn the source-specific information through generative factors $\boldsymbol{z}$, and thus achieve improvement in separating mixed speech.

(2) MLP vs. CNN in the encoder/decoder/constructor

In this subsection, we discuss the results of our methods with different networks (MLP or CNN) in the encoder/decoder/constructor. From the results in Table 1, we can see that SDR, SIR, and SAR increase at varying degrees when CNN is used in the encoder/decoder/constructor. The reason is that the common MLP can handle only one frame of speech signal at a time. If we need to treat a segment of speech signal once by MLP, the segment must be flattened into one-dimensional data. However, CNN can deal with a two-dimensional segment directly, so it can learn better features than an MLP. In addition, if MLP addresses one segment of speech at a time, then the model needs more parameters, which may lead to slow convergence. CNN can process a segment of speech signal at the same time with fewer parameters. Therefore, we can conclude that

Table 1 Performance comparison on TIMIT

Setting	Model	SDR (dB)	SIR (dB)	SAR (dB)
Speaker-dependent	No-mask	4.25	8.84	6.91
	Binary-mask	3.51	12.50	4.33
	Soft-mask	4.25	8.44	6.91
	Autoencoder	3.01	6.77	6.31
	VAE	6.03	8.80	7.02
	Deep-VAE	6.13	8.85	7.16
	VAE	5.84	10.32	8.21
	Deep-VAE	6.19	10.56	8.61
	$\boldsymbol{z}$AE-MLP	8.40	12.19	11.33
	$\boldsymbol{z}$AE-CNN	9.16	13.42	11.63
	WFAE-no_reg	9.43	13.79	**11.82**
	WFAE	**9.70**	**14.77**	11.69
Speaker-independent	$\boldsymbol{z}$AE-MLP	7.94	11.67	11.00
	$\boldsymbol{z}$AE-CNN	8.34	12.44	11.05
	WFAE-no_reg	8.67	12.88	**11.26**
	WFAE	**8.81**	**13.75**	10.99

The first half of the speaker-dependent results comes from Pandey et al. (2018), and the second half of the speaker-dependent results comes from implemented baselines and our models with our created mixtures (seen speakers). The speaker-independent results come from our models with created unseen speakers. The best results are in bold

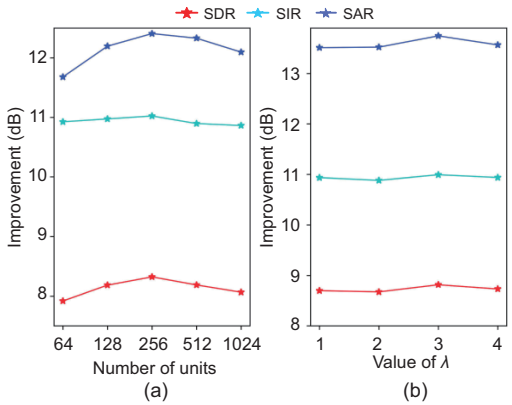


Fig. 3 Model ablation on the number of units in generative factors (a) and the hyper-parameter $\boldsymbol{\lambda}$ (b)

CNN is more suitable as the structure of the encoder/decoder, and is also more suitable for constructing the target source in the autoencoder and related architectures.

(3) Benefit of the latent attention mechanism and regularization term

A major contribution of our structure is to introduce the latent attention mechanism. With the attention mechanism, WFAE-no_reg can improve the SDR, SIR, and SAR over zAE-CNN, as shown in Table 1. This indicates that the attention mechanism helps the model learn discriminatively source-specific generative factors for speakers of different sexes, so that the mixed speech can be better separated. In addition, by adding a regularization term in the loss function, one source can be isolated from the other easily, resulting in higher SDR and SIR. Although SAR decreases a little compared with the case without the regularization term, it is in line with the results in Huang et al. (2014). Ultimately, WFAE significantly outperforms the deep-VAE approach with 3.51-dB SDR, 4.21-dB SIR, and 3.08-dB SAR improvements.

3. Speaker-independent case

We compare the methods of zAE-MLP, zAE-CNN, WFAE-no_reg, and WFAE in the speaker-independent case, and the results are shown in Table 1. In the speaker-independent case, our proposed WFAE achieves better performance than the other three methods. It is demonstrated that the latent source-specific generative factors learned by WFAE still work better than those learned by the other three methods in the speaker-independent case. In addition, the changing trend of the experimental results of these four methods in the speaker-independent case is similar to that in the speaker-dependent case. We can easily see that the three metrics of WFAE in the speaker-independent case are slightly lower than those of WFAE in the speaker-dependent case, which is also the case for zAE-MLP, zAE-CNN, and WFAE-no_reg. Importantly, all the results in the speaker-independent case are far higher than those of the best baseline deep-VAE in the speaker-dependent case. This indicates that WFAE has great success on the relatively challenging MSS case, namely, speaker-independent MSS.

4. Analysis of the training cost and practicality

The experimental setup of Pandey et al. (2018) combines the voices of male and female speakers and trains an individual model for each target speaker. That is to say, they have to train a large number of models. Then, the cost of computing resources is high, and it is difficult to apply the results to real-life scenarios. Different from Pandey et al. (2018), we drive our model to separate the mixed speech of female and male speakers with a single speaker-independent model. Compared with the scheme in Pandey et al. (2018), a lot of computational resources can be saved. Our proposed model is more suitable for real applications.

4.3.2 Performance comparison between deep-VAE and WFAE on Mini LibriSpeech and Common Voice CN

In addition, we conduct comparisons to evaluate the performance of the proposed model on the Mini LibriSpeech corpus. We reproduce the framework of the best baseline deep-VAE, and list SDR, SIR, and SAR improvements in Table 2. The results are consistent with those on the TIMIT dataset. However, as can be seen from Table 2, the performances of deep-VAE and WFAE on the Mini LibriSpeech corpus are both slightly inferior to that on the TIMIT dataset. The reason is that Mini LibriSpeech is derived from the LibriSpeech corpus, which leads to an inconsistent data distribution.

Experiments on Common Voice CN are performed to prove the generalization of the proposed method, and similar performance superiority can be seen in Table 3. For SDR, SIR, and SAR, the results of WFAE are 11.02, 16.90, and 12.54 dB, with improvements of 2.77, 1.98, and 3.10 dB over deep-VAE, respectively.

Table 2 SDR, SIR, and SAR results on Mini LibriSpeech

Model	SDR (dB)	SIR (dB)	SAR (dB)
Deep-VAE	5.86	12.35	7.29
WFAE	**7.76**	**13.29**	**9.66**

The best results are in bold

Table 3 SDR, SIR, and SAR results on Common Voice CN

Model	SDR (dB)	SIR (dB)	SAR (dB)
Deep-VAE	8.25	14.92	9.44
WFAE	**11.02**	**16.90**	**12.54**

The best results are in bold

4.3.3 Discriminative feature visualization

To gain intuitive understanding of the learned source-specific generative factors and discriminative features, we visualize and compare the distributions of the features learned by deep-VAE and WFAE. Fig. 4 shows the feature distributions for 750 samples on the TIMIT dataset. We clearly see that WFAE learns a discriminative and compact feature representation, ultimately leading to better speech separation performance. It is easier for WFAE to find a more straightforward boundary between the female features and male features than deep-VAE. Thus, the effectiveness of our method for MSS is clearly demonstrated.

4.3.4 Visualization of clean and separated speeches with WFAE

To further demonstrate the performance of WFAE, we randomly choose a set of separated results, including original mixed speech, reconstructed mixed speech, clean speeches, and separated speeches (Fig. 5). The waveforms in Fig. 5 clearly show that the reconstructed mixed speech is almost identical to the original mixed speech, which means that the compressed features capture the generative factors of mixed speech successfully. Comparing the separated speeches with the corresponding clean speeches, we can observe that the waveform of the separated speech 1 is nearly the same as that of clean speech 1. This also works for the separated speech 2 and clean speech 2. This demonstrates the separation capability of our model for MSS intuitively.

5 Conclusions and future work

Generative factor is an important research issue for MSS. In this study, we have proposed a novel model, WFAE, to learn source-specific generative factors and a set of discriminative features for each audio source. Compared with current autoencoder-based approaches, our proposed architecture has achieved superior speech separation performance with great SDR, SIR, and SAR improvements on a more difficult MSS case, speaker-independent MSS.

In our opinion, these improvements may be useful in other speech processing tasks such as speech enhancement and speech denoising. We believe that the WFAE framework can be extended to recurrent neural network (RNN) and convolutional RNN (CRNN) to obtain more discriminative features that also capture temporal information. We are currently working on this idea. In the future, we plan to extend the proposed method to separate gender-independent mixed speeches. Introducing time-domain separation methods into our model is also a direction worth studying.

Contributors

Jing-jing CHEN and Qi-rong MAO designed the research. Jing-jing CHEN processed the data. Jing-jing CHEN and Qi-rong MAO drafted the manuscript. You-cai QIN, Shuang-qing QIAN, and Zhi-shen ZHENG helped organize the manuscript. Jing-jing CHEN and Qi-rong MAO revised and finalized the paper.

Compliance with ethics guidelines

Jing-jing CHEN, Qi-rong MAO, You-cai QIN, Shuang-

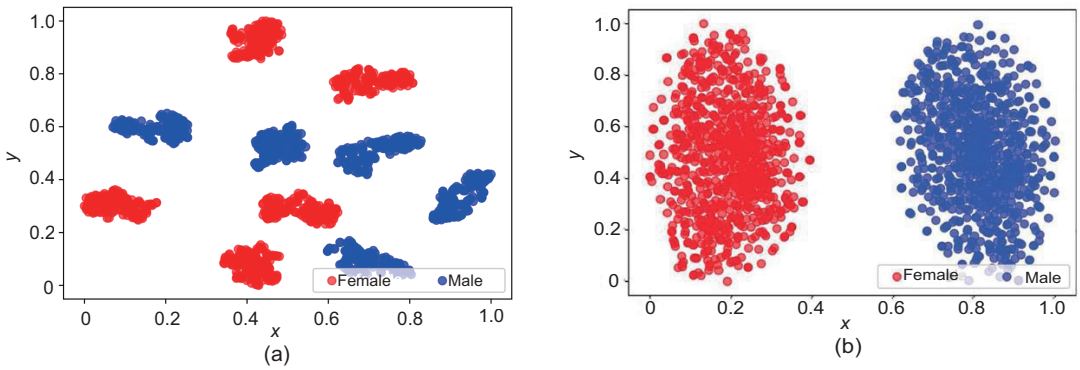


Fig. 4 Feature distributions from deep-VAE (a) and our WFAE method (b)

Features are extracted from the first hidden layer after the layer of generative factors and then mapped to a two-dimensional space using t-SNE (van der Maaten and Hinton, 2008)

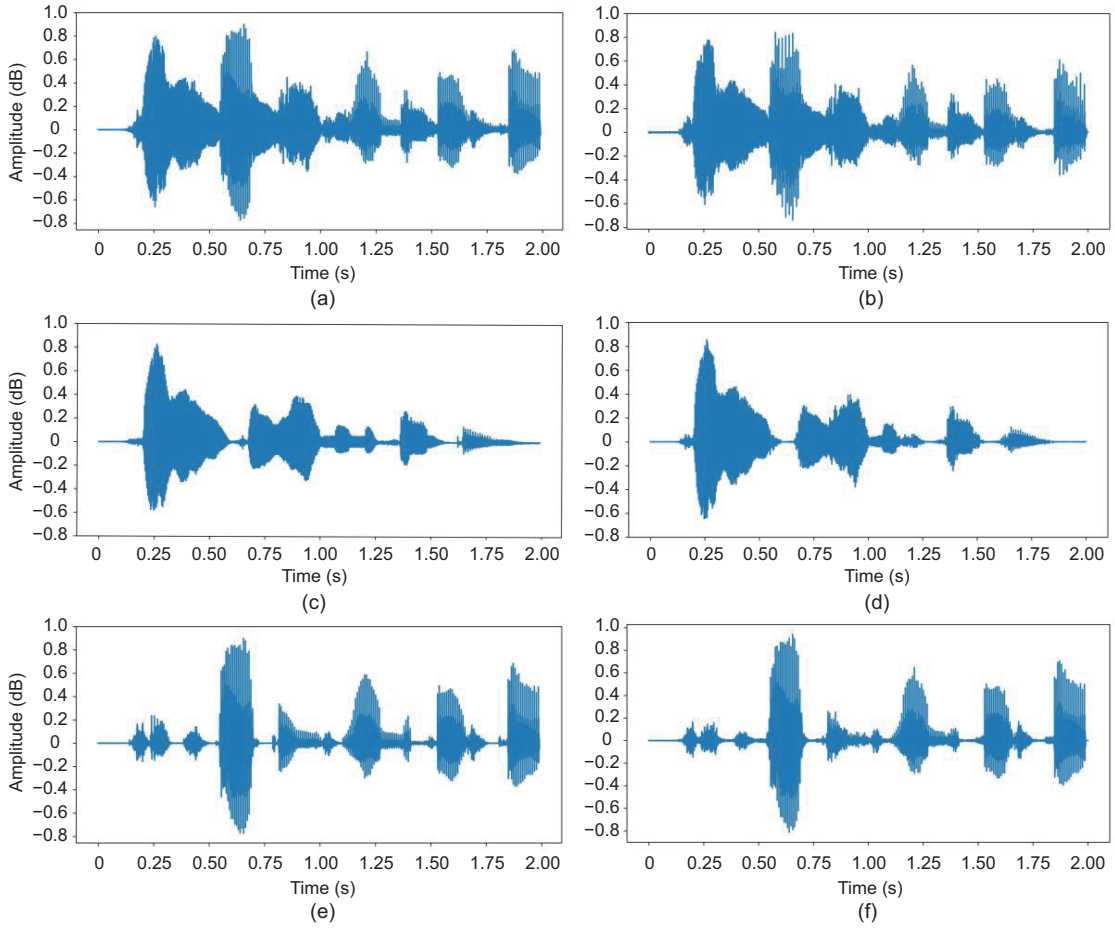


Fig. 5 Speech separation examples using the TIMIT dataset: (a) mixture of two speakers; (b) reconstructed mixed speech; (c) clean speech of speaker 1; (d) separated speech of speaker 1; (e) clean speech of speaker 2; (f) separated speech of speaker 2

qing QIAN, and Zhi-shen ZHENG declare that they have no conflict of interest.

References

Araki S, Sawada H, Mukai R, et al., 2007. Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. *Signal Process*, 87(8):1833-1847. https://doi.org/10.1016/j.sigpro.2007.02.003

Benesty J, Chen JD, Huang YT, 2008. Microphone Array Signal Processing. Springer, Berlin, Germany.

Bregman AS, 1990. Auditory Scene Analysis: the Perceptual Organization of Sound. The MIT Press, Cambridge.

Brown GJ, Cooke M, 1994. Computational auditory scene analysis. *Comput Speech Lang*, 8(4):297-336. https://doi.org/10.1006/csla.1994.1016

Chen Z, Luo Y, Mesgarani N, 2017. Deep attractor network for single-microphone speaker separation. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.246-250. https://doi.org/10.1109/ICASSP.2017.7952155

Erdogan H, Hershey JR, Watanabe S, et al., 2015. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.708-712. https://doi.org/10.1109/ICASSP.2015.7178061

Garofolo JS, Lamel LF, Fisher WM, et al., 1993. DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1. NASA STI/Recon Technical Report, NASA, USA.

Ghahramani Z, Jordan MI, 1997. Factorial hidden Markov models. *Mach Learn*, 29(2-3):245-273. https://doi.org/10.1023/A:1007425814087

Gou JP, Yi Z, Zhang D, et al., 2018. Sparsity and geometry preserving graph embedding for dimensionality reduction. *IEEE Access*, 6:75748-75766. https://doi.org/10.1109/ACCESS.2018.2884027

Grais EM, Plumbley MD, 2017. Single channel audio source separation using convolutional denoising autoencoders. Proc IEEE Global Conf on Signal and Information Processing, p.1265-1269. https://doi.org/10.1109/GlobalSIP.2017.8309164

Hershey JR, Chen Z, Le Roux J, et al., 2016. Deep clustering: discriminative embeddings for segmentation and separation. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.31-35. https://doi.org/10.1109/ICASSP.2016.7471631

Hsu WN, Zhang Y, Glass J, 2017. Learning latent representations for speech generation and transformation. 18th Annual Conf of the Int Speech Communication Association, p.1273-1277.

Hu K, Wang DL, 2013. An unsupervised approach to cochannel speech separation. *IEEE Trans Audio Speech Lang Process*, 21(1):122-131. https://doi.org/10.1109/TASL.2012.2215591

Huang PS, Kim M, Hasegawa-Johnson M, et al., 2014. Deep learning for monaural speech separation. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.1562-1566. https://doi.org/10.1109/ICASSP.2014.6853860

Hyvärinen A, Oja E, 2000. Independent component analysis: algorithms and applications. *Neur Netw*, 13(4-5):411-430. https://doi.org/10.1016/S0893-6080(00)00026-5

Karamatlı E, Cemgil AT, Kırbız S, 2019. Weak label supervision for monaural source separation using non-negative denoising variational autoencoders. Proc 27th Signal Processing and Communications Applications Conf, p.1-4. https://doi.org/10.1109/SIU.2019.8806536

Kolbæk M, Yu D, Tan ZH, et al., 2017. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Trans Audio Speech Lang Process*, 25(10):1901-1913. https://doi.org/10.1109/TASLP.2017.2726762

Luo Y, Mesgarani N, 2019. Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Trans Audio Speech Lang Process*, 27(8):1256-1266. https://doi.org/10.1109/TASLP.2019.2915167

Luo Y, Chen Z, Yoshioka T, 2019. Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. https://arxiv.org/abs/1910.06379

Nadas A, Nahamoo D, Picheny MA, 1989. Speech recognition using noise-adaptive prototypes. *IEEE Trans Acoust Speech Signal Process*, 37(10):1495-1503. https://doi.org/10.1109/29.35387

Osako K, Mitsufuji Y, Singh R, et al., 2017. Supervised monaural source separation based on autoencoders. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.11-15. https://doi.org/10.1109/ICASSP.2017.7951788

Panayotov V, Chen GG, Povey D, et al., 2015. LibriSpeech: an ASR corpus based on public domain audio books. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.5206-5210. https://doi.org/10.1109/ICASSP.2015.7178964

Pandey L, Kumar A, Namboodiri V, 2018. Monaural audio source separation using variational autoencoders. Proc Interspeech, p.3489-3493. https://doi.org/10.21437/Interspeech.2018-1140

Qian YM, Weng C, Chang XK, et al., 2018. Past review, current progress, and challenges ahead on the cocktail party problem. *Front Inform Technol Electron Eng*, 19(1):40-63. https://doi.org/10.1631/FITEE.1700814

Radford A, Metz L, Chintala S, 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. https://arxiv.org/abs/1511.06434

Roweis ST, 2001. One microphone source separation. Proc 13th Int Conf on Neural Information Processing Systems, p.793-799.

Schmidt MN, Olsson RK, 2006. Single-channel speech separation using sparse non-negative matrix factorization. Proc 9th Int Conf on Spoken Language Processing.

Smaragdis P, 2007. Convolutive speech bases and their application to supervised speech separation. *IEEE Trans Audio Speech Lang Process*, 15(1):1-12. https://doi.org/10.1109/TASL.2006.876726

van der Maaten L, Hinton G, 2008. Visualizing data using t-SNE. *J Mach Learn Res*, 9(11):2579-2605.

Vincent E, Gribonval R, Fevotte C, 2006. Performance measurement in blind audio source separation. *IEEE Trans Audio Speech Lang Process*, 14(4):1462-1469. https://doi.org/10.1109/TSA.2005.858005

Wang DL, Brown GJ, 2006. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press, Hoboken, USA.

Wang YN, Du J, Dai LR, et al., 2016. Unsupervised single-channel speech separation via deep neural network for different gender mixtures. Asia-Pacific Signal and Information Processing Association Annual Summit and Conf, p.1-4. https://doi.org/10.1109/APSIPA.2016.7820736

Wang YX, Narayanan A, Wang DL, 2014. On training targets for supervised speech separation. *IEEE/ACM Trans Audio Speech Lang Process*, 22(12):1849-1858. https://doi.org/10.1109/TASLP.2014.2352935

Williamson DS, 2018. Monaural speech separation using a phase-aware deep denoising auto encoder. Proc IEEE 28th Int Workshop on Machine Learning for Signal Processing, p.1-6. https://doi.org/10.1109/MLSP.2018.8516918

Xia LM, Wang H, Guo WT, 2019. Gait recognition based on Wasserstein generating adversarial image inpainting network. *J Cent South Univ*, 26(10):2759-2770. https://doi.org/10.1007/s11771-019-4211-7

Yu D, Kolbæk M, Tan ZH, et al., 2017. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.241-245. https://doi.org/10.1109/ICASSP.2017.7952154

Zhang QJ, Zhang L, 2018. Convolutional adaptive denoising autoencoders for hierarchical feature extraction. *Front Comput Sci*, 12(6):1140-1148. https://doi.org/10.1007/s11704-016-6107-0