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

Review:

Past review, current progress, and challenges ahead on the cocktail party problem*

Yan-min QIAN†‡[1], Chao WENG[1], Xuan-kai CHANG[2], Shuai WANG[2], Dong YU[1]

[1]*Tencent AI Lab, Tencent, Bellevue 98004, USA*

[2]*Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*

†E-mail: yanminqian@tencent.com

Received Dec. 8, 2017; Revision accepted Jan. 17, 2018; Crosschecked Jan. 25, 2018

Abstract: The cocktail party problem, i.e., tracing and recognizing the speech of a specific speaker when multiple speakers talk simultaneously, is one of the critical problems yet to be solved to enable the wide application of automatic speech recognition (ASR) systems. In this overview paper, we review the techniques proposed in the last two decades in attacking this problem. We focus our discussions on the speech separation problem given its central role in the cocktail party environment, and describe the conventional single-channel techniques such as computational auditory scene analysis (CASA), non-negative matrix factorization (NMF) and generative models, the conventional multi-channel techniques such as beamforming and multi-channel blind source separation, and the newly developed deep learning-based techniques, such as deep clustering (DPCL), the deep attractor network (DANet), and permutation invariant training (PIT). We also present techniques developed to improve ASR accuracy and speaker identification in the cocktail party environment. We argue effectively exploiting information in the microphone array, the acoustic training set, and the language itself using a more powerful model. Better optimization objective and techniques will be the approach to solving the cocktail party problem.

Key words: Cocktail party problem; Computational auditory scene analysis; Non-negative matrix factorization; Permutation invariant training; Multi-talker speech processing

https://doi.org/10.1631/FITEE.1700814 **CLC number:** TP391.4

1 Introduction

Although the accuracy of automatic speech recognition (ASR) systems has surpassed the threshold for adoption for many real-world applications (Hinton et al., 2012; Abdel-Hamid et al., 2014; Yu and Deng, 2014; Bi et al., 2015; Peddinti et al., 2015; Sainath et al., 2015; Qian et al., 2016; Sercu et al., 2016; Xiong et al., 2016; Yu and Li, 2017), there are still difficulties to be solved to make ASR systems more robust and more widely deployed (Qian et al., 2018). The cocktail party problem, i.e., tracing and recognizing the speech from a specific speaker when multiple speakers talk simultaneously and when other background noise is involved, is one such problem. The cocktail party problem has been widely observed. Solving it could enable many scenarios and applications, such as meeting transcription, multi-party human–machine interaction, and hearing impairment assistants, where overlapped speech cannot be ignored.

There is a long history of research on the cocktail party problem (Cherry, 1953; Wang and Brown, 2006; Kolbæk et al., 2017a; Yu et al., 2017b). Although the processing mechanisms seem clear and related tasks are easy for humans, researchers have found it surprisingly difficult to give machines the same ability. Although many approaches

‡ Corresponding author

* Project supported by the Tencent and Shanghai Jiao Tong University Joint Project

ORCID: Yan-min QIAN, http://orcid.org/0000-0002-0314-3790

© Zhejiang University and Springer-Verlag GmbH Germany, part of Springer Nature 2018

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

were proposed and attempted in the early days, including those based on signal processing techniques (Ephraim and Malah, 1985; Hu and Loizou, 2007, 2008), computational auditory scene analysis (CASA) (Brown and Cooke, 1994; Ellis, 1996; Wang and Brown, 2006), non-negative matrix factorization (NMF) (Raj et al., 2010; Schuller et al., 2010; Chen et al., 2014), and microphone array techniques (Fischer and Simmer, 1996; Kellermann, 1997; Anguera et al., 2007; Benesty et al., 2007), a few of these approaches achieved robust performance with a high separation quality, especially when only a single channel of the mixed signal is available or the speakers are facing the same direction.

Inspired by the great success of deep learning in speech recognition (Sainath et al., 2013; Xiong et al., 2016; Yu et al., 2016) and speaker identification (Lei et al., 2014; Variani et al., 2014; Liu et al., 2015), deep learning-based techniques have been developed recently to address the cocktail party problem. These new techniques significantly outperform the conventional approaches, and performance improvements are particularly impressive for recent techniques such as deep clustering (DPCL) (Hershey et al., 2016), the deep attractor network (DANet) (Chen et al., 2017b), and permutation invariant training (PIT) (Yu et al., 2017a,b). The preliminary success ignites new hope and provides important stepping stones towards eventually solving the cocktail party problem.

This paper aims to provide a comprehensive survey of the popular and effective solutions to the cocktail party problem developed in the past two decades. We focus on the recent progress achieved with deep learning technologies and the remaining difficulties and challenges ahead. We hope this survey can help readers become familiar with this active research area, and gain insights into the possible research directions for addressing this interesting and important problem.

2 Cocktail party problem

Natural auditory environments, such as cocktail parties, usually contain many concurrently existing sounds, including speech signals from multiple speakers and other sounds such as music and instruments. The cocktail party problem is the task of separating these mixed sounds and paying attention to only one or two sounds of interest, often speech signals, in such complex auditory environments (Fig. 1). The cocktail party problem is quite interesting yet difficult to solve. Although there is a long history of research on how humans behave in the cocktail party environment and many attempts have been make to develop computer algorithms to match a machine's ability to that of humans in such environments, the cocktail party problem remains a challenge to be solved to enable a truly free conversation between humans and computers.

Fig. 1 A typical cocktail party scene (image from Daniel Hagerman: *High Society Cocktail Party—End of Prohibition 1933*)

Although the cocktail party problem is difficult for computers, it seems to be easy for humans. Humans can separate a signal consisting of multiple sources and attend to recognize one single source (Mesgarani and Chang, 2012; Chen, 2017). For instance, at a typical cocktail party, people can easily concentrate on the speech of the conversational talkers, the song from the singers, or the melody from the musical instruments. Mesgarani and Chang (2012) conducted a research on the cortical representation of multi-talker mixed speech, and concluded that the human auditory system restores the representation of speaker of interest while suppressing irrelevant competing speech. In fact, this ability exists in not only humans but also other species. For example, animals can easily identify the sounds from mates or enemies in crowded environments where many animals vocalize at the same time (McDermott, 2009).

To match a computer's ability to that of humans and animals in the cocktail party environment, we need to attack two distinct challenges. The first challenge is how to separate sounds from the mixed

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

signal, which is the sum of all sounds in the complex auditory scene. Humans are typically interested in and capable of concentrating on only one or two sound sources at the same time and thus need only to separate these sounds from the mixture. However, computers can multi-task, and thus it is desirable to separate all sound sources from the mixture. The second challenge, which is very important in multi-talker conversation, is how to trace and hold attention to the sound of interest source and switch attention among sources. In most cases, these two challenges are intertwined: the attention to the target source of interest can benefit from good separation and the separation can benefit from speaker tracing.

The term 'cocktail party problem' was coined in Cherry's classic paper (Cherry, 1953). This paper studied whether humans can select one speech signal over another, whether they retain anything about the non-selected signal, and how they can switch their attention between signals. About four decades later, Bregman (1990) began studying sound segregation, termed 'auditory scene analysis'. In fact, most of the past and current work on the cocktail party problem focused on the first challenge (Du et al., 2014; Xu et al., 2014; Wang et al., 2014; Weninger et al., 2015; Chen, 2017), i.e., sound segregation, which is also the main focus of this paper.

To evaluate the performance of the solution to the cocktail party problem, many metrics have been proposed to measure the ability of sound separation (usually speech separation) and target source attention (usually target speaker tracing). For example, for the speech separation task, the metrics for speech quality, such as perceptual evaluation of speech quality (PESQ) (Rix et al., 2001), source-to-noise ratio (SNR), source-to-distortion ratio (SDR), source-to-artifacts ratio (SAR) (Vincent et al., 2006), and short-time objective intelligibility (STOI) (Taal et al., 2010), are commonly used. In some scenarios, the performance measurement is task dependent. For example, in the multi-talker speech recognition task, speech separation is just an intermediate step and the essential metric of the system is the recognition accuracy measured with, e.g., the word error rate (WER). In the multi-talker speaker identification task, the equal error rate (EER) is often used to evaluate the performance of the solution in the cocktail party environment.

Although researchers have not achieved a solution yet, many technologies have been proposed to attack the cocktail party problem over the past two decades. In Sections 3–7 we will review the most popular ones.

3 Conventional single-channel techniques

3.1 Computational auditory scene analysis

Although speech separation has proved to be difficult for computers, it is remarkably easy for the human auditory system. An obvious idea is to study how humans separate speech and learn from them. CASA follows this idea exactly.

In psychoacoustic research, the perceptual process of separating mixtures of sound sources is called 'auditory scene analysis (ASA)' (Bregman, 1990). Research in ASA has inspired CASA (Hu and Wang, 2004; Wang, 2005; Wang and Brown, 2006), in which certain segmentation rules based on perceptual grouping cues are (often semi-manually) designed to operate on low-level features to estimate a time–frequency (T-F) mask that isolates the signal components belonging to different speakers. This mask is then used to reconstruct the signal. For example, natural speech contains both voiced and unvoiced portions, and voiced portions account for about 75%–80% of spoken English (Hu and Wang, 2008). Because voiced speech is characterized by periodicity (or harmonicity), harmonicity has been used as a primary cue in many CASA systems for segregating voiced speech (Brown and Cooke, 1994).

Although CASA was proposed more than a decade ago, techniques based on the same principles are still being developed. Hu and Wang (2010) used a tandem algorithm to generate multiple simultaneous speech streams, and then grouped them sequentially by maximizing a joint speaker recognition score where speakers are described with Gaussian mixture models (GMMs). Hu and Wang (2013) proposed to use the information from a co-channel signal to improve the segmentation and grouping in CASA. An input scene is decomposed into T-F segments, each of which originates primarily from a single sound source. Grouping selectively aggregates segments to form streams corresponding to sound sources. Both

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

simultaneous and sequential grouping techniques are used. Simultaneous grouping organizes sound components across frequencies to produce simultaneous streams, and sequential grouping links them across time to form final sound streams.

Although CASA simulates the high-level behavior of human listening, it suffers from many drawbacks. First, it works on only speech and may fail in the broader perspective of audio source separation. Second, most of the rules are manually designed based on a limited number of observations and generalize poorly. Third, since the final separation is based on T-F segmentation (i.e., each T-F bin belongs to only one sound source), the best possible result is agreement with the oracle binary mask, which has been shown to be suboptimal in most scenarios (Wang, 2005; Kjems et al., 2009). Fourth, the entire system heavily depends on the accuracy of the pitch tracker, which is not robust under complex acoustic conditions. Fifth, it is limited because it cannot learn from data automatically.

3.2 Non-negative matrix factorization

In CASA, the T-F bins are grouped together based mainly on the hand-designed rules from human observations. To find the complex inherent characteristics from data, the data-driven methods were proposed. NMF (Lee and Seung, 2001), along with other matrix decomposition models, was built based on the assumption that the audio spectrogram has a low rank structure that can be represented with a small number of bases. Under certain conditions, the decomposition in NMF is unique and no other orthogonality or independence assumptions are needed. Specifically, in NMF,

$$\boldsymbol{Y} = \sum_s \boldsymbol{W}_s \boldsymbol{H}_s, \tag{1}$$

where each source s is modeled by the low rank approximation with non-negative matrices $\boldsymbol{W}_s$ and $\boldsymbol{H}_s$ and then summed to form mixture $\boldsymbol{Y}$. Because of the non-negativity of the decomposition matrices, there is no cancellation between sources in the reconstruction of mixture spectra $\boldsymbol{Y}$, which models the additivity between mixed sources.

Fig. 2 illustrates the basic NMF process. In the training stage, each clean source, e.g., speech, noise, and music, is decomposed and mapped into a set of bases and activations, and a source-specific

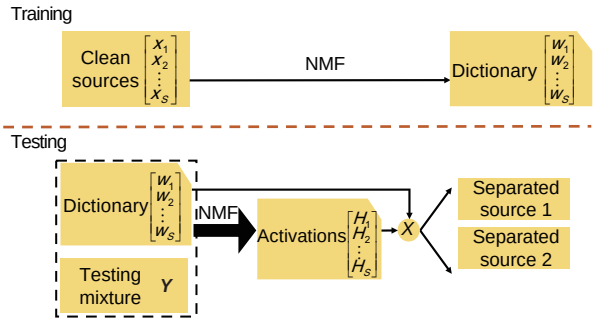

Fig. 2 The training phase where a dictionary set is learned for each individual source, and the testing phase where activation is inferred from non-negative matrix factorization, which is then used to reconstruct the source signals, given the dictionary and testing data

dictionary $\boldsymbol{W}$ is formed. During the testing stage, all the source-specific dictionaries learned are merged into a combined dictionary. This combined dictionary is fixed and only activation $\boldsymbol{H}$ is optimized for each source, in which case the optimization is convex and a global optimum can be achieved. Each source in the mixture is then reconstructed by the bases and the corresponding activations. The basic NMF algorithm is

$$\min_{\boldsymbol{W},\boldsymbol{H}} D(\boldsymbol{Y} \| \boldsymbol{W}\boldsymbol{H}) \tag{2}$$

$$\text{s.t. } \boldsymbol{W}, \boldsymbol{H} \geq 0. \tag{3}$$

Several variations of the NMF methods have been proposed. For example, the sparse NMF (Hoyer, 2004; Schmidt and Olsson, 2006; Virtanen, 2007) forces activation $\boldsymbol{H}$ to be sparse. In the convolutional NMF (Behnke, 2003; Bello, 2010; Chen et al., 2014), the spectrogram is decomposed into the convolution (instead of multiplication) of the basis and the activation. The robust NMF (Zhang et al., 2011; Chen and Ellis, 2013) combines NMF with robust principal component analysis.

The success of NMF is limited by a few facts. First, it is limited by the basis. Other attributes and regularities (e.g., temporal dynamics) of speech signals are not exploited. Second, the power of the model is limited by its linear system formulation, which prevents it from achieving a high separation quality. Third, the complexity of the decomposition during testing is expensive, limiting its application in real-time scenarios. Fourth, the size of the model parameters is determined by, and increases linearly with, the number of clean sources in the training set. This deteriorates its effectiveness in using a large

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

training set. Fifth, during testing, each source has to have a dictionary learned during the training stage (i.e., the source is included in the training set), which is not feasible in most real-world applications.

3.3 Generative models

NMF cannot model temporal dynamics. To address this limitation, several studies have been conducted (Kristjansson et al., 2006; Virtanen, 2006; Hershey et al., 2007; Cooke et al., 2010; Hershey et al., 2010; Rennie et al., 2010), most of which are based on the Gaussian mixture model-hidden Markov model (GMM-HMM) framework, a popular generative model in single-talker speech recognition. Among all these GMM-HMM separation models, the factorial hidden Markov model (FHMM) (Ghahramani and Jordan, 1996) is the most interesting and performs best. In FHMM, each source signal is modeled with an HMM trained on the data for that source. For each signal source s, if we define the clean signal as $\{\boldsymbol{x}_t^s\}$ ($t\in\{1,2,\cdots,T\}$), hidden states as $\{\boldsymbol{v}_t^s\}$, and the discrete mixture state as $\{\boldsymbol{m}_t^s\}$, HMM has the characteristics

$$p(\boldsymbol{v}_t^s|\boldsymbol{v}_{1:t-1}^s)=p(\boldsymbol{v}_t^s|\boldsymbol{v}_{t-1}^s), \tag{4}$$

$$\begin{aligned}p(\boldsymbol{x}_t^s|\boldsymbol{v}_{1:T}^s)&=p(\boldsymbol{x}_t^s|\boldsymbol{v}_t^s)\\&=\sum_{\boldsymbol{s}_t^s}p(\boldsymbol{x}_t^s|\boldsymbol{m}_t^s)p(\boldsymbol{m}_t^s|\boldsymbol{v}_t^s),\end{aligned} \tag{5}$$

where Eq. (4) describes the transition probability and Eq. (5) describes the observation probability under the Markov independence assumption. Given the mixed signal $\{\boldsymbol{y}_t\}$ of S signal sources, the new generative model, called the 'interaction model', can be defined as

$$\begin{aligned}&p(\{\boldsymbol{y}_t\},\{\boldsymbol{x}_t\},\{\boldsymbol{m}_t\},\{\boldsymbol{v}_t\})\\=&\prod_{t=1}^{\mathrm{T}}p(\boldsymbol{y}_t|\{\boldsymbol{x}_{(t)}^s\})\\&\cdot\prod_{t=1}^{\mathrm{T}}\prod_{s=1}^{S}p(\boldsymbol{x}_t^s|\boldsymbol{m}_t^s)p(\boldsymbol{m}_t^s|\boldsymbol{v}_t^s)p(\boldsymbol{v}_t^s|\boldsymbol{v}_{t-1}^s),\end{aligned} \tag{6}$$

where $\{\boldsymbol{x}_t^s\}$ is not observable.

The process of inferring hidden state sequence $\{\hat{\boldsymbol{v}}_t^{(s)}\}$ for each source s using the maximum a posterior (MAP) criterion requires computing $p(\boldsymbol{y}_t|\{\boldsymbol{v}_{(t)}^s\})$ as

$$\begin{aligned}&p(\boldsymbol{y}_t|\{\boldsymbol{v}_{(t)}^s\})\\&=\sum_{\boldsymbol{m}_t^1,\boldsymbol{m}_t^2,\cdots,\boldsymbol{m}_t^S}p(\boldsymbol{y}_t|\{\boldsymbol{m}_{(t)}^i\})\prod_s p(\boldsymbol{m}_t^s|\boldsymbol{v}_t^s)\\&=\sum_{\{\boldsymbol{m}_{(t)}^i\}}p(\boldsymbol{y}_t|\{\boldsymbol{m}_{(t)}^i\})\prod_s p(\boldsymbol{m}_t^s|\boldsymbol{v}_t^s),\end{aligned} \tag{7}$$

where

$$\begin{aligned}&p(\boldsymbol{y}_t|\{\boldsymbol{m}_{(t)}^s\})\\&=\int\int\cdots\int p(\boldsymbol{y}_t,\{\boldsymbol{x}_{(t)}^s\}|\{\boldsymbol{m}_{(t)}^s\})\mathrm{d}\boldsymbol{x}_t^1\mathrm{d}\boldsymbol{x}_t^2\cdots\mathrm{d}\boldsymbol{x}_t^S.\end{aligned} \tag{8}$$

$p(\boldsymbol{y}_t|\{\boldsymbol{v}_{(t)}^s\})$ does not factor over the speakers. The exact MAP state sequences of the speakers must be jointly estimated.

To reconstruct the features of source s at time t, the posterior expected value needs to be computed as

$$\begin{aligned}E(\boldsymbol{x}_t^s|\boldsymbol{y}_t,\{\hat{\boldsymbol{v}}_{(t)}^i\})=&\sum_{\{\boldsymbol{m}_{(t)}^i\}}p(\{\boldsymbol{m}_{(t)}\}^i|\boldsymbol{y}_t,\{\hat{\boldsymbol{v}}_{(t)}^i\})\\&\cdot E(\boldsymbol{x}_t^s|\boldsymbol{y}_t,\{\boldsymbol{m}_{(t)}^i\}),\end{aligned} \tag{9}$$

where

$$\begin{aligned}&E(\boldsymbol{x}_t^s|\boldsymbol{y}_t,\{\boldsymbol{m}_{(t)}^i\})\\&=\int\int\cdots\int\boldsymbol{x}_t^s p(\{\boldsymbol{x}_{(t)}^i\}|\boldsymbol{y}_t,\{\boldsymbol{m}_{(t)}^i\})\mathrm{d}\boldsymbol{x}_t^1\mathrm{d}\boldsymbol{x}_t^2\cdots\mathrm{d}\boldsymbol{x}_t^S.\end{aligned} \tag{10}$$

The computation process is very complicated and intractable because all these estimates are coupled over the states of the speakers. Several approximations for the interaction function have been developed to allow the integral in Eq. (10) to be computed analytically. The computation process can be divided into two parts, i.e., computing acoustic state likelihoods $p(\boldsymbol{y}_t|\{\boldsymbol{m}_{(t)}^s\})$ and combining these likelihoods to infer the MAP configuration of dynamic state variables $\{\hat{\boldsymbol{v}}_t^s\}$. The former part includes approximation using the log-sum model and the max-model, and the latter part includes loopy belief propagation.

Table 1 compares FHMM with other conventional techniques on the 2006 two-talker speech separation and recognition challenge (SSC) task (Cooke et al., 2010). All generative models outperform CASA and NMF. Among the generative models, FHMM (Hershey et al., 2010) performs the best

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

and even surpassed human listeners on this task (which is not a natural task for humans). Additional details can be found in Rennie et al. (2010) and Cooke et al. (2010).

Table 1 Word error rate (WER) on the 0-dB portion of the 2006 two-talker speech separation and recognition challenge (SSC) task (Cooke et al., 2010)

Algorithm	WER (%)
ALGONQUIN (log-sum model) (Hershey et al., 2010)	**22.7**
Max-model (Hershey et al., 2010)	**23.7**
Humans (Cooke et al., 2010)	28.5
PMC iterative Viterbi (log-sum model) (Virtanen, 2006)	35.1
CASA and fragment decoding (Barker et al., 2010)	38.2
NMF (Schmidt and Olsson, 2006)	44.2

Bold numbers indicate results exceeding human performance. PMC: parallel model combination; CASA: computational auditory scene analysis; NMF: non-negative matrix factorization

Although FHMM shows promising performance, it has several limitations that prevent it from being used in real-world applications. First, the computation cost during inference is very high, especially with an increased number of speakers, even with the approximate inference technique. Second, the interaction model becomes exponentially more complex when the number of speakers in the mixed signal increases. Third, because each speaker in the test set needs to have an HMM model, FHMM cannot effectively handle speakers and acoustic environments unseen in the training set.

4 Conventional multi-channel techniques

The techniques discussed in Section 3 require only a single channel of the mixed signal. The speech separation and tracing are carried out based on manually designed rules or joint distributions and models learned from the training data. However, if we can access multiple channels of the mixed signal, just as each person has two ears, we can exploit spatial sound source information of the sound source to improve the performance in the cocktail party environment.

Conventionally, there are two main categories of multi-channel techniques in dealing with the cocktail party problem: beamforming and multi-channel blind source separation.

4.1 Conventional beamforming approaches

A beamformer is a spatial filter that operates on the outputs of a microphone array and forms a beam (directivity) pattern to enhance the desired speech coming from one direction when suppressing interfering speech or noise from other directions. Such a spatial filtering operation can be divided into two subprocesses, synchronization and weight-and-sum (Benesty et al., 2008). According to the time difference of arrival (TDOA) information, the synchronization subprocess introduces a proper amount of delay to each channel to align the signal components coming from the desired direction. The weight-and-sum subprocess is to weight aligned signals and then add them to form one channel output. The synchronization controls the steering direction, and the weight-and-sum process controls the beamwidth of the mainlobe and the characteristics of the sidelobes. Conventional beamforming can be divided broadly into two categories. A fixed beamformer, as its name indicates, uses a fixed set of weighting coefficients and time delays once the array geometry and the desired steering direction are determined, while an adaptive beamformer automatically adapts its coefficients to different situations based on the characteristics of signal and noise.

4.1.1 Fixed beamforming

Delay-and-sum is the most widely known fixed beamforming approach. Denoted by $\boldsymbol{x}(t)$, the source signal at time t, the received signals by each of N microphones in the array can be expressed as

$$\begin{aligned}\boldsymbol{y}_n(t) &= a_n\boldsymbol{x}[t-\tau_n]+\boldsymbol{v}_n(t),\\ &= \boldsymbol{x}_n(t)+\boldsymbol{v}_n(t),\ n=1,2,...,N,\end{aligned} \tag{11}$$

where a_n and τ_n are attenuation factors and delays due to the propagation, respectively, and $\boldsymbol{v}_n(t)$ is the interfering speech or ambient noise signals. The first step of delay-and-sum is to introduce proper time shifts to align the microphone signals that correspond to the TDOA estimation. Without loss of generality, we consider the first microphone signal as the reference; i.e., after applying time shifts, the

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

aligned array output signals will be

$$\begin{aligned}\boldsymbol{y}_n^{\text{align}}(t) &= a_n\boldsymbol{x}[t-\tau_n+\Delta\tau_n]+\boldsymbol{v}_n(t+\Delta\tau_n)\\ &= a_n\boldsymbol{x}[t-\tau_1]+\boldsymbol{v}_n(t+\Delta\tau_n)\\ &= \boldsymbol{x}_n^{\text{align}}(t)+\boldsymbol{v}_n^{\text{align}}(t),\ n=1,2,...,N,\end{aligned} \tag{12}$$

Then the second step will sum up the aligned signals:

$$\begin{aligned}\boldsymbol{z}(t) &= \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{y}_n^{\text{align}}(t)\\ &= a_s\boldsymbol{x}[t-\tau_1]+\boldsymbol{v}_s(t),\end{aligned} \tag{13}$$

where

$$\begin{aligned}a_s &= \frac{1}{N}\sum_{n=1}^{N}a_n,\\ \boldsymbol{v}_s(t) &= \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{v}_n(t+\Delta\tau_n).\end{aligned} \tag{14}$$

Assuming all the signals are zero-mean and stationary, SNRs for reference input and summed output signals can be written as

$$\text{SNR}_{\text{in}} = \frac{\sigma_{\boldsymbol{x}_1}^2}{\sigma_{\boldsymbol{v}_1}^2} = a_1^2\frac{\sigma_s^2}{\sigma_{\boldsymbol{v}_1}^2}, \tag{15}$$

$$\text{SNR}_{\text{out}} = N^2a_s^2\frac{\sigma_s^2}{\sigma_{\boldsymbol{v}_s}^2} = \left(\sum_{n=1}^{N}a_n\right)^2\frac{\sigma_s^2}{\sigma_{\boldsymbol{v}_s}^2}. \tag{16}$$

With a few derivations and under the assumptions that all the interfering signals have the same energy and $\forall n, a_n = 1$, Benesty et al. (2008) showed that the relationship between input and output SNRs can be concisely expressed as

$$\text{SNR}_{\text{out}} = \frac{N}{1+\rho}\text{SNR}_{\text{in}}, \tag{17}$$

where $\rho = \dfrac{2}{N}\sum\limits_{i=1}^{N-1}\sum\limits_{j=i+1}^{N}\rho_{\boldsymbol{v}_i\boldsymbol{v}_j}$, and $\rho_{\boldsymbol{v}_i\boldsymbol{v}_j}$ is the correlation coefficient between $\boldsymbol{v}_i(t)$ and $\boldsymbol{v}_j(t)$. If $\boldsymbol{v}_n(t)$ are uncorrelated, we will have $\text{SNR}_{\text{out}} = N\cdot\text{SNR}_{\text{in}}$. If $\boldsymbol{v}_n(t)$ is completely correlated (i.e, $\rho_{\boldsymbol{v}_i\boldsymbol{v}_j}=1$ and $\rho = N-1$), we will not have any SNR gains from delay-and-sum beamforming.

One issue of a delay-and-sum beamformer is its narrowband nature; i.e., the beam pattern will change in different frequencies while speech is a broadband signal. A natural extension of a delay-and-sum beamformer is the filter-and-sum beamforming techniques first proposed by Frost (1972). Instead of introducing only a time shift and attenuation scale to each channel before summing up the signals, filter-and-sum applies a finite impulse response (FIR) filter to each channel output to enable a response-invariant beamforming design (Sydow, 1994). Another advantage of filter-and-sum is that all the FIR design algorithms, such as dereverberation, can be applied to each individual channel.

4.1.2 Adaptive beamforming

Instead of optimizing beam patterns in the fixed beamforming approaches, an adaptive beamformer is usually designed by first formulating a criterion and then adapting the filtering coefficients to optimize it statistically based on the arriving source speech and interfering signals. Some popular criteria used to design an adaptive beamformer include maximum-SNR (Applebaum, 1976), least squares error (Doclo and Moonen, 2003), minimum variance distortionless response (MVDR) (Capon, 1969; Souden et al., 2010, 2013), and linearly constrained minimum variance (LCMV) (Frost, 1972).

MVDR is perhaps the most widely used adaptive beamformer in speech recognition. An MVDR beamformer estimates the coefficients to minimize the power of the output signals under the constraint that the desired source speech signals are not distorted. In the STFT domain, denoting the frame and frequency index by t and f, we can express the observations at a microphone array in vector-matrix form as

$$\boldsymbol{Y}(t,f) = \boldsymbol{h}_f\boldsymbol{X}(t,f)+\boldsymbol{V}(t,f), \tag{18}$$

where $\boldsymbol{X}(t,f)$ and $\boldsymbol{V}(t,f)$ are the source and interfering signals, respectively. $\boldsymbol{h}_f = [h_1(f),h_2(f),...,h_N(f)]^{\text{T}}$ is the steering vector representing the propagation from the source to the microphones. The array output can be written as

$$\hat{\boldsymbol{X}}(t,f)=\boldsymbol{w}_f^H\boldsymbol{Y}(t,f)=\boldsymbol{w}_f^H(\boldsymbol{h}_f\boldsymbol{X}(t,f))+\boldsymbol{w}_f^H\boldsymbol{V}(t,f). \tag{19}$$

According to MVDR, $\boldsymbol{w}_f$ is solved by

$$\begin{aligned}\boldsymbol{w}_f &= \arg\min_{\boldsymbol{w}_f}E|\boldsymbol{w}_f^H\boldsymbol{V}(t,f)|^2\\ &\text{s.t.}\quad \boldsymbol{w}_f^H\boldsymbol{h}_f = 1.\end{aligned} \tag{20}$$

By solving Eq. (20) with Lagrange multipliers and

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

setting the derivatives to zero, we have

$$\boldsymbol{w}_f = \frac{(\boldsymbol{R}_f^{vv})^{-1}\boldsymbol{h}_f}{\boldsymbol{h}_f^H(\boldsymbol{R}_f^{vv})^{-1}\boldsymbol{h}_f}. \tag{21}$$

As can be seen in Eqs. (18)–(21), the key components of MVDR coefficients computation are steering vector $\boldsymbol{h}_f$ and spatial correlation matrix $\boldsymbol{R}_f^{vv}$ of the interfering signals. In a practical MVDR system, both of the coefficients can be calculated based on the estimation of a T-F mask $\boldsymbol{M}(t,f)$:

$$\boldsymbol{M}(t,f) = \begin{cases} 1, \text{ if } |\boldsymbol{V}(t,f)| > |\boldsymbol{X}(t,f)|, \\ 0, \text{ otherwise.} \end{cases} \tag{22}$$

With the mask available, we have

$$\boldsymbol{R}_f^{vv} = \frac{\sum\limits_t \boldsymbol{M}(t,f)\boldsymbol{y}_{t,f}\boldsymbol{y}_{t,f}^H}{\sum\limits_t \boldsymbol{M}(t,f)}. \tag{23}$$

Souden et al. (2013) showed that the steering vector can be estimated using the principal eigenvector of the spatial correlation matrix of source signals $\boldsymbol{R}_f^{ss}$, which can be further obtained by

$$\boldsymbol{R}_f^{xx} = \boldsymbol{R}_f^{yy} - \boldsymbol{R}_f^{vv}. \tag{24}$$

A reliable mask estimation often plays a key role in an MVDR beamformer in real applications. A simple mask estimation can be conducted based on the voice activity detection (VAD) module, while more sophisticated schemes have been explored by Souden et al. (2010). Recently, a few neural network-based approaches have been proposed (Erdogan et al., 2016; Heymanna et al., 2017; Xiao et al., 2017) to improve the mask estimation accuracy. Gannot et al. (2001) showed that MVDR is a special case of the LCMV beamforming approach and is closely related to a generalized sidelobe canceller (GSC) (Gannot et al., 2004).

4.1.3 Direction-of-arrival estimation

Obviously, a good direction-of-arrival (DOA) estimation is crucial to generic beamforming algorithms. When the geometry of a microphone array is determined, DOA estimation is equivalent to time-difference-of-arrival (TDOA) estimation. Most TDOA estimation algorithms are based on the cross-correlation method where delay estimation is obtained as the lag time that maximizes the cross-correlation function between two channels. The generalized cross correlation (GCC) (Knapp and Carter, 1976) unifies various cross correlation-based algorithms into one framework as

$$\hat{\tau}_{\mathrm{GCC}} = \arg\max_m \Psi_{\mathrm{GCC}}[m], \tag{25}$$

where

$$\Psi_{\mathrm{GCC}}[m] = \sum_{k=0}^{K-1} \Phi[k] S_{\boldsymbol{x}_0\boldsymbol{x}_1} \mathrm{e}^{\mathrm{j}2\pi mk/K} \tag{26}$$

is the generalized cross-correlation function (GCCF), $S_{\boldsymbol{x}_0\boldsymbol{x}_1}[k] = E\{\boldsymbol{X}_0[k]\boldsymbol{X}_1[k]^*\}$ is the cross spectrum, '$*$' denotes the complex conjugate operator, $\boldsymbol{X}[k]$ is the DFT of $\boldsymbol{x}[t]$, K is the length of DFT, and $\Phi[k]$ is a weighting function. In a practical system, $S_{\boldsymbol{x}_0\boldsymbol{x}_1}[k]$ is usually estimated by replacing the expected values by the corresponding instantaneous ones. Commonly used weighting functions $\Phi[k]$ include the smoothed coherence transform (SCOT) (Carter et al., 1973), the Roth processor (Roth, 1971), the Echart filter (Knapp and Carter, 1976), the phase transform (PHAT), the maximum-likelihood (ML) processor (Applebaum, 1976), and the Hassab-Boucher transform (Hassab and Boucher, 1981). A GCC with a PHAT TDOA estimation is practically robust to the reverberation where the PHAT transform weighting function is used, $\Phi_{\mathrm{PHAT}}[k] = 1/|S_{\boldsymbol{x}_0\boldsymbol{x}_1}[k]|$. A systematic overview of TDOA algorithms ranging from cross-correlation methods to blind channel identification-based techniques can be found in Chen et al. (2006).

4.2 Multi-channel blind source separation

State-of-the-art multi-channel blind source separation (BSS) approaches can be categorized into two main classes, mask based and independent component analysis (ICA). Mask-based BSS approaches separate multiple sources of speech usually by performing clustering algorithms on the STFT representation of a speech signal and estimating a T-F mask for each source, while an ICA-based approach conducts the separation based on the assumption of independence between the sources.

4.2.1 Mask-based blind source separation

Mask-based BSS was first studied in Yilmaz and Rickard (2004) where the W-disjoint

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

orthogonality assumption (i.e., each T-F bin is dominated by no more than one source) was introduced. Instead of estimating the binary T-F mask simply via an ML estimator of attenuations and delays in the original masked-based BSS work, more advanced probabilistic models have been adopted to perform mask estimation (Sawada et al., 2007; Mandel et al., 2010; Souden et al., 2013). In Sawada et al. (2007), the frequency bin-wise mixtures were classified based on GMM fitting when a permutation alignment needs to be conducted to make sure that the same class index from each frequency bin corresponds to the same source. In model-based EM source separation and localization (MESSL) (Mandel et al., 2010), each source in a broadband mixture is directly modeled by a probabilistic model of inter-aural parameters and the mask is estimated iteratively when fitting this mixture model to the data using an expectation maximization (EM) algorithm; therefore, the permutation alignment step is no longer needed in the later stage.

Most mask-based approaches can be used as a single channel filter; however, they work more effectively when multi-channel inputs become available and more optimally if combined with other multi-channel techniques such as MVDR beamforming. Traditionally, mask-based BSS approaches are unsupervised in the sense that mask estimation is based mainly on the clustering algorithms. Recently, more and more deep learning approaches, such as deep clustering (Hershey et al., 2016; Isik et al., 2016) and permutation invariant training (Kolbæk et al., 2017b; Yu et al., 2017b), have been introduced to more reliably estimate masks for a better separation quality. These techniques formulate the mask estimation problem as a supervised learning problem that requires source labels during the training stage. Actually, some CASA-based approaches can also be regarded as mask-based BSS, and deep learning-based CASA approaches have been explored in much research (Narayanan and Wang, 2013; Wang et al., 2014; Erdogan et al., 2017). See Section 5 for details on mask-based BSS using deep learning.

4.2.2 Independent component analysis

ICA-based BSS approaches are a family of methods for finding statistically independent sources given a mixture of sources by using higher-order statistics (Lee, 1998; Hyvarinen et al., 2001). They usually work more effectively in a case where there are more microphones than in a case where there are sources to be separated. A typical ICA-based BSS approach requires two steps: narrow-band source separation and permutation alignment. Independent vector analysis (IVA) (Ono, 2011), however, is formulated in a single step without requiring the permutation alignment step. The simplest form of ICA algorithms assumes the instantaneous mixing model; i.e., all the signals arrive at the sensors at the same time without any reverberation, whereas in a real environment, the mixing models involve both delays and convolutions. To deal with more realistic convolutive mixtures in both time and frequency domains, several approaches have been proposed (Kim et al., 2006). A good survey of ICA-based convolutive BSS can be found in Pedersen et al. (2007).

5 Deep learning techniques

In recent years, inspired by the success in speech recognition, deep learning techniques have been introduced to solve the cocktail party problem. Most research has been conducted on monaural speech separation tasks.

5.1 Separation with supervised regression

Deep learning models are mostly effective when the problem can be formulated as a supervised learning problem. In the monaural speech separation task, a linearly mixed single-microphone signal $\boldsymbol{y}[t] = \sum_{s=1}^{S} \boldsymbol{x}_s[t]$ is given, where t is the time index, and $\boldsymbol{x}_s[t]$ ($s \in \{1, 2, \cdots, S\}$ are S individual source signals. The goal of the task is to recover the source signals. In most cases, the separation is carried out on the corresponding short-time Fourier transformation (STFT) $\boldsymbol{Y}(t, f)$ to recover each source $\boldsymbol{X}_s(t, f)$ in the T-F domain for every time point t and frequency bin f. Because there are an infinite number of possible source signal combinations to obtain the same mixed signal, we need to learn regularities in the speech signals from a training set to rule out impossible combinations.

This task can be formulated as a multi-class regression problem, in which the regression module can be a deep learning model. More specifically, given spectral feature $\boldsymbol{Y}(t, f)$ of the mixed speech as the input, the deep model aims to predict the individual

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

spectral feature stream $\boldsymbol{X}_s(t,f)$. The key here is to provide the right supervision information to the network. One approach is to record both the individual source signals and the mixed signal synchronously. However, this can be very expensive and in most cases feasible only for collecting the test data. As an alternative, the training set is usually constructed by recording the source signals and mixing them artificially. This approach, although not optimal, turns out to be very effective because it allows us to generate a huge amount of training data at almost no cost. When the task is to separate speech from other sounds such as noise and music, one of the regression targets can simply be the clean speech and the other noise or music. When the task is to separate multiple streams of speech signals, special techniques need to be used as we will discuss later in Sections 5.4–5.6.

It is possible to directly estimate magnitude spectra $\boldsymbol{X}_s(t,f)$ of each source using a deep learning model (Tu et al., 2014a,b) as shown in Fig. 3. However, it is well known (e.g., in Wang et al. (2014) and Erdogan et al. (2017)) that better results can be achieved if, instead of estimating $|\boldsymbol{X}_s(t,f)|$ directly, we first estimate a set of masks $\boldsymbol{M}_s(t,f)$ using a deep learning model $h(v(\boldsymbol{Y});\Phi)=\hat{\boldsymbol{M}}_s$ and reconstruct magnitude spectra $|\boldsymbol{X}_s|$ as $|\hat{\boldsymbol{X}}_s|=\hat{\boldsymbol{M}}_s\circ|\boldsymbol{Y}|$, where $\circ$ is the element-wise product of two operands, as shown in Fig. 4. This is because masks are well constrained and are invariant to input variabilities caused by, e.g., energy differences.

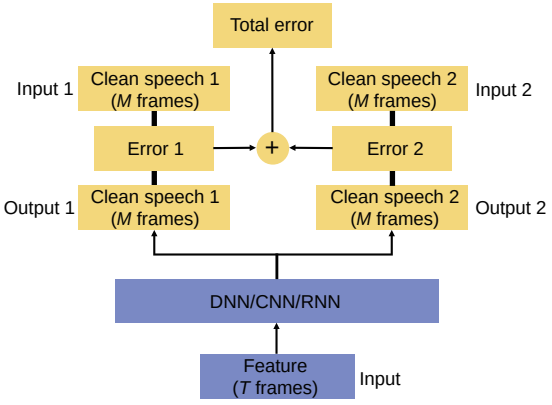

Fig. 3 Direct estimation of the magnitude spectra of each source

DNN: deep neural network; CNN: convolutional neural network; RNN: recurrent neural network

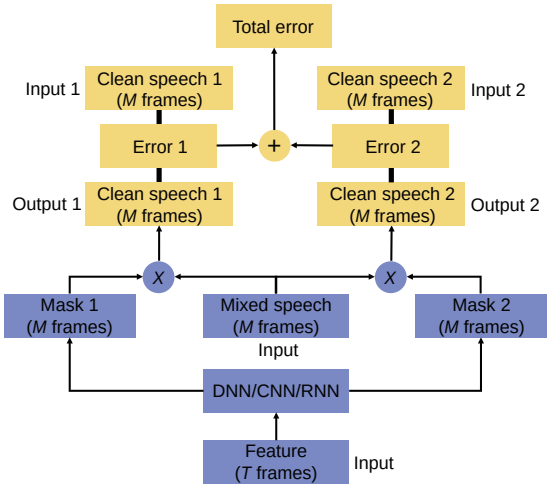

Fig. 4 Estimation of masks and reconstruction of the source spectra from the masks

DNN: deep neural network; CNN: convolutional neural network; RNN: recurrent neural network

5.2 Masks and training criteria in supervised regression

Because masks are estimated and used to reconstruct the magnitude spectrogram of source signals, the choice of masks is important. Several masks have been developed (Narayanan and Wang, 2013; Wang et al., 2014; Erdogan et al., 2015, 2017; Kolbæk et al., 2017b), including the ideal ratio mask (IRM), ideal amplitude mask (IAM), and phase sensitive mask (PSM).

The ideal ratio mask (IRM) for each source is defined as

$$\boldsymbol{M}_s^{\mathrm{IRM}}(t,f)=\frac{|\boldsymbol{X}_s(t,f)|}{\sum_{s=1}^{S}|\boldsymbol{X}_s(t,f)|}. \tag{27}$$

It was shown that IRM maximizes SDR (Vincent et al., 2006) when all sources have the same phase, which is not realistic. IRMs have the constraints $0\le\boldsymbol{M}_s^{\mathrm{IRM}}(t,f)\le 1$ and $\sum_{s=1}^{S}\boldsymbol{M}_s^{\mathrm{IRM}}(t,f)=1$ for all T-F bins (t,f)'s, which can be satisfied with the softmax activation function. However, because $\sum_{s=1}^{S}|\boldsymbol{X}_s(t,f)|$ is unknown in the mixed speech, IRM cannot be used practically to reconstruct the source streams.

Practically, we can use IAM defined as

$$\boldsymbol{M}_s^{\mathrm{IAM}}(t,f)=\frac{|\boldsymbol{X}_s(t,f)|}{|\boldsymbol{Y}(t,f)|} \tag{28}$$

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

to reconstruct $\boldsymbol{X}_s$, because the magnitude spectra of the mixed speech $\boldsymbol{Y}$ is known during testing. IAMs have the constraint $0 \leq \boldsymbol{M}_s^{\mathrm{IAM}}(t,f) \leq \infty$, although it is found empirically that the majority of the T-F units are in the range of $0 \leq \boldsymbol{M}_s^{\mathrm{IAM}}(t,f) \leq 1$. Accordingly, softmax, sigmoid, and ReLU are possible activation functions for estimating IAMs in the implementation.

IAM is suboptimal because it does not consider the phase differences between source signals and the mixture. PSM proposed by Erdogan et al. (2015, 2017) and Kolbæk et al. (2017b) as

$$\boldsymbol{M}_s^{\mathrm{PSM}}(t,f) = \frac{|\boldsymbol{X}_s(t,f)|\cos(\theta_y(t,f)-\theta_s(t,f))}{|\boldsymbol{Y}_s(t,f)|}, \tag{29}$$

on the other hand, takes phase difference into consideration, where θ_y and θ_s are the phases of mixed speech $\boldsymbol{y}$ and source $\boldsymbol{x}_s$, respectively. Due to the phase-correcting term, PSM sums to one, i.e., $\sum_{s=1}^{S} \boldsymbol{M}_s^{\mathrm{PSM}}(t,f) = 1$.

Once the mask is chosen, a deep model can be optimized to minimize the mean square error (MSE) between estimated mask $\hat{\boldsymbol{M}}_s$ and target mask

$$J_m = \frac{1}{B}\sum_{s=1}^{S} \|\hat{\boldsymbol{M}}_s - \boldsymbol{M}_s\|_{\mathrm{F}}^2, \tag{30}$$

where the denominator $B = T \cdot F \cdot S$ is the total number of T-F bins over all sources and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. However, directly optimizing for the mask error comes with two drawbacks. First, in the silent segments, $|\boldsymbol{X}_s(t,f)| = 0$ and $|\boldsymbol{Y}(t,f)| = 0$; thus, target masks $\boldsymbol{M}_s(t,f)$ are not well defined. Second, a smaller error in mask estimation does not always translate to a smaller reconstruction error between the reconstructed source signal and the true source signal.

To overcome the above limitations, Wang et al. (2014) proposed to directly minimize MSE between the estimated magnitude and the true magnitude:

$$J_x = \frac{1}{B}\sum_{s=1}^{S} \||\hat{\boldsymbol{X}}_s| - |\boldsymbol{X}_s|\|_{\mathrm{F}}^2 \tag{31}$$

$$= \frac{1}{B}\sum_{s=1}^{S} \|\hat{\boldsymbol{M}}_s \circ |\boldsymbol{Y}| - |\boldsymbol{X}_s|\|_{\mathrm{F}}^2. \tag{32}$$

When the phase-sensitive mask is used, J_x can be reformulated as J_p (Erdogan et al., 2017; Kolbæk et al., 2017b):

$$J_p = \frac{1}{B}\sum_{s=1}^{S} \|\hat{\boldsymbol{M}}_s \circ |\boldsymbol{Y}| - |\tilde{\boldsymbol{X}}_s|\|_{\mathrm{F}}^2 \tag{33}$$

$$= \frac{1}{B}\sum_{s=1}^{S} \|\hat{\boldsymbol{M}}_s \circ |\boldsymbol{Y}| - |\boldsymbol{X}_s| \circ \cos(\theta_y - \theta_s)\|_{\mathrm{F}}^2, \tag{34}$$

where $|\tilde{\boldsymbol{X}}_s = |\boldsymbol{X}_s| \circ \cos(\theta_y - \theta_s)$ is the phase-discounted magnitude target. Eq. (34) essentially indicates that to use PSM we need only to provide the phase-discounted magnitude as the training target. Experiments in Erdogan et al. (2017) and Kolbæk et al. (2017b) show that PSM consistently outperforms IAM.

5.3 Label permutation problem

In the multi-class regression framework, we need to provide the correct reference (or target) magnitude $|\boldsymbol{X}_1|$ and $|\boldsymbol{X}_2|$ to the corresponding output layer segments for supervision during training. Assigning the supervision with a fixed order (Tu et al., 2014a; Wang et al., 2014; Weninger et al., 2015) (Figs. 3 and 4), works well for separating speech from other sounds, but not for separating mixed speech in the cocktail party environment due to the label permutation problem. Assume that there are two speakers in the mixed speech. Because speech sources are symmetric given the mixture ($\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have the same characteristics), we do not know whether it is $(\boldsymbol{X}_1 + \boldsymbol{X}_2)$ or $(\boldsymbol{X}_2 + \boldsymbol{X}_1)$. There is thus no predetermined way to assign the correct target to the corresponding output layer segment. This problem becomes serious when the number of speakers in the mixed speech increases. It prevents the supervised regression framework from being used to solve the speaker-independent cocktail party problem.

Recently several strategies have been proposed to address the label permutation problem. We will discuss the most promising ones, including DPCL (Hershey et al., 2016; Isik et al., 2016), DANet (Chen et al., 2017b), and PIT (Kolbæk et al., 2017b; Yu et al., 2017b), one by one.

5.4 Deep clustering

Hershey et al. (2016) proposed a speech separation framework called 'DPCL' to address the label permutation problem. Different from the

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

supervised regression framework, they cast the separation problem as a segmentation problem. Specifically, they assumed that each T-F bin (t, f) of the mixed speech belongs to only one speaker. If we assign the same unique color to the bins belonging to the same speaker, the spectrogram is segmented into clusters, one for each speaker. The key observation in this framework is that during training we need only to know which bins belong to the same speaker (or cluster), and which is unambiguous, thus avoiding the label permutation problem.

Because clustering is defined based on some distance between bins, Hershey et al. (2016) proposed to define the distance in the embedding space of the bins that the system can learn from the training data. If two bins belong to the same speaker, their distance in the embedding space is small, and if two bins belong to different speakers, their distance in the embedding space is large.

Precisely, given a raw input signal $\boldsymbol{y}$, its feature vector is defined as $\boldsymbol{Y}_i = g_i(\boldsymbol{y})$ $(i \in \{1, 2, \cdots, N\})$, where i is the T-F index (t, f) in the case of audio signals. A deep neural networkwork is used to transform input signal $\boldsymbol{x}$ into D-dimensional embeddings $\boldsymbol{V} = f_\theta(\boldsymbol{Y}) \in \mathbb{R}^{N \cdot D}$, where each row vector $\boldsymbol{v}_i$ has unit norm. Performing clustering in the embedding space will likely lead to a partition of $\{1, 2, \cdots, N\}$, which is close to the target. Embeddings $\boldsymbol{V}$ is considered to implicitly represent an $N \times N$ estimated affinity matrix $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}$. The target partition is represented by indicator $\boldsymbol{E} = \{\boldsymbol{e}_{i,s}\}$, mapping each element i to each of S clusters; thus, $\boldsymbol{e}_{i,s} = 1$ if element i is in cluster c. In this case, $\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}$ is considered as a binary affinity matrix that represents the cluster assignments in a permutation-independent way: $(\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}})_{i,j} = 1$ if elements i and j belong to the same cluster, and $(\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}})_{i,j} = 0$ otherwise, and $(\boldsymbol{E}\boldsymbol{P})(\boldsymbol{E}\boldsymbol{P})^{\mathrm{T}} = \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}$ for any permutation matrix $\boldsymbol{P}$.

Thus, we can learn affinity matrix $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}$, as a function of inputs $\boldsymbol{X}$, to match affinities $\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}$, by minimizing the training cost function, with respect to $\boldsymbol{V} = f_\theta(\boldsymbol{Y})$:

$$C_{\boldsymbol{E}}(\boldsymbol{V}) = \left\|\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 = \sum_{i,j} \left(\langle \boldsymbol{v}_i, \boldsymbol{v}_j\rangle - \langle \boldsymbol{e}_i, \boldsymbol{e}_j\rangle\right)^2 \quad (35)$$

$$= \sum_{i,j:\boldsymbol{e}_i=\boldsymbol{e}_j} \left(|\boldsymbol{v}_i - \boldsymbol{v}_j|^2 - 1\right) + \sum_{i,j} \langle \boldsymbol{v}_i, \boldsymbol{v}_j\rangle^2 \quad (36)$$

summed over training examples, where $\|\cdot\|_{\mathrm{F}}^2$ is the squared Frobenius norm.

For the inference, embeddings $\boldsymbol{V}$ is first computed on input signal $\boldsymbol{Y}$, and row $\boldsymbol{v}_i$ is clustered using K-means. The resulting cluster assignments are used as binary masks to separate the sources. The sources are estimated by applying the final masks based on cluster assignments $\bar{\boldsymbol{E}}$ to the mixture signal. One interesting property is that if we know the number of speakers in the mixed speech signal, DPCL can essentially separate mixed speech of different numbers of speakers in the same model.

This basic DPCL framework can recover only binary masks for each source, while each bin is a mixture of speech from multiple speakers, as we know. Isik et al. (2016) addressed this limitation with a second-stage enhancement network that directly improves the signal reconstruction. In addition, the enhancement stage can be trained with the deep clustering embeddings in an end-to-end mode using the signal reconstruction objective instead of the original mask-based deep clustering objective.

DPCL is the first interesting technique that solves the label permutation problem. However, due to the clustering step, it comes with several drawbacks. First, the whole training and inference pipeline is complex. This makes it difficult to incorporate other techniques into the framework. Second, because clustering usually occurs only after the embeddings of all bins are available to avoid performance degradation, it is not suitable for a real-time streaming process. To solve this problem, an online GMM/K-means technique can be used to reduce latency at the cost of some performance degradation. Note that even with online GMM/K-means, some latency may still be unavoidable due to the need to estimate the number of speakers. Third, when the same model is used to separate two- and three-speaker mixed speech, an estimation of the number of speakers needs to be conducted first, usually introducing additional errors.

5.5 Deep attractor network

One of the drawbacks of the original DPCL is its inefficiency in performing end-to-end mapping, because it optimizes for the affinity between the sources in the embedded space rather than the separated signals themselves. Chen et al. (2017b) proposed an improved framework called 'DANet' as shown in Fig. 5.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

It can be seen that during training, an ideal mask is applied to form the attractors; during testing, K-means is used to form the attractor.

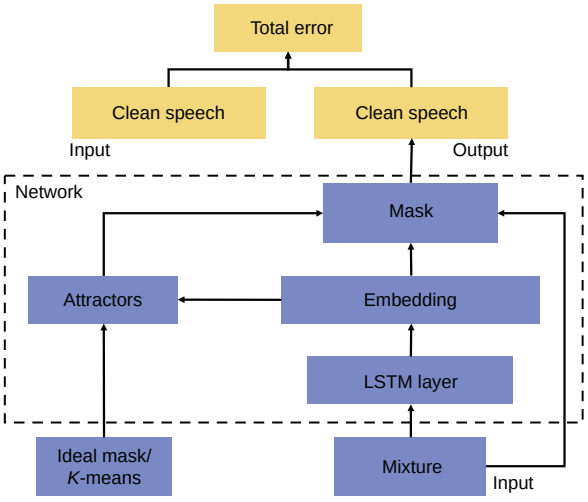

Fig. 5 System architecture of the deep attractor network (LSTM: long short-term memory)

The term 'attractor' refers to the well-studied perceptual effects in human speech perception. It is believed that the brain circuits create perceptual attractors that warp the stimulus space to draw the sound that is closest to it (Kuhl, 1991). DANet follows a similar principle by forming a reference attractor, which draws all the T-F bins toward itself, for each source in the embedding space. Using the similarity between the embedded points and each attractor, a mask is estimated for each source in the mixture. Similar to DPCL, DANet does not have the label permutation problem because it can be considered as a soft-clustering technique just like GMM. Because the number of masks is determined by the number of attractor points, the proposed framework can potentially be extended to an arbitrary number of sources. Compared with DPCL, the mask learning in DANet enables more efficient end-to-end training.

Given the mixture signal with S sources, a K-dimensional embedding $\boldsymbol{V}\in\mathbb{R}^{F\cdot T\cdot K}$ of the mixed acoustic signal $\boldsymbol{Y}=[F\cdot T]$, where F is the frequency and T is the time, is learned by the neural network. During training, attractors $\boldsymbol{A}\in\mathbb{R}^{S\cdot K}$ are learned as

$$\boldsymbol{A}_{s,k}=\frac{\sum\limits_{f,t}\boldsymbol{V}_{k,ft}\cdot\boldsymbol{E}_{s,ft}}{\sum\limits_{f,t}\boldsymbol{E}_{s,ft}} \tag{37}$$

in the embedding space, where $\boldsymbol{E}\in\mathbb{R}^{F\cdot T\cdot S}$ is the dominant source membership function for each T-F bin. A mask $\boldsymbol{M}$ is then estimated in the embedding space as

$$\boldsymbol{M}_{f,t,s}=\mathrm{Softmax}\left(\sum_{K}\boldsymbol{A}_{s,k}\cdot\boldsymbol{V}_{ft,k}\right). \tag{38}$$

Finally, the neural network is trained to minimize

$$\mathcal{L}=\sum_{f,t,s}\|\boldsymbol{X}_{f,t,s}-\boldsymbol{Y}_{f,t}\cdot\boldsymbol{M}_{f,t,s}\|_2^2. \tag{39}$$

where $\boldsymbol{X}$ is the clean spectrogram of S sources.

During the test, true assignment $\boldsymbol{E}$ is unknown; thus, the attractor points need to be estimated differently, either using the K-means algorithm or reusing the attractor points from the training stage. The latter method is based on an observation that the location of the attractors in the embedding space is relatively stable between training and testing. Details can be found in Chen et al. (2017b).

DANet is a direct extension of the hard clustering in DPCL and thus has drawbacks similar to DPCL. For example, it needs to estimate the number of speakers in the mixed signal before it can compute the masks. In addition, using either the attractors learned during training or those estimated with K-means is suboptimal.

5.6 Permutation invariant training

A different approach was proposed by Yu et al. (2017b) and Kolbæk et al. (2017b) to address the label permutation problem. As shown in Fig. 6, the key ingredient of this approach is PIT, illustrated in the dashed rectangle in Fig. 6. PIT casts speech separation as a multi-class segregation problem where the supervision is provided as a set instead of an ordered list.

In this architecture, feature vectors of the mixed signal $\boldsymbol{Y}(t,f)=\sum\limits_{s=1}^{S}\boldsymbol{X}_s(t,f)$ are input to the deep learning model to estimate mask $\hat{\boldsymbol{M}}_s(t,f)$ for each talker. Each mask $\hat{\boldsymbol{M}}_s$ is then used to construct a single-source speech

$$\hat{\boldsymbol{X}}_s=\hat{\boldsymbol{M}}_s\circ|\boldsymbol{Y}| \tag{40}$$

at the corresponding output layer segment.

During training, the correct reference magnitudes $\boldsymbol{X}_s$ ($s\in\{1,2,\cdots,S\}$) (or their phase-adjusted counterparts) of each source stream are given as a

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

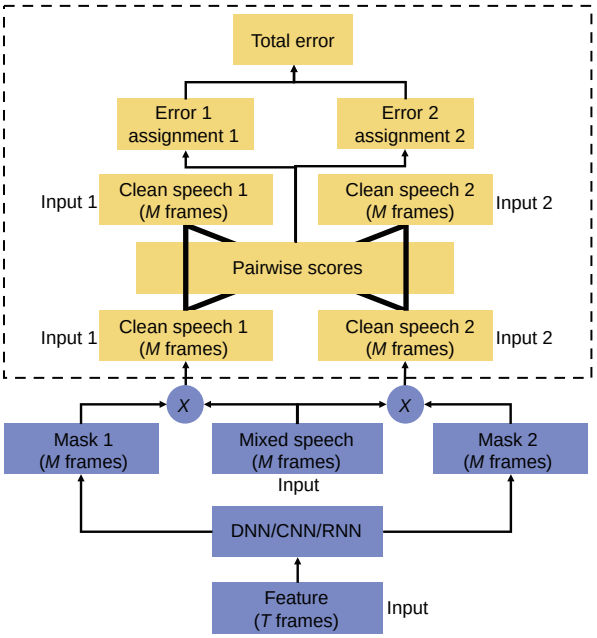

Fig. 6 The two-talker speech separation model with permutation invariant training

DNN: deep neural network; CNN: convolutional neural network; RNN: recurrent neural network

set. The pairwise MSE is first computed between each pair of reference $|\boldsymbol{X}_s|$ and estimated source $|\hat{\boldsymbol{X}}_s|$ for any possible assignment. Then the assignment with the least MSE is chosen and the model is optimized to further reduce this least MSE:

$$J=\frac{1}{F\cdot T\cdot S}\min_{s'\in\mathrm{permu}(S)}\sum_{s=1}^{S}\||\hat{\boldsymbol{X}}_s|-|\boldsymbol{X}_s|\|_{\mathrm{F}}^2, \quad (41)$$

where permu(S) is a permutation of $1,2,\cdots,S$. Kolbæk et al. (2017b) showed that when recurrent neural networks with utterance-level training loss are used, PIT can optimize effectively for speech separation and speaker tracing in one shot. During evaluation, each separated speech stream can be easily constructed by assembling estimated frames from the same output layer segment sequentially. In contrast to first impression, here the extra computation introduced by permutation evaluation is quite limited. When there are S speakers in the mixed speech, the computation of distance between each pair of true and estimated spectra happens only S^2 times. The permutation score evaluation indeed happens $S!$ times. However, it takes only $O(S!)$ summations, which can be almost ignored completely (compared with the deep learning model itself) during training. During separation, even this tiny extra computation does not exist because permutation evaluation is needed only during training.

Fig. 7 summarizes the speech separation quality of DPCL, DANet, and PIT. All the deep learning techniques (including DPCL, DANet, and PIT) that avoid the label permutation problem significantly outperform the conventional techniques such as CASA and NMF. All the three deep learning techniques can separate two- and three-talker mixed speech with comparable quality using a single model. The result of one DANet in Chen et al. (2017b) is better than that of PIT. However, it was obtained with a complicated training schedule that includes curriculum training. However, PIT is much simpler to implement, easier to integrate with other techniques, and more efficient during testing. With PIT, one does not need to estimate the number of speakers in the mixed speech before separation. This means that the estimation error on the number of speakers will not affect the separation quality, even if users do want to estimate the number of speakers in some applications, e.g., by checking the energy of each output segment, whereas for approaches such as DPCL and DANet, the estimation error on the number of speakers will affect the separation result.

PIT also has limitations. Similar to other deep learning-based techniques, PIT performs much better with an opposite-gender mixed speech than it does with a same-gender mixed speech. Another limitation is that the maximum number of mixing streams that the model can handle is determined by the network architecture; e.g., the PIT with two output segments will not work for three-talker speech separation. This is inferior to DPCL and DANet in which the network architecture is independent of the number of speakers, although their model size may need to be increased to support more speakers. Fortunately, this limitation is not a big concern in practice, because in the majority of the environments the system needs to pay attention to at most three overlapping streams and treat the rest (usually low-energy streams if they exist) as speech-like noises. Kolbæk et al. (2017a) reported that this can be very nicely handled by PIT. This essentially means that the model needs at most four output segments to cover 99% of scenarios. For example, even if one wants to separate six overlapping streams, the separation quality is poor no matter which existing technique is used, and thus the limitation from the output number of PIT has a limited practical

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

importance. In addition, even though DPCL and DANet can still output separation results with more mixing streams larger than those have been seen during training, the separation quality is poor. For example, according to Table 5 in Isik et al. (2016), DPCL trained on two-talker mixed speech achieves only an SDR improvement of 2.1 dB on three-talker mixed speech, whereas the system trained on three-talker mixed speech can achieve an SDR improvement of 7.1 dB. This result indicates that to better separate S-talker mixed speech, one needs to train the model with up to S mixing streams no matter which technique is used.

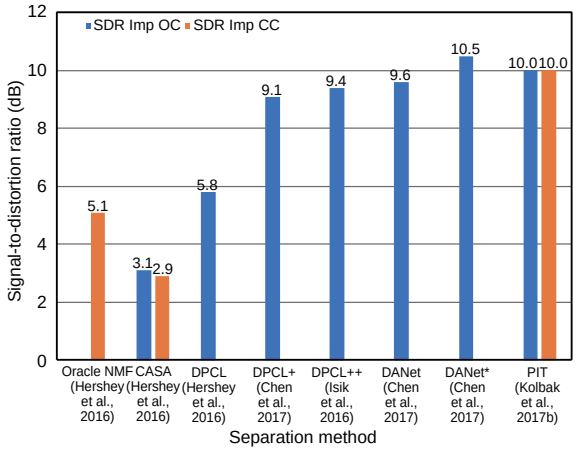

Fig. 7 Signal-to-distortion ratio (SDR) improvements for different separation methods on the WSJ0-2MIX dataset without additional tracing

* indicates curriculum training. CC: closed condition; OC: open condition

6 Multi-talker speech recognition

The speech separation techniques described in Section 5 can be used not only to enhance speech for improved human–human communication, but also to improve the ASR performance in the cocktail party environment. A simple solution used in Isik et al. (2016) is to use single- or multi-channel speech separation techniques to estimate each speech source stream in the mixed signal, and then use a single-talker ASR system to recognize each stream. To improve the recognition accuracy, single-talker ASR system can be adapted using the reconstructed speech streams.

Alternatively, the mixed speech signal can be recognized directly without an explicit speech separation stage. Similar to the speech separation task, there exists the label permutation problem when doing so.

Weng et al. (2015) proposed to solve the label permutation problem by assigning the senone labels of the talker with higher instantaneous energy to output one and the other to output two. Although this addresses the label ambiguity problem, it causes frequent speaker switch across frames. To deal with the speaker switch problem, a two-speaker joint-decoder with a speaker switching penalty was used by Weng et al. (2015) to trace speakers. This approach has two limitations. First, energy may not be the best information for assigning labels under all conditions. In fact, under most conditions, pitch is a better indication. Second, the frame switching problem introduces additional burden to the decoder.

In Yu et al. (2017a) and Chen et al. (2017c), PIT was extended for direct recognition of overlapped speech without first separating the signal into speech streams. Given the mixed signal of S speech sources, a deep neural networkwork with S output layer segments is used as the classification model, as shown in Fig. 8.

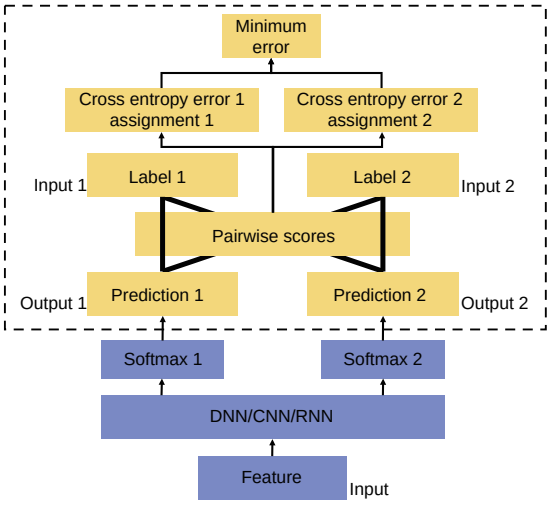

Fig. 8 The two-talker speech recognition model with permutation invariant training

DNN: deep neural network; CNN: convolutional neural network; RNN: recurrent neural network

Each output layer segment represents an estimate of the senone posterior probability for a speech stream. Similar to that in speech separation, the model is optimized by minimizing the objective function

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$J=\frac{1}{S}\min_{s'\in\mathrm{permu}(S)}\sum_{s}\sum_{t}\mathrm{CE}(\boldsymbol{\ell}_t^{s'},\boldsymbol{O}_t^{s}),\ s=1,2,\cdots,S, \tag{42}$$

where permu(S) is a permutation of $1,2,\cdots,S$, $\boldsymbol{O}_s^t$ is the output for stream s at timestep t, and $\boldsymbol{\ell}_t^{s'}$ are the labels for source s'. In other words, their approach first computes the average CE over the whole sequence for each possible assignment of labels. The one with the minimum CE is then picked for optimization. Experiments conducted by Chen et al. (2017c) and Qian et al. (2017) showed that PIT-based direct mixed speech recognition can cut WER by 45% over the baseline. In addition, similar to that in speech separation, the PIT-based ASR system can effectively recognize single- and multi-talker speech with a single model without knowing or estimating the number of speakers in the mixed speech.

In addition to the superiority on the accuracy and the simplicity of the architecture, another advantage of the PIT-based multi-talker speech recognition technique is its ability to combine other techniques to further improve the recognition accuracy. For example, it has been improved with adaptation (Chang et al., 2018), sequence discriminative training (Chen et al., 2017c), and knowledge distillation (Chen et al., 2017c; Tan et al., 2018).

7 Multi-talker speaker identification

Multi-talker speaker identification (SID) aims to recognize the identities of multiple talkers when they speak simultaneously. It is important in many applications, such as meeting transcription, and can help improve speech separation by paying attention to specific speakers. Similar to the separation and recognition tasks, there has been research on single- and multi-channel conditions, and most of the work focuses on single-channel setup.

It is obvious that although the state-of-the-art SID techniques, such as Gaussian mixture model-universal background model (GMM-UBM) (Reynolds et al., 2000), *i*-vector (Dehak et al., 2011), and *d*/j-vector (Variani et al., 2014; Chen et al., 2015), have achieved an impressive accuracy in the single-talker scenario (Dehak et al., 2011; Liu et al., 2015; Zhang and Koishida, 2017; Huang et al., 2018), they perform poorly on highly overlapped speech. If the individual speech stream in the multi-talker mixed speech can be separated with a good quality, single-talker SID algorithms may be directly applied. Unfortunately, multi-talker speech separation itself is a very challenging problem.

For this reason, most of the current work has performed this task without an explicit separation stage, and used a pattern recognition technique to recognize the identities directly under the closed-set supervised setup, in which all speakers in the testing phase are shown in the training set.

7.1 Conventional techniques

Different from multi-talker speech recognition, multi-talker (also called 'co-channel') SID can be conducted on a subset of homogeneous speech segments, called 'usable speech' (Lovekin et al., 2001). To take advantage of this property, some researchers focus on extracting usable speech by casting it as a sequential grouping problem. For example, Shao and Wang (2003, 2006) used a multi-pitch tracker to find frames with only one pitch point and treated them as an usable speech. They then jointly evaluated all the grouping hypotheses and speaker candidates, and selected the optimal one. Mowlaee et al. (2010) proposed an iterative multi-talker SID and separation procedure, which was later improved by fusing adapted GMM and Kullback–Leibler divergence (KLD) scores (Mowlaee et al., 2012).

The key to these conventional techniques is the algorithm that identifies the most probable speaker pair $(\hat{\lambda}_a,\hat{\lambda}_b)$ from all possible pairs. This is usually conducted by modeling the conditional joint probability of each pair of speakers in the training set given the observation O. If we assume that the prior probability of each speaker pair $P(\lambda_a,\lambda_b)$ is equal, we can optimize for

$$\begin{aligned}\hat{\lambda}_a,\hat{\lambda}_b&=\arg\max_{\lambda_a,\lambda_b} P(\lambda_a,\lambda_b|O)\\&=\arg\max_{\lambda_a,\lambda_b}\frac{p(O|\lambda_a,\lambda_b)P(\lambda_a,\lambda_b)}{p(O)}\\&=\arg\max_{\lambda_a,\lambda_b} p(O|\lambda_a,\lambda_b).\end{aligned} \tag{43}$$

In most work (Hershey et al., 2010; Li et al., 2010), the joint probability of two speakers was modeled using GMMs. Unfortunately, the search space increases exponentially with the number of speakers in this approach. Hershey et al. (2010) proposed to first model each speaker λ given observation O individually as

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$p(\lambda|O)=\frac{p(O|\lambda)P(\lambda)}{\sum_s p(O|\lambda_s)P(\lambda_s)}, \tag{44}$$

where s is the speaker index, and $P(\lambda)$ and $P(\lambda_s)$ are prior probabilities. The top speaker is then paired with the rest for expectation-maximization (EM) based gain estimation. The output is the speaker pair whose gain adapted model maximizes the likelihood of the test utterance.

Li et al. (2010) proposed a simpler yet more effective method, in which the top speaker model is directly combined with each of the other candidates. These speech pairs are then evaluated directly, one by one, for SID without the EM step. These two approaches are the best conventional techniques for this task, and the multi-talker SID results on the SSC task can be found in the first two rows of Table 2.

7.2 Deep learning techniques

Co-channel SID can also be formulated as a multi-class classification problem, in which the goal is to predict the target speakers provided that the utterance and deep learning models can be used. The flow chart of the typical deep learning based co-channel SID system is shown in Fig. 9. In this architecture, the frame-level multi-talker mixed feature is used as the input, and the soft speaker identities on that frame, computed using the frame-level energy ratio, are used as the training labels. Similar to that in speech separation, artificially generated multi-talker speech is usually used for training. Details can be found in Zhao et al. (2015a).

Zhao et al. (2015b) chose a simple DNN as the deep model and optimized it with the KLD between the model prediction and ground truth soft labels. In the testing phase, utterance-level prediction was obtained by averaging frame-level scores. Given a test utterance $\boldsymbol{O}$ consisting of T frames $\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_T$,

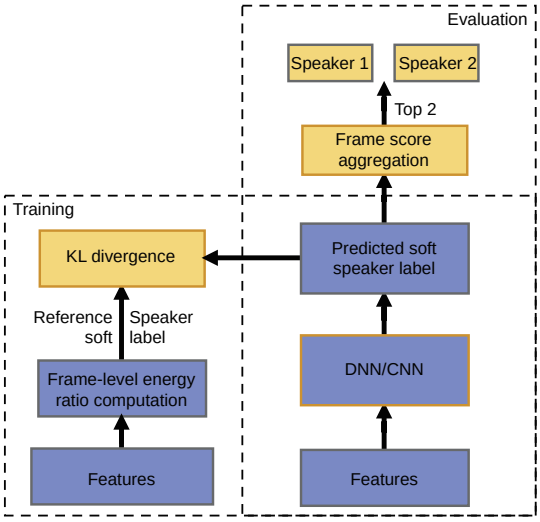

Fig. 9 Illustration of two-talker speaker identification using deep models

DNN: deep neural network; CNN: convolutional neural network

the utterance-level probability for speaker s can be computed as

$$\mathcal{J}(s)=\frac{1}{T}\sum_{t=1}^{\mathrm{T}}P(s|\boldsymbol{o}_t), \tag{45}$$

where $P(s|\boldsymbol{o}_t)$ represents the probability that frame $\boldsymbol{o}_t$ comes from speaker s. The predicted speaker identities can be obtained by selecting the top-S (e.g., $S=2$ for two-talker and $S=3$ for three-talker) speakers with the largest probabilities. The results of this approach in the SSC task are shown in the third row in Table 2. We can see that the deep model based system is consistently better than the conventional ones under all SNR conditions.

Recently, Wang et al. (2018) further improved the system with three refinements, compared with the DNN-based system proposed by Zhao et al. (2015b). First, the basic DNN was replaced by a dilated CNN (D-CNN) (Yu and Koltun, 2015) with a

Table 2 Speaker identification accuracy comparison of different methods on the speech separation and recognition challenge task

Method	Speaker identification accuracy (%)						
	−9 dB	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB
GMM with EM gain estimation (Hershey et al., 2010)	96.5	98.1	98.2	99.0	99.1	98.4	98.2
GMM w/o EM gain estimation (Li et al., 2010)	97.3	98.8	99.5	99.7	99.7	98.8	99.0
KLD-DNN (Zhao et al., 2015b)	98.3	99.5	100.0	99.8	100.0	99.0	99.4
FKLD-DCNN (Wang et al., 2018)	–	–	–	100.0	–	–	–

GMM: Gaussian mixture model; EM: expectation–maximization; KLD-DNN: Kullback–Leibler divergence deep neural network; FKLD-DCNN: focal KL-divergence dilated convolutional neural network

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

superior ability in learning structured features. Second, a focal KL-divergence (FKLD) loss function was proposed for model optimization, which can reduce the relative loss for well-classified samples (simple samples) and pay more attention to the misclassified ones (hard samples). Finally, rather than simply averaging frame-wise scores to obtain an utterance-level prediction, a post-processing task assigning different frames with individual weights was adopted during the evaluation, assigning more confidence to non-overlapped speech frames and less to the overlapped ones.

This newly proposed method was evaluated on an artificially generated multi-talker RSR co-channel SID corpus, which was based on the background part of the RSR2015 dataset (Larcher et al., 2014). Details on the data generation procedure can be found in Wang et al. (2018). A comparison of results with different deep learning techniques is shown in Table 3, for both two- and three-talker speaker identifications. It is observed that each new technique can further improve the performance significantly based on the previous framework, and the final system using the dilated CNN structure with both focal KLD and post-processing can increase the accuracy from 87.16% and 47.79% to 92.47% and 55.83% for two- and three-talker identification, respectively (Wang et al., 2018).

Table 3 Speaker identification accuracy comparison of different deep learning techniques on the multi-talker RSR dataset

Model	Focal	PF	Speaker identification accuracy (%)	
			Two stalkers	Three stalkers
DNN	×	×	87.16	47.79
	✓	×	88.59	51.91
	✓	✓	**89.24**	**52.51**
D-CNN	×	×	88.65	50.86
	✓	×	91.31	55.74
	✓	✓	**92.47**	**55.83**

Bold numbers indicate the best number with each model. PF: post filtering proposed by Wang et al. (2018); DNN: deep neural network; D-CNN: dilated convolutional neural network

8 Discussions and conclusions

In this paper, we have described past efforts in attacking the cocktail party problem. We can observe that the majority of the efforts focus on the speech separation task, and some of the work targets speaker tracing and speech recognition.

Various techniques have been proposed to address the speech separation problem. These techniques were developed to target two different scenarios. The scenario that has received the most attention is monaural speech separation, in which it is assumed that only a single channel of the mixed speech is available. Speech separation under this condition is obviously underdetermined. Additional constraints and regularities are thus needed. These regularities may be obtained through manually designed rules after observing and analyzing mixed speech signals as in CASA. However, research along this direction depends heavily on expert knowledge and cannot exploit the complicated relationship between the T-F bins in the mixed speech. Its success is thus quite limited. More recent work is all data-driven. Non-negative matrix factorization and generative models such as factorial HMMs, are two representative models along this direction. Both techniques outperform CASA even with a moderately sized training set. Unfortunately, both techniques are essentially shallow models with limited modeling power and an indirect optimization criterion. They require building models for each mixing source and are thus not effective under the speaker-independent setup.

Great advancement was observed in monaural speech separation when the problem was converted into a supervised regression problem in which the optimization objective is closely related to the separation task, especially when the regression model is deep learning based. To obtain the reference target without incurring too much data collection effort, synthesized data are used typically for model training. This framework and methodology works well for separating speech from noise and music, or the speech of a specific known speaker from that of other speakers. Unfortunately, it encounters the label ambiguity (or permutation) problem when applied to separate multiple speech streams from the mixed speech signal. As a result, the model cannot be effectively optimized and performs poorly in the cocktail party problem.

Techniques that can solve or avoid the label permutation problem have since been proposed. Deep clustering, the deep attractor network, and permutation invariant training are representative techniques.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

All these techniques significantly outperform the earlier models such as NMF and factorial HMM. More interestingly, all these models support separation of various numbers of mixing speech streams with a single model. Although they achieved a similar signal-to-distortion ratio improvement in various speech separation tasks, their different mechanism leads to different complexities. Among these three techniques, PIT is the simplest to implement and the easiest to integrate with other techniques and thus the most promising from our point of view.

Another scenario that has attracted great attention in recent years is multi-channel speech separation. The two most popular techniques are beamforming and multi-channel blind source separation. Both techniques exploit correlated information existing in different channels to conduct spatial filtering. Beamforming, which is a more widely used technique, guides the attention of the system in one specific direction, which is often estimated using direction of arrival techniques. Thus, signals from that direction are enhanced while those from other directions are attenuated. If the system is so designed that it can allow for multiple beamforming at the same time, it can potentially separate speech streams from different directions. To design a better beamformer, deep learning techniques have been exploited recently to improve mask estimation (Erdogan et al., 2016; Heymanna et al., 2017; Xiao et al., 2017). In addition to segregation, having the system track and hold attention on the speaker of interest is another challenge, for which making the beamformer target-aware (speaker-aware) is one possible solution (Zmolikova et al., 2017; Zhou and Qian, 2018).

Because a microphone array is much cheaper and is more widely deployed than before, multi-channel techniques will become more and more important. However, the single-channel techniques are still indispensable. This is because many recording devices still do not have built-in microphone arrays, and also because when the two speakers talk in the same direction, spatial filtering techniques cannot separate them. Furthermore, the majority of multi-channel techniques exploit only acoustic physics and signal processing techniques. They often do not learn regularities from the training set like single-channel techniques do. One important research direction for solving the cocktail party problem is thus to find solutions that combine single- and multi-channel techniques, and signal processing and machine learning methods. Chen et al. (2017a) and Drude and Haeb-Umbach (2017) made some attempts and led to clear improvements. Because PIT affects only the training phase and can be integrated freely with other techniques, we believe that the combination of PIT and multi-channel techniques is promising.

Speech separation is just the first step in solving the cocktail party problem, although it can improve human–human communication. In many scenarios, the goal is to improve human–computer interaction, in which case ASR is an important component. As we have shown in this paper, unlike humans who can pay attention to one or at most two speakers at the same time, computers can separate mixed speech into multiple speech streams and recognize all of them simultaneously.

Recognizing speech in the cocktail party environment usually takes one of two approaches. In the first approach which is more widely used due to its flexibility, we first separate the mixed speech into multiple speech streams and then recognize each stream using a single-talker recognizer. Because the speech separation module is not perfect and usually introduces nonlinear distortions to the separated speech streams, the recognizer needs to be refined using the separated speech. In the second approach, there is no explicit separation component. Instead, the system is optimized end to end to improve the recognition accuracy. Under this setup, PIT is an indispensable component because other similar techniques cannot solve the label permutation problem at the recognition level. Because mixed speech data can be synthesized as much as we would want, none of these two approaches would suffer from the data sparsity problem. The end-to-end direct approach has three advantages. First, it is simple architecturally, which may be beneficial from an engineering point of view. Second, it may automatically sacrifice separation quality for a better overall recognition accuracy, similar to what is observed in humans' behavior. Third, the system may be combined more easily with other techniques such as speaker adaptation and sequence discriminative training. On the other hand, the two-stage approach can benefit both human–human communication and human–computer interaction with one system. In addition, such system may be easier to explain because the

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

quality of the intermediate separated speech can be evaluated independently. We believe that research will continue in both directions and some mixture of these two approaches will finally win.

Although we have made some advances in the cocktail party problem, e.g., we can now use a single system to recognize single-, two-, and three-talker mixed speech without sacrificing the recognition accuracy on the single-talker speech, there is still room for further improvement. For example, analysis has shown that for the opposite-gender mixed speech, both DPCL and PIT can improve SDR by about 12 dB. However, for the same-gender mixed speech, the improvement is only 8 dB. Our study suggests that this is because in the same-gender case the speech tracing is harder. Besides using multi-channel information, we foresee several other possible ways to improve the performance, including using finer speech features, operating at the raw wave signal level, using more powerful deep learning models, simultaneously identifying speakers and separating speech, and exploiting the language model and decoding graph information.

References

Abdel-Hamid O, Mohamed A, Jiang H, et al., 2014. Convolutional neural networks for speech recognition. Annual Conf of Int Speech Communication Association, p.1533-1545.

Anguera X, Wooters C, Hernando J, 2007. Acoustic beamforming for speaker diarization of meetings. *IEEE Trans Audio Speech Lang Process*, 15(7):2011-2022. https://doi.org/10.1109/TASL.2007.902460

Applebaum S, 1976. Adaptive arrays. *IEEE Trans Antennas Propag*, 24(9):585-598. https://doi.org/10.1109/TAP.1976.1141417

Barker J, Ma N, Coy A, et al., 2010. Speech fragment decoding techniques for simultaneous speaker identification and speech recognition. *Comput Speech Lang*, 24(1):94-111. https://doi.org/10.1016/j.csl.2008.05.003

Behnke S, 2003. Discovering hierarchical speech features using convolutional non-negative matrix factorization. Int Joint Conf on Neural Networks, p.2758-2763. https://doi.org/10.1109/IJCNN.2003.1224004

Bello RWJ, 2010. Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization. 11th Int Society for Music Information Retrieval Conf, p.123-128.

Benesty J, Chen J, Huang Y, et al., 2007. On microphone-array beamforming from a MIMO acoustic signal processing perspective. *IEEE Trans Audio Speech Lang Process*, 15(3):1053-1065. https://doi.org/10.1109/TASL.2006.885251

Benesty J, Chen J, Huang Y, 2008. Automatic Speech Recognition: a Deep Learning Approach. Springer Berlin Heidelberg, New York, USA.

Bi M, Qian Y, Yu K, 2015. Very deep convolutional neural networks for LVCSR. 16th Annual Conf of Int Speech Communication Association, p.3259-3263.

Bregman AS, 1990. Auditory scene analysis. In: Smelzer NJ, Bates PB (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier, Amsterdam.

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

Capon J, 1969. High resolution frequency-wavenumber spectrum analysis. *Proc IEEE*, 57:1408-1418. https://doi.org/10.1109/PROC.1969.7278

Carter GC, Nuttall AH, Cable PG, 1973. The smoothed coherence transform. *Proc IEEE*, 61:1497-1498. https://doi.org/10.1109/PROC.1973.9300

Chang X, Qian Y, Yu D, 2018. Adaptive permutation invariant training with auxiliary information for monaural multi-talker speech recognition. Int Conf on Acoustics, Speech, and Signal Processing, in press.

Chen J, Benesty J, Huang Y, 2006. Time delay estimation in room acoustic environments: an overview. *EURASIP J Adv Signal Process*, 2006:026503. https://doi.org/10.1155/ASP/2006/26503

Chen N, Qian Y, Yu K, 2015. Multi-task learning for text-dependent speaker verification. Annual Conf of Int Speech Communication Association, p.185-189.

Chen Z, 2017. Single Channel Auditory Source Separation with Neural Network. PhD Thesis, Columbia University, New York, USA.

Chen Z, Ellis DP, 2013. Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition. Workshop on Applications of Signal Processing to Audio and Acoustics, p.1-4. https://doi.org/10.1109/WASPAA.2013.6701883

Chen Z, McFee B, Ellis DP, 2014. Speech enhancement by low-rank and convolutive dictionary spectrogram decomposition. Annual Conf of Int Speech Communication Association, p.2833-2837.

Chen Z, Li J, Xiao X, et al., 2017a. Cracking the cocktail party problem by multi-beam deep attractor network. IEEE Workshop on Automatic Speech Recognition and Understanding, p.437-444.

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

Chen Z, Droppo J, Li J, et al., 2017c. Progressive joint modeling in unsupervised single-channel overlapped speech recognition. http://arxiv.org/abs/1707.07048

Cherry EC, 1953. Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Am*, 25(5):975-979. https://doi.org/10.1121/1.1907229

Cooke M, Hershey JR, Rennie SJ, 2010. Monaural speech separation and recognition challenge. *Comput Speech Lang*, 24(1):1-15. https://doi.org/10.1016/j.csl.2009.02.006

Dehak N, Kenny PJ, Dehak R, et al., 2011. Front-end factor analysis for speaker verification. *IEEE Trans Audio Speech Lang Process*, 19(4):788-798. https://doi.org/10.1109/TASL.2010.2064307

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Doclo S, Moonen M, 2003. Design of far-field and near-field broadband beamformers using eigenfilters. *IEEE Signal Process Lett*, 83(12):2641-2673. https://doi.org/10.1016/j.sigpro.2003.07.005

Drude L, Haeb-Umbach R, 2017. Tight integration of spatial and spectral features for BSS with deep clustering embeddings. Annual Conf of Int Speech Communication Association, p.2650-2654.

Du J, Tu Y, Xu Y, et al., 2014. Speech separation of a target speaker based on deep neural networks. Int Conf on Signal Processing, p.473-477. https://doi.org/10.1109/ICOSP.2014.7015050

Ellis DPW, 1996. Prediction-Driven Computational Auditory Scene Analysis. PhD Thesis, Massachusetts Institute of Technology, Cambridge, USA.

Ephraim Y, Malah D, 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans Audio Speech Lang Process*, 33(2): 443-445. https://doi.org/10.1109/TASSP.1985.1164550

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

Erdogan H, Hershey J, Watanabe S, et al., 2016. Improved MVDR beamforming using single-channel mask prediction networks. Annual Conf of Int Speech Communication Association, p.1981-1985.

Erdogan H, Hershey JR, Watanabe S, et al., 2017. Deep recurrent networks for separation and recognition of single-channel speech in nonstationary background audio. New Era for Robust Speech Recognition, p.165-186. https://doi.org/10.1007/978-3-319-64680-0_7

Fischer S, Simmer KU, 1996. Beamforming microphone arrays for speech acquisition in noisy environments. *Speech Commun*, 20(3-4):215-227. https://doi.org/10.1016/S0167-6393(96)00054-4

Frost OL, 1972. An algorithm for linearly constrained adaptive array processing. *Proc IEEE*, 60(8):926-935. https://doi.org/10.1109/PROC.1972.8817

Gannot S, Burshtein D, Weinstein E, 2001. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Trans Signal Process*, 49(8): 1614-1626. https://doi.org/10.1109/78.934132

Gannot S, Burshtein D, Weinstein E, 2004. Analysis of the power spectral deviation of the general transfer function GSC. *IEEE Trans Signal Process*, 52(4):1115-1120. https://doi.org/10.1109/TSP.2004.823487

Ghahramani Z, Jordan MI, 1996. Factorial hidden Markov models. NIPS, p.472-478.

Hassab JC, Boucher RE, 1981. Performance of the generalized cross correlator in the presence of a strong spectral peak in the signal. *IEEE Trans Audio Speech Lang Process*, 29(3):549-555. https://doi.org/10.1109/TASSP.1981.1163613

Hershey JR, Kristjansson T, Rennie S, et al., 2007. Single channel speech separation using factorial dynamics. NIPS, p.593-600.

Hershey JR, Rennie SJ, Olsen PA, et al., 2010. Super-human multi-talker speech recognition: a graphical modeling approach. *Comput Speech Lang*, 24(1):45-66. https://doi.org/10.1016/j.csl.2008.11.001

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

Heymanna J, Drudea L, Haeb-Umbacha R, 2017. A generic neural acoustic beamforming architecture for robust multi-channel speech processing. *Comput Speech Lang*, 46(C):374-385. https://doi.org/10.1016/j.csl.2016.11.007

Hinton G, Deng L, Yu D, et al., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Mag*, 29(6):82-97. https://doi.org/10.1109/MSP.2012.2205597

Hoyer PO, 2004. Non-negative matrix factorization with sparseness constraints. *J Mach Learn Res*, 5:1457-1469.

Hu G, Wang D, 2004. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans Neur Netw*, 15(5):1135-1150. https://doi.org/10.1109/TNN.2004.832812

Hu G, Wang D, 2008. Segregation of unvoiced speech from nonspeech interference. *J Acoust Soc Am*, 124(2): 1306-1319. https://doi.org/10.1121/1.2939132

Hu G, Wang D, 2010. A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Trans Audio Speech Lang Process*, 18(8):2067-2079. https://doi.org/10.1109/TASL.2010.2041110

Hu K, Wang D, 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

Hu Y, Loizou PC, 2007. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Commun*, 49(7):588-601. https://doi.org/10.1016/j.specom.2006.12.006

Hu Y, Loizou PC, 2008. Evaluation of objective quality measures for speech enhancement. *IEEE Trans Audio Speech Lang Process*, 16(1):229-238. https://doi.org/10.1109/TASL.2007.911054

Huang Z, Wang S, Qian Y, 2018. Joint *i*-vector with end-to-end system for short duration text-independent speaker verification. Int Conf on Acoustics, Speech, and Signal Processing, in press.

Hyvarinen A, Karhunen J, Oja E, 2001. Independent Component Analysis. John Wiley & Sons, Inc, New York, USA.

Isik Y, Roux JL, Chen Z, et al., 2016. Single-channel multi-speaker separation using deep clustering. Annual Conf of Int Speech Communication Association, p.545-549. https://doi.org/10.21437/Interspeech.2016-1176

Kellermann W, 1997. Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays. Int Conf on Acoustics Speech and Signal Processing, p.219-222. https://doi.org/10.1109/ICASSP.1997.599608

Kim T, Attias HT, Lee SY, et al., 2006. Blind source separation exploiting higher-order frequency dependencies. *IEEE Trans Audio Speech Lang Process*, 15(4):70-79. https://doi.org/10.1109/TASL.2006.872618

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Kjems U, Boldt JB, Pedersen MS, et al., 2009. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *J Acoust Soc Am*, 126(3):1415-1426. https://doi.org/10.1121/1.3179673

Knapp CK, Carter GC, 1976. The generalized correlation method for estimation of time delay. *IEEE Trans Audio Speech Lang Process*, 24(4):320-327. https://doi.org/10.1109/TASSP.1976.1162830

Kolbæk M, Yu D, Tan ZH, et al., 2017a. Joint separation and denoising of noisy multi-talker speech using recurrent neural networks and permutation invariant training. IEEE Int Workshop on Machine Learning for Signal Processing. http://arxiv.org/abs/1708.09588

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

Kristjansson T, Hershey J, Olsen P, et al., 2006. Super-human multi-talker speech recognition: the IBM 2006 speech separation challenge system. Int Conf on Spoken Language Processing, Paper 1775-Mon1WeS.7.

Kuhl PK, 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Percept Psychol*, 50(2):93-107. https://doi.org/10.3758/BF03212211

Larcher A, Lee KA, Ma B, et al., 2014. Text-dependent speaker verification: classifiers, databases and RSR2015. *Speech Commun*, 60:56-77. https://doi.org/10.1016/j.specom.2014.03.001

Lee DD, Seung HS, 2001. Algorithms for non-negative matrix factorization. NIPS, p.556-562.

Lee TW, 1998. Independent Component Analysis—Theory and Applications. Kluwer Academic Publishers, Boston, USA.

Lei Y, Scheffer N, Ferrer L, et al., 2014. A novel scheme for speaker recognition using a phonetically-aware deep neural network. Int Conf on Acoustics Speech and Signal Processing, p.1695-1699. https://doi.org/10.1109/ICASSP.2014.6853887

Li P, Guan Y, Wang S, et al., 2010. Monaural speech separation based on MAXVQ and CASA for robust speech recognition. *Comput Speech Lang*, 24(1):30-44. https://doi.org/10.1016/j.csl.2008.05.005

Liu Y, Qian Y, Chen N, et al., 2015. Deep feature for text-dependent speaker verification. *Speech Commun*, 73:1-13. https://doi.org/10.1016/j.specom.2015.07.003

Lovekin JM, Yantorno RE, Krishnamachari KR, et al., 2001. Developing usable speech criteria for speaker identification technology. Int Conf on Acoustics Speech and Signal Processing, p.421-424. https://doi.org/10.1109/ICASSP.2001.940857

Mandel MI, Weiss RJ, Ellis DPW, 2010. Model-based expectation maximization source separation and localization. *IEEE Trans Audio Speech Lang Process*, 18(2):382-394. https://doi.org/10.1109/TASL.2009.2029711

McDermott JH, 2009. The cocktail party problem. *Curr Biol*, 19(22):R1024-R1027.

Mesgarani N, Chang EF, 2012. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233-236. https://doi.org/10.1038/nature11020

Mowlaee P, Saeidi R, Tan ZH, et al., 2010. Joint single-channel speech separation and speaker identification. Int Conf on Acoustics Speech and Signal Processing, p.4430-4433. https://doi.org/10.1109/ICASSP.2010.5495619

Mowlaee P, Saeidi R, Christensen MG, et al., 2012. A joint approach for single-channel speaker identification and speech separation. *IEEE Trans Audio Speech Lang Process*, 20(9):2586-2601. https://doi.org/10.1109/TASL.2012.2208627

Narayanan A, Wang D, 2013. Ideal ratio mask estimation using deep neural networks for robust speech recognition. Int Conf on Acoustics Speech and Signal Processing, p.7092-7096. https://doi.org/10.1109/ICASSP.2013.6639038

Ono N, 2011. Stable and fast update rules for independent vector analysis based on auxiliary function technique. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. https://doi.org/10.1109/ASPAA.2011.6082320

Peddinti V, Povey D, Khudanpur S, 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. Annual Conf of Int Speech Communication Association, p.3214-3218.

Pedersen MS, Larsen J, Kjems U, et al., 2007. A Survey of Convolutive Blind Source Separation Methods. Springer Press, New York, USA.

Qian YM, Bi M, Tan T, et al., 2016. Very deep convolutional neural networks for noise robust speech recognition. *IEEE Trans Audio Speech Lang Process*, 24(12):2263-2276. https://doi.org/10.1109/TASLP.2016.2602884

Qian YM, Chang XK, Yu D, 2017. Single-channel multi-talker speech recognition with permutation invariant training. http://arxiv.org/abs/1707.06527

Qian YM, Tan T, Hu H, et al., 2018. Noise robust speech recognition on Aurora4 by humans and machines. Int Conf on Acoustics, Speech, and Signal Processing, in press.

Raj B, Virtanen T, Chaudhuri S, et al., 2010. Non-negative matrix factorization based compensation of music for automatic speech recognition. Annual Conf of Int Speech Communication Association, p.717-720.

Rennie SJ, Hershey JR, Olsen PA, 2010. Single-channel multitalker speech recognition. *IEEE Signal Process Mag*, 27(6):66-80. https://doi.org/10.1109/MSP.2010.938081

Reynolds DA, Quatieri TF, Dunn RB, 2000. Speaker verification using adapted gaussian mixture models. *Dig Signal Process*, 10(1-3):19-41. https://doi.org/10.1006/dspr.1999.0361

Rix AW, Beerends JG, Hollier MP, et al., 2001. Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs. Int Conf on Acoustics, Speech, and Signal Processing, p.749-752. https://doi.org/10.1109/ICASSP.2001.941023

Roth P, 1971. Effective measurements using digital signal analysis. *IEEE Spectr*, 8(4):62-70.

Sainath TN, Mohamed A, Kingsbury B, et al., 2013. Deep convolutional neural networks for LVCSR. Int Conf on Acoustics Speech and Signal Processing, p.8614-8618.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Sainath TN, Vinyals O, Senior A, et al., 2015. Convolutional, long short-term memory, fully connected deep neural networks. Int Conf on Acoustics Speech and Signal Processing, p.4580-4584. https://doi.org/10.1109/ICASSP.2015.7178838

Sawada H, Araki S, Makino S, 2007. A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, p.139-142. https://doi.org/10.1109/ASPAA.2007.4393012

Schmidt MN, Olsson RK, 2006. Single-channel speech separation using sparse non-negative matrix factorization. Annual Conf of Int Speech Communication Association, Paper 1652-Thu2FoP.10.

Schuller B, Weninger F, Wöllmer M, et al., 2010. Non-negative matrix factorization as noise-robust feature extractor for speech recognition. Int Conf on Acoustics Speech and Signal Processing, p.4562-4565. https://doi.org/10.1109/ICASSP.2010.5495567

Sercu T, Puhrsch C, Kingsbury B, et al., 2016. Very deep multilingual convolutional neural networks for LVCSR. Int Conf on Acoustics Speech and Signal Processing, p.4955-4959. https://doi.org/10.1109/ICASSP.2016.7472620

Shao Y, Wang D, 2003. Co-channel speaker identification using usable speech extraction based on multi-pitch tracking. Int Conf on Acoustics, Speech, and Signal Processing, p.205-208. https://doi.org/10.1109/ICASSP.2003.1202330

Shao Y, Wang D, 2006. Model-based sequential organization in cochannel speech. *IEEE Trans Audio Speech Lang Process*, 14(1):289-298. https://doi.org/10.1109/TSA.2005.854106

Souden M, Benesty J, Affes S, 2010. On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Trans Signal Process*, 18(2):260-276. https://doi.org/10.1109/TASL.2009.2025790

Souden M, Araki S, Kinoshita K, et al., 2013. A multichannel MMSE-based framework for speech source separation and noise reduction. *IEEE Trans Signal Process*, 21(9):1913-1928. https://doi.org/10.1109/TASL.2013.2263137

Sydow C, 1994. Broadband beamforming for a microphone array. *J Acoust Soc Am*, 96(8):845-849. https://doi.org/10.1121/1.410323

Taal CH, Hendriks RC, Heusdens R, et al., 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. Int Conf on Acoustics Speech and Signal Processing, p.4214-4217. https://doi.org/10.1109/ICASSP.2010.5495701

Tan T, Qian Y, Yu D, 2018. Knowledge transfer in permutation invariant training for single-channel multi-talker speech recognition. Int Conf on Acoustics, Speech, and Signal Processing, in press.

Tu Y, Du J, Xu Y, et al., 2014a. Deep neural network based speech separation for robust speech recognition. Int Conf on Signal Processing, p.532-536. https://doi.org/10.1109/ICOSP.2014.7015061

Tu Y, Du J, Xu Y, et al., 2014b. Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers. Int Symp on Chinese Spoken Language Processing, p.250-254. https://doi.org/10.1109/ISCSLP.2014.6936615

Variani E, Lei X, McDermott E, et al., 2014. Deep neural networks for small footprint text-dependent speaker verification. Int Conf on Acoustics Speech and Signal Processing, p.4052-4056. https://doi.org/10.1109/ICASSP.2014.6854363

Vincent E, Gribonval R, Févotte 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

Virtanen T, 2006. Speech recognition using factorial hidden Markov models for separation in the feature space. Annual Conf of Int Speech Communication Association, Paper 1850-Mon1WeS.5.

Virtanen T, 2007. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans Audio Speech Lang Process*, 15(3):1066-1074. https://doi.org/10.1109/TASL.2006.885253

Wang D, 2005. On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi P (Ed.), Speech Separation by Humans and Machines. Springer, Boston, USA, p.181-197. https://doi.org/10.1007/0-387-22794-6_12

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

Wang S, Qian Y, Yu K, 2018. Focal KL-divergence based dilated convolutional neural networks for co-channel speaker identification. Int Conf on Acoustics, Speech, and Signal Processing, in press.

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

Weng C, Yu D, Seltzer ML, et al., 2015. Deep neural networks for single-channel multi-talker speech recognition. *IEEE Trans Audio Speech Lang Process*, 23(10):1670-1679. https://doi.org/10.1109/TASLP.2015.2444659

Weninger F, Erdogan H, Watanabe S, et al., 2015. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. Int Conf on Latent Variable Analysis and Signal Separation, p.91-99. https://doi.org/10.1007/978-3-319-22482-4_11

Xiao X, Zhao SK, Jones DL, et al., 2017. On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition. Int Conf on Acoustics Speech and Signal Processing, p.3246-3250. https://doi.org/10.1109/ICASSP.2017.7952756

Xiong W, Droppo J, Huang X, et al., 2016. Achieving human parity in conversational speech recognition. http://arxiv.org/abs/1610.05256

Xu Y, Du J, Dai LR, et al., 2014. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process Lett*, 21(1):65-68. https://doi.org/10.1109/LSP.2013.2291240

Yilmaz O, Rickard S, 2004. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans Signal Process*, 52(7):1830-1847. https://doi.org/10.1109/TSP.2004.828896

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Yu D, Deng L, 2014. Automatic Speech Recognition: a Deep Learning Approach. Springer, New York, USA.

Yu D, Li, JY, 2017. Recent progresses in deep learning based acoustic models. *IEEE/CAA J Automat Sin*, 4(3):396-409. https://doi.org/10.1109/JAS.2017.7510508

Yu D, Xiong W, Droppo J, et al., 2016. Deep convolutional neural networks with layer-wise context expansion and attention. Annual Conf of Int Speech Communication Association, p.17-21. https://doi.org/10.21437/Interspeech.2016-251

Yu D, Chang X, Qian Y, 2017a. Recognizing multi-talker speech with permutation invariant training. Annual Conf of Int Speech Communication Association, p.2456-2460.

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

Yu F, Koltun V, 2015. Multi-scale context aggregation by dilated convolutions. http://arxiv.org/abs/1511.07122

Zhang C, Koishida K, 2017. End-to-end text-independent speaker verification with triplet loss on short utterances. Annual Conf of Int Speech Communication Association, p.1487-1491. https://doi.org/10.21437/Interspeech.2017-1608

Zhang L, Chen Z, Zheng M, et al., 2011. Robust non-negative matrix factorization. *Front. Electr. Electron. Eng. China*, 6(2):192-200. https://doi.org/10.1007/s11460-011-0128-0

Zhao X, Wang Y, Wang D, 2015a. Cochannel speaker identification in anechoic and reverberant conditions. *IEEE Trans Audio Speech Lang Process*, 23(11):1727-1736. https://doi.org/10.1109/TASLP.2015.2447284

Zhao X, Wang Y, Wang D, 2015b. Deep neural networks for cochannel speaker identification. Int Conf on Acoustics, Speech and Signal Processing, p.4824-4828. https://doi.org/10.1109/ICASSP.2015.7178887

Zhou Y, Qian Y, 2018. Robust mask estimation by integrating neural network-based and clustering-based approaches for adaptive acoustic beamforming. Int Conf on Acoustics, Speech, and Signal Processing, in press.

Zmolikova K, Delcroix M, Kinoshita K, et al., 2017. Speaker-aware neural network based beamformer for speaker extraction in speech mixtures. Annual Conf of Int Speech Communication Association, p.2655-2659. https://doi.org/10.21437/Interspeech.2017-667

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net