Speech Communication 106 (2019) 31–43

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

An iterative mask estimation approach to deep learning based multi-channel speech recognition

Yan-Hui Tu[a], Jun Du[a,*], Lei Sun[a], Feng Ma[b], Hai-Kun Wang[b], Jing-Dong Chen[c], Chin-Hui Lee[d]

[a] University of Science and Technology of China, Hefei, Anhui, China
[b] iFlytek Co., Ltd., Hefei, Anhui, China
[c] Northwestern Polytechnical University, Xian, Shaanxi, China
[d] Georgia Institute of Technology, Atlanta, Georgia, USA

ARTICLE INFO

Keywords:
CHiME challenge
Deep learning
Ideal ratio mask (IRM)
Microphone array
Robust speech recognition

ABSTRACT

We propose a novel iterative mask estimation (IME) framework to improve the state-of-the-art complex Gaussian mixture model (CGMM)-based beamforming approach in an iterative manner by leveraging upon the complementary information obtained from different deep models. Although CGMM has been recently demonstrated to be quite effective for multi-channel, automation speech recognition (ASR) in operational scenarios, the corresponding mask estimation, however, is not always accurate in adverse environments due to the lack of prior or context information. To address this problem, in this study, a neural-network-based ideal ratio mask estimator learned from a multi-condition data set is first adopted to incorporate prior information, obtained from the speech/noise interactions and the long acoustic context, into CGMM-based beamformed speech that has a higher signal-to-noise ratio (SNR) than the original noisy speech signal. Next, to further utilize the rich context information in deep acoustic and language models, voice activity detection information, obtained from speech recognition results, is then used to refine mask estimation, yielding a significant reduction in insertion errors. During testing on the recently launched CHiME-4 Challenge ASR task of recognizing 6-channel microphone array speech, the proposed IME approach significantly and consistently outperforms the CGMM approach under different configurations, with relative word error rate reductions ranging from 20% to 30%. Furthermore, the IME approach plays a key role in the ensemble system that achieves the best performance in the CHiME-4 Challenge.

1. Introduction

Recently, hands-free speech communication is in high demand for many applications, such as multi-microphone portable devices and automatic speech recognition (ASR) systems, due to the provided convenience and flexibility. However, the ASR performance is often severely degraded when the target speech signals are corrupted by interfering speakers, background noises and room reverberation. Speech enhancement algorithms that reduce noise without considerably damaging the target speech are therefore highly desired for improving the ASR performance and robustness. Over the past several decades, many algorithms have been developed, and they can be divided into two broad categories. The first is single-channel speech enhancement, which exploits only the temporal and spectral information. Representative algorithms in this category include spectral subtraction (Boll, 1979), Wiener filtering (Lim and Oppenheim, 1979), minimum mean-square error (MMSE) estimator (Ephraim and Malah, 1984), and the optimally modified log-spectral amplitude (OM-LSA) speech estimator (Cohen and Berdugo, 2001).

The second category is multi-channel speech enhancement, which uses spatial information in addition to the temporal and spectral information. Representative algorithms in this category include multi-channel Wiener filtering (Meyer and Simmer, 1997; Spriet et al., 2004), blind source separation (Jutten and Herault, 1991; Buchner et al., 2005; Wang et al., 2011), and beamforming (Van Veen and Buckley, 1988; Cox et al., 1987; Hoshuyama et al., 1999; Talmon et al., 2009; Souden et al., 2010; Krueger et al., 2011; Higuchi et al., 2016). Beamforming is a popular approach for multi-channel speech enhancement, where its performance dependents on constructing a steering vector that represents the acoustic propagation (Veen and Buckley, 1988). Conventionally, the beamformers utilizing *a priori* knowledge, e.g., the geometry of the microphone array and the direction of arrival (DOA) information, to construct the steering vector. They may work well for simulated data where the prior information is accessible and accurate, e.g., the baseline beamformer (Anguera et al., 2007) provided by the CHiME-3 challenge (Barker et al., 2017). However, the robustness of such beamform-

* Corresponding author.
E-mail address: jundu@ustc.edu.cn (J. Du).

https://doi.org/10.1016/j.specom.2018.11.005
Received 8 March 2018; Received in revised form 9 October 2018; Accepted 26 November 2018
Available online 26 November 2018

ers often becomes an issue in real-life environments where the acoustic propagation information is unknown and is difficult to estimate accurately. Recently, a complex Gaussian mixture model (CGMM)-based time-frequency (T-F) mask estimation algorithm was used to steer a beamformer in Higuchi et al. (2016), which was demonstrated to be beneficial to ASR in RealData scenarios, e.g., in some top-performing CHiME-4 systems (Tu et al., 2017; Menne et al., 2016).

On the other hand, deep learning techniques are becoming increasingly popular in speech recognition areas (Hinton et al., 2012; Mohamed et al., 2012). Different deep neural network (DNN) architectures have been adopted in single-channel speech enhancement for ASR, and they have demonstrated a significant increase in recognition performance (Du et al., 2016a; Weninger et al., 2015; Tu et al., 2015; Gao et al., 2015). Some preliminary studies on using deep learning approaches for multi-channel speech enhancement have also been conducted. In Gao et al. (2016), the signals obtained using multi-channel speech enhancement algorithms were directly used as the input signals for neural-network-based enhancement models. In Heymann et al. (2015), bidirectional long short-term memory (BLSTM) (Hochreiter and Schmidhuber, 1997) was adopted to estimate signal statistics to steer the beamformer for multi-channel speech enhancement. It was also demonstrated in Nugraha et al. (2016) that deep neural network (DNN)-based source spectra estimation is helpful for steering a multi-channel filter. In Sainath et al. (2017), they proposed multichannel enhancement jointly with acoustic modeling in a deep neural network framework. The raw time-domain waveform was directly modeled by beamforming, which leverages upon differences in the fine time structure of the signal at different microphones to filter energy arriving from different directions. In Ochiai et al. (2017), an end-to-end framework was proposed by encompassing microphone array signal processing for noise suppression and speech enhancement within the acoustic encoding network, allowing the beamforming components to be optimized jointly within the recognition architecture to improve the end-to-end speech recognition objective.

In this paper, we propose an iterative mask estimation approach to beamforming by leveraging the information obtained via iterative neural-network-based ideal ratio mask (IRM) (Wang and Wang, 2013) estimation and ASR-based voice activity detection (VAD) (Sohn et al., 1999). The proposed approach has four major contributions to front-end beamforming. First, we use the estimated IRM based on the trained NN model to improve the time-frequency masks estimated using the CGMM-based approach. The CGMM parameters are optimized based on the maximum a posteriori (MAP) estimation (Gauvain and Lee, 1994), and the parameters are generally adjusted based on the estimated speech presence probability at each T-F unit. Thus, we expect that the approach is insensitive to non-stationary noise and also robust to stationary noise. The acoustic context information, including the full frequency band and expanding context frames, can be appropriately utilized for NN-based IRM estimation, avoiding misjudgment in non-stationary noise. Second, an ASR-based VAD from the segmentation results of the recognizer is also used to improve beamforming. The ASR-based VAD can directly provide speech and non-speech segmentation information from ASR, which makes our beamformed speech adapt to ASR. Third, the estimated IRM and ASR-based VAD are updated with better beamformed speech such that they can form a closed-loop optimization by leveraging the information of NN-based IRM and ASR-based VAD. Finally, our approach combines the CGMM-based approach which is an adaptive algorithm to test data, a powerful ability of learning NNs, and the feedback of recognition results to iteratively improve the beamforming performance. Several factors can affect the beamforming and ASR performance of the proposed system, including the number of closed-loop iterations, different NN architectures (including the DNNs and LSTMs used to estimate the T-F mask), and the combination of the estimated masks, which will also be investigated in this paper. In Heymann et al. (2017) and Xiao et al. (2017), the NN-based mask estimator was tuned using an ASR-based cost function, and the front-end enhancement was combined with a back-end recognizer. Although the ASR information are also utilized in these methods, the main motivation, the specific information adopted, and the way to combine multiple types of information in our approach are quite different. Moreover, the experiments show that our approach can achieve better results on the CHiME-4 challenge task according to Heymann et al. (2017) and Xiao et al. (2017).

This work is comprehensively extended from our recent paper Tu et al. (2017) with new contributions listed as follows. First, Tu et al. (2017) only briefly describes the proposed multi-channel ASR system for CHiME-4 evaluation with both front-end and back-end design. However, in this study we focus on elaborating our proposed IME-based beamformer more generally in Section 2, including, in detail, mask-based beamforming in Section 3, the motivations of the proposed IME algorithm in Section 4. Second, in Section 4.2, LSTM-based IRM estimation is also adopted and its complementarity with DNN-based IRM estimation via ensemble systems is shown while Tu et al. (2017) only considers DNN-based IRM estimation to improve mask estimation. Finally, for the experimental design, this work gives detailed descriptions of the comparisons and analyses among the state-of-the-art beamformers and the proposed IME-based beamformer from different perspectives. But Tu et al. (2017) lists very limited and the best results of integrated front-end/back-end systems to show the promising system performance for the CHiME-4 challenge without giving any detailed analyses.

The remainder of this paper is organized as follows. In Section 2, we present an overview of the system. In Section 3, we provide a brief introduction to the conventional beamformers. In Section 4, we present a detailed description of our proposed iterative mask estimation approach. Section 5 presents the ASR performance of our proposed approach on the CHiME-4 challenge. Finally, we summarize our findings in Section 6.

2. The IME framework

The overall system flowchart is shown in Fig. 1. For the IRM estimation, the NN-IRM models are trained using the log-power spectral (LPS) features of data from the reference channel of the microphone array as the input features and the corresponding IRMs as the output features. The LPS features that offer perceptually relevant parameters are adopted (Xie and Van Compernolle, 1994; Du and Huo, 2008; Wan and Nelson, 1998), while IRM is interpreted as the speech presence probability at each time-frequency point in speech separation (Hummersone et al., 2014). For ASR, both log mel-filterbank (LMFB) and feature-space maximum likelihood regression (fMLLR) features (Du et al., 2015) are adopted as acoustic features. Then, acoustic models based on DNN or deep CNN (DCNN) with hidden Markov models (HMMs) are trained using beamformed data and data from all channels. Finally, the LSTM-based language models are constructed. For more ASR details, the readers can refer to Du et al. (2015, 2016b).

The IME-based beamformer is divided into four successive steps, namely, beamforming initialization, NN-based signal statistics (IRM and ASR-based VAD) estimation, beamforming, and recognition. First, beamformed speech is initialized, and a T-F mask of test speech is obtained by CGMM-based beamforming with poorly estimated initial prior values. Then, the IRM estimated using a trained NN-IRM model is used to improve the initial mask, where the NN-IRM model uses the LPS features of the initial beamformed speech and the ASR-based VAD information from the segmentation results of a recognizer with beamformed speech. Next, the improved mask is adopted to estimate the initial values of the CGMM-based approach to generate the estimated mask that steers the beamformer, thereby obtaining the beamformed speech for ASR (Nakatani et al., 2017). Finally, multiple acoustic models are first fused at the state level, and then first-pass decoding is performed with the HMM and 3-gram to generate lattices as the hypotheses, which subsequently serve for the second-pass decoding with an LSTM-based LM. The details are presented in the following subsections.

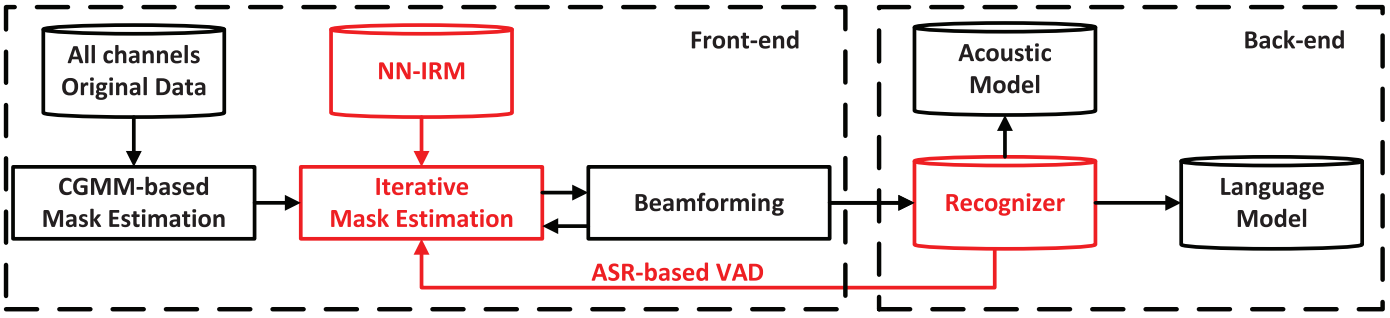


Fig. 1. A block diagram of the entire system, which consists of front-end beamforming and back-end acoustic/language models.

3. Background

In this section, we first present the signal model and then briefly introduce the well-known time-frequency-mask-based beamformer, which serves as the basis for the proposed system.

3.1. Signal model

Given speech $s(t)$ in the target speaker position, the signals received by an array of J microphones are time-delayed and amplitude-attenuated versions of $s(t)$ with extra noises and interferences, which are modeled in the time domain as follows:

$$y_i(t) = g_i s(t-\tau_i) + n_i(t) = x_i(t) + n_i(t), \tag{1}$$

where $i = 1, 2, \ldots, J$; τ_i denotes the time that it takes for the sound to propagation from the speaker location to the ith microphone location; g_i is the acoustic impulse response to model the effects of propagation attenuation, the amplification gain of the corresponding microphone setting and the directionality of the source and the ith microphone; $x_i(t)$ is the convolved speech signal at the ith microphone; and $n_i(t)$ is the noise plus inference signal received by the ith microphone. In the short-time Fourier transform (STFT) domain (Mcaulay and Quatieri, 1986), the signal model in (1) can be expressed as follows (Zhang et al., 2008):

$$\boldsymbol{y}(k,l) = \boldsymbol{g}(k)s(k,l) + \boldsymbol{n}(k,l) = \boldsymbol{x}(k,l) + \boldsymbol{n}(k,l), \tag{2}$$

where k is the frequency bin index; l is the frame index; $\boldsymbol{x}(k, l)$ and $\boldsymbol{n}(k, l)$ are J-dimensional complex vectors that consist of the STFT-domain representations of $x_i(t)$ and $n_i(t)$, respectively; $s(k, l)$ is the STFT of $s(t)$; and $\boldsymbol{g}(k)$ is the signal propagation vector, which is in the same form as the so-called steering vector in the array beamforming literature (Veen and Buckley, 1988). We assume that the analysis window is longer than all the channel impulse responses and that $\boldsymbol{n}(k, l)$ is relatively stationary.

3.2. MVDR beamformer

The MVDR beamformer applies a set of weights $\boldsymbol{w}(k)$ to the vector $\boldsymbol{y}(k, l)$ such that the variance of the noise component in the beamformer's output is minimized subject to a constraint of unity gain in the target direction, i.e.,

$$\min_{\boldsymbol{w}(k)} \ \boldsymbol{w}^H(k)\boldsymbol{R}_{\boldsymbol{nn}}(k)\boldsymbol{w}(k), \qquad \text{s.t. } \boldsymbol{w}^H(k)\boldsymbol{g}(k) = 1, \tag{3}$$

where the superscript H denotes the conjugate transpose operator and

$$\boldsymbol{R}_{\boldsymbol{nn}}(k) = E\left[\boldsymbol{n}(k,l)\boldsymbol{n}^H(k,l)\right] \tag{4}$$

is the spatial correlation matrix of the noise and interference. A closed-form solution of Eq. (3) is the following (Capon, 1969):

$$\boldsymbol{w}(k) = \frac{\boldsymbol{R}_{\boldsymbol{nn}}^{-1}(k)\boldsymbol{g}(k)}{\boldsymbol{g}^H(k)\boldsymbol{R}_{\boldsymbol{nn}}^{-1}(k)\boldsymbol{g}(k)}. \tag{5}$$

3.3. Beamforming based on time-frequency mask

Implementing the MVDR beamformer given in (5) requires knowledge of the signal propagation vector $\boldsymbol{g}(k)$, which is not accessible in practical environments. A *de facto* standard practice in real applications is to replace this propagation vector with a steering vector. However, this replacement may lead to significant performance degradation because there is always a mismatch between the steering and propagation vectors. In the literature, substantial efforts have been devoted to developing adaptive beamformers that are robust to uncertainty in DOA (Keyi et al., 2005; Zhao et al., 2014) and microphone gains (Zhao et al., 2015). In this work, we adopt the so-called time-frequency-mask-based beamformer (Yoshioka et al., 2015; Higuchi et al., 2016), which uses a spectral-mask-based steering vector without relying on the *a prior* information of either DOA or acoustic propagation. The principal eigenvector of the spatial correlation matrix of the target speech signal is directly used as an estimate of the steering vector. The spatial correlation matrix can be estimated using a time-frequency mask as follows.

Let us assume that the speech signal and noise are statistically independent. The spatial correlation matrix of $\boldsymbol{x}(k, l)$, i.e., $\boldsymbol{R}_{\boldsymbol{xx}}(k)$, can be estimated as follows:

$$\boldsymbol{R}_{\boldsymbol{xx}}(k) = \sum_{l=1}^{T} M(k,l)\boldsymbol{y}(k,l)\boldsymbol{y}^H(k,l), \tag{6}$$

where $M(k, l)$ denotes the time-frequency mask that represents the probability of the T-F unit (k, l) containing the target speech signal. The quantity $1 - M(k, l)$ then represents the probability of the T-F unit (k, l) containing only noise. Thus, the spatial correlation matrix of $\boldsymbol{n}(k, l)$, $\boldsymbol{R}_{\boldsymbol{nn}}(k)$, can be estimated as follows:

$$\boldsymbol{R}_{\boldsymbol{nn}}(k) = \sum_{l=1}^{T} [1 - M(k,l)]\boldsymbol{y}(k,l)\boldsymbol{y}^H(k,l). \tag{7}$$

Based on the assumption that speech and noise are not correlated and the cross term can be ignored, the $\boldsymbol{R}_{\boldsymbol{xx}}(k)$ matrix can be written according to the signal model given in Eq. (2) as follows:

$$\boldsymbol{R}_{\boldsymbol{xx}}(k) = E[\boldsymbol{x}(k,l)\boldsymbol{x}^H(k,l)] = \sigma_s^2(k)\boldsymbol{g}(k)\boldsymbol{g}^H(k), \tag{8}$$

where $\sigma_s^2(k)$ is the variance of $s(k, l)$. Clearly, the positive semi-definite matrix $\boldsymbol{R}_{\boldsymbol{xx}}(k)$ is of rank 1. Consequently, an estimate of the signal propagation vector $\boldsymbol{g}(k)$ can be obtained by computing the principal eigenvector of the $\boldsymbol{R}_{\boldsymbol{xx}}(k)$ estimate from Eq. (6) (Jones and Ratnam, 2009).

The key to this approach becomes the unsupervised and accurate estimation of the spectral masks that indicate the presence and absence of speech T-F units.

3.4. CGMM-based time-frequency mask estimation

In Higuchi et al. (2016), an approach that uses a speech spectral model based on CGMM was proposed to estimate the time-frequency masks. The parameters of the CGMM are full-rank spatial correlation

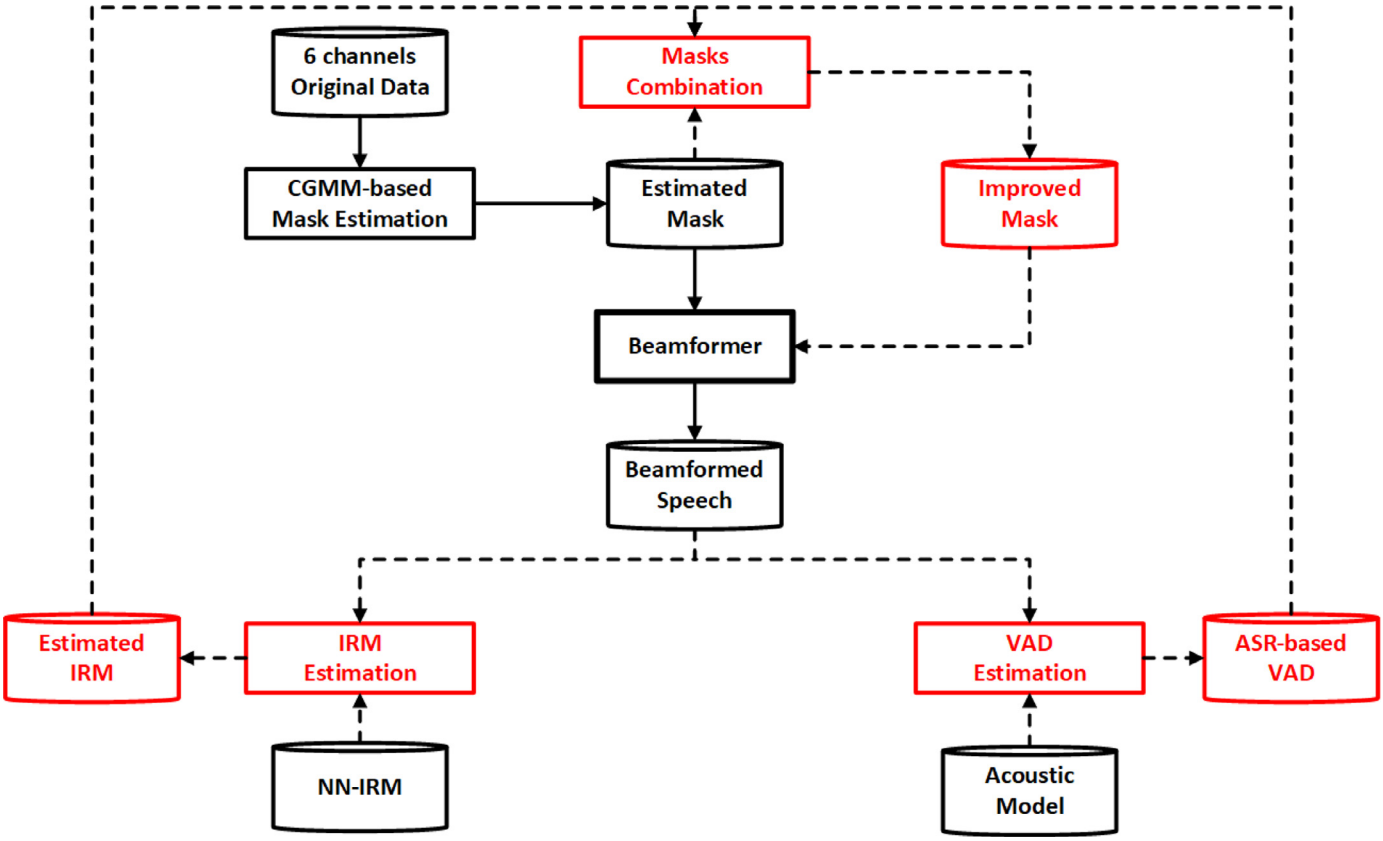


Fig. 2. The framework of iterative mask estimation.

matrices, which provide some flexibility to address the spatial fluctuation of the steering vector. The CGMM parameters, i.e., $\boldsymbol{R}_{\boldsymbol{xx}}(k)$ and $\sigma_s^2(k)$, can be estimated using the expectation-maximization (EM) algorithm (Higuchi et al., 2016) with poorly estimated initial prior values. For example, the initial value of $\boldsymbol{R}_{\boldsymbol{xx}}(k)$ was set as the covariance matrix of the observed signal, and $\boldsymbol{R}_{\boldsymbol{nn}}(k)$ was initialized using an identity matrix.

4. Iterative mask estimation for beamforming

In Section 3, the estimation of the time-frequency masks is only based on the statistical CGMM model. It is an algorithm that is adaptive to the test signal, which is generally not sufficiently robust in adverse environments, particularly when there is burst-type noise. In this section, we discuss NN-based IRM estimation and VAD based on ASR results to improve the masks estimated using the CGMM-based approach. Our experimental results demonstrate that the improved masks can yield significant improvement in the ASR performance. In the next three subsections, the procedure of iterative mask estimation is presented, followed by the elaboration of the NN-based IRM estimation and ASR-based VAD estimation.

4.1. The proposed iterative mask estimation procedure

The iterative mask estimation consists of the following steps:

Step 1: Estimate the initial mask for each T-F unit (k, l), denoted as $M_{\text{CGMM}}(k, l)$, using the CGMM-based approach.
Step 2: Steer the beamformer with the estimated mask and obtain the beamformed speech.
Step 3: Feed the NN-IRM model with the beamformed speech from Step 2 to estimate IRM, denoted as $M_{\text{NN}}(k, l)$.
Step 4: Perform the first-pass decoding with the beamformed speech from Step 2 to get the ASR-based VAD, denoted as $M_{\text{ASR}}(k, l)$.
Step 5: Combine $M_{\text{CGMM}}(k, l)$ in Step 1 with $M_{\text{NN}}(k, l)$ in Step 3 or/and $M_{\text{ASR}}(k, l)$ in Step 4 to generate the improved mask.
Repeat Steps 2–5 for N iterations.

As an illustration in Fig. 2, the solid-line linking modules correspond to **Steps 1–2** of the above procedure while the dotted-line linking modules refer to **Steps 3–5**, namely the combination of CGMM-based, NN-based and ASR-based masks.

4.2. Improving mask estimation by NN-based IRM

First, we use an NN-IRM to predict the mask representing the speech presence probability at every T-F unit given the input LPS features of enhanced speech obtained at Step 2 in Section 4.1. Acoustic context information along both the time axis (with multiple neighboring frames) and frequency axis (with full frequency bins) can be fully exploited by the NN to obtain a good mask estimate in adverse environments, which is strongly complementary with the conventional CGMM-based approach to retain robustness. The estimated IRMs are restricted to be in the range from zero to one, which can be directly used to represent the speech presence probability. In the training stage, the IRM as the learning target is defined as follows:

$$M_{\text{ref}}(k, l) = \sqrt{s^{\text{PS}}(k, l) / \left[s^{\text{PS}}(k, l) + n^{\text{PS}}(k, l) \right]}, \tag{9}$$

where $s^{\text{PS}}(k, l)$ and $n^{\text{PS}}(k, l)$ are clean and noise versions of power spectral features at the T-F unit (k, l). Because the training of this NN-IRM model requires a large amount of time-synchronized stereo data with the IRM and LPS of enhanced training data pairs, the training data are synthesized by adding different types of noise to the clean speech utterances with different SNR levels. Note that the specified SNR levels in

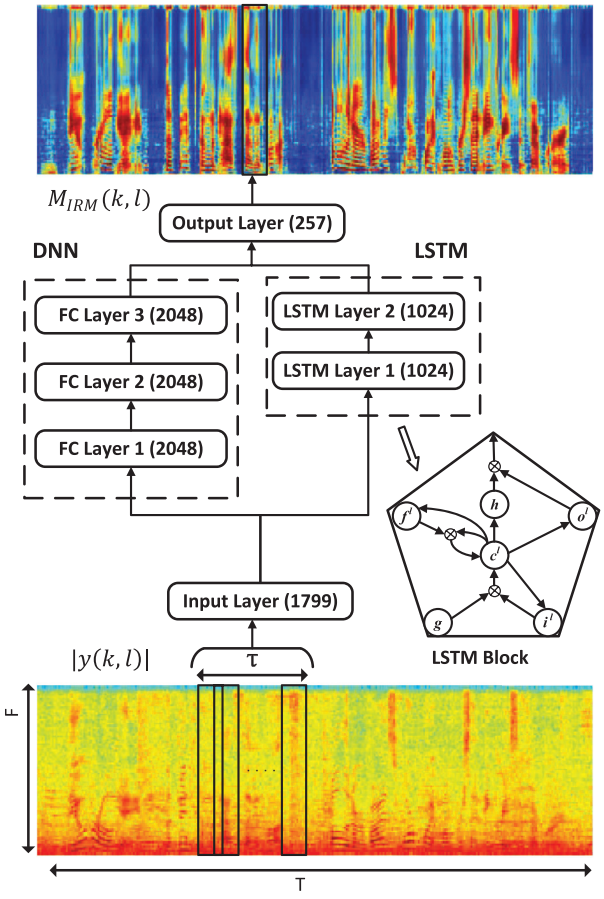


Fig. 3. The architectures of DNN and LSTM for IRM estimation.

the training stage are expected to address the problem of SNR variation in the test stage with real speech data. Then, the estimated $M_{\mathrm{NN}}(k, l)$ is combined with $M_{\mathrm{CGMM}}(k, l)$ to yield an improved mask $M_1(k, l)$, i.e.,

$$M_1(k,l) = \sqrt{M_{\mathrm{CGMM}}(k,l) M_{\mathrm{NN}}(k,l)}. \tag{10}$$

This process can repeat iteratively following Steps 2–5 in Section 4.1.

To train the NN model, supervised fine-tuning is used to minimize the mean squared error (MSE) between the NN-IRM output $M_{\mathrm{NN}}(k, l)$ and the reference IRM $M_{\mathrm{ref}}(k, l)$, which is defined as

$$E_{\mathrm{NN}} = \sum_k \sum_l \left[M_{\mathrm{NN}}(k,l) - M_{\mathrm{ref}}(k,l) \right]^2. \tag{11}$$

This MSE is optimized using the stochastic gradient descent-based back-propagation method in a mini-batch mode.

4.2.1. Architecture of NN models

The NN architecture is shown in Fig. 3 (DNN on the left-hand side and LSTM nn the right-hand side). The input layer of both the DNN and LSTM is a 1799-dimensional vector of noisy LPS features with 7 frame expansions and 257 frequency bins. Each node of the output layer adopts a sigmoid activation function. The estimated IRM at T-F unit (k, l) is denoted as $M_{\mathrm{NN}}(k, l)$. The three hidden layers of the DNN, each with 2048 nodes, are fully connected (FC) with a sigmoid activation function, while two consecutive unidirectional LSTM layers, each with 1024 cells, are adopted. The key components, namely, memory cell state $\boldsymbol{c}^l$, input gate $\boldsymbol{i}^l$, forget gate $\boldsymbol{f}^l$, and output gate $\boldsymbol{o}^l$, are shown in Fig. 3. With this architecture, the network can determine what information to store, update, discard, and output. Furthermore, with a longer acoustic context of history and future information, the LSTM may be robust toward non-stationary noises due to its ability to capture the inherent statistical properties of speech and noise.

4.2.2. The ensemble of DNN-based and LSTM-based IRMs

In Tu et al. (2016), it was shown that different networks may have a strong complementarity and that the corresponding ensemble can result in a considerable improvement in ASR performance. The main difference between the architectures of the DNN and LSTM models is that LSTM introduces the concepts of memory cell and a series of gates to dynamically control the information flow. Thus, the information of neighboring frames is fully utilized in training, whereas the DNN only uses this information as input features. However, LSTM usually with a larger number of parameters than DNN can more easily result in overfitting when the training data are limited. In CHiME-4, the ASR performance of the LSTM model is slightly worse than that of the DNN model for unseen noise cases. The fusion of DNN- and LSTM-based IRM estimates is conducted to further improve the mask estimation. Specifically, a linear combination of $M_{\mathrm{DNN}}(k, l)$ from DNN and $M_{\mathrm{LSTM}}(k, l)$ from LSTM is computed as the ensemble IRM:

$$M_{\mathrm{E}}(k,l) = \alpha_2 M_{\mathrm{DNN}}(k,l) + (1-\alpha_2) M_{\mathrm{LSTM}}(k,l), \tag{12}$$

where α determines the balance between the masks estimated by the two networks. The value of α is set to 0.5 in our experiments.

Fig. 4 presents an utterance example from the RealData test set of CHiME-4 to illustrate the motivation of using NN-based IRM. Fig. 4(a) and (b) plot the spectrograms from channel 0 (the close-talking microphone to record the reference "clean" speech) and channel 5 (one main microphone to record the noisy speech). The CGMM-based approach clearly plays only a limited role in reducing the non-stationary noise, as shown in the marked regions in Fig. 4(c). Fig. 4(d) and (e) plot the two masks estimated by the same DNN model with channel 5 data and CGMM-based beamformed speech as the input of DNN, respectively. Comparing these two plots reveals that the mask estimated directly from channel 5 may misclassify the T-F region dominated by speech to non-speech/noise [e.g., the circled region in Fig. 4(d)], particularly in low SNR conditions, whereas the mask estimated from beamformed speech can generate considerably better results. This result demonstrates the superiority of our approach. Finally, the mask estimated by LSTM with CGMM-based beamformed speech is plotted in Fig. 4(f). This mask retains the high-frequency parts of speech compared with the mask estimated by the DNN, but it may misclassify some noise-only regions. This result illustrates the complementarity of the two estimated masks.

4.3. Improving mask estimation by ASR-based VAD

In some adverse environments, the mask estimated by the CGMM- or NN-based approaches may result in a high false-alarm probability, misclassifying the T-F region dominated by noise, whereas the segmentation results of the speech recognizer are more accurate to handle this problem by using considerably longer acoustic context information in acoustic and language models. Therefore, in our system, the VAD information from the segmentation results of the speech recognizer using beamformed speech at each frame is used to further improve the mask estimation. The VAD-based mask at each T-F unit (k, l) from the ASR results is defined as

$$M_{\mathrm{ASR}}(k,l) \begin{cases} = 1 & \text{if the } l\text{th frame is speech} \\ = 0 & \text{else} \end{cases}. \tag{13}$$

Then, the improved mask using the $M_{\mathrm{ASR}}(k, l)$ is obtained as

$$M_2(k,l) = M_{\mathrm{CGMM}}(k,l) M_{\mathrm{ASR}}(k,l). \tag{14}$$

Note that Eq. (14) only uses the ASR-based VAD information to improve the CGMM-based mask. According to Step 5 in Section 4.1, if both the NN-based mask and ASR-based mask are adopted, $M_{\mathrm{CGMM}}(k, l)$ in Eq. (14) should be replaced by $M_1(k, l)$, i.e.,

$$M_3(k,l) = M_1(k,l) M_{\mathrm{ASR}}(k,l). \tag{15}$$

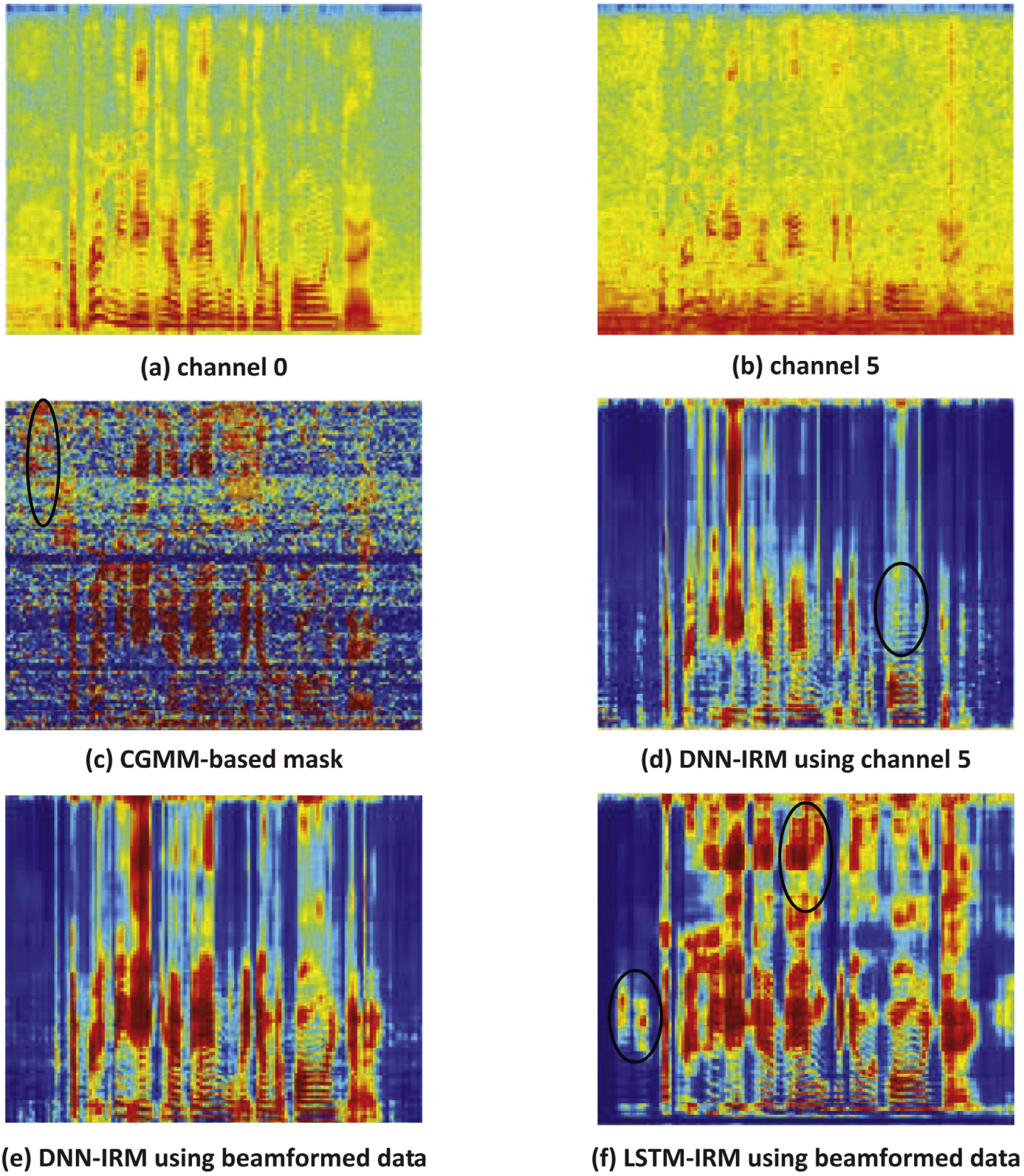
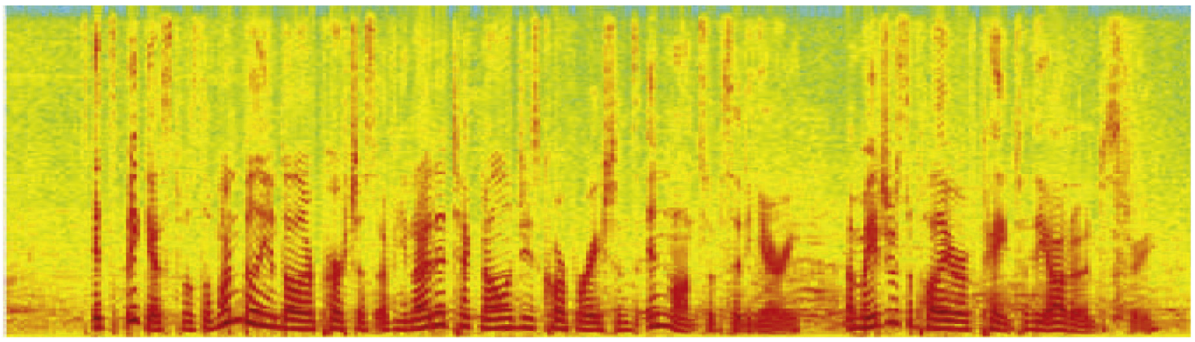


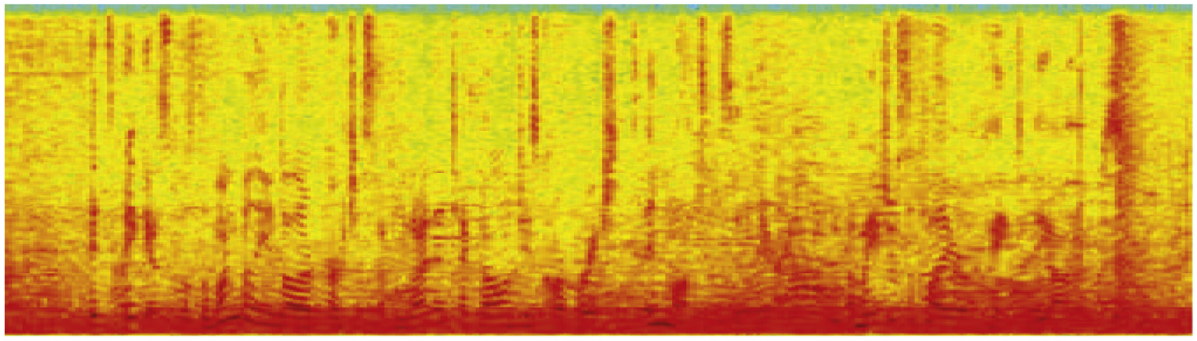
Fig. 4. The comparison of estimated masks from different approaches for an utterance of the CHiME-4 RealData test set.

Similar to $M_1(k, l)$, $M_2(k, l)$ and $M_3(k, l)$ can be iteratively refined by repeating Steps 2–6 of Section 4.1.

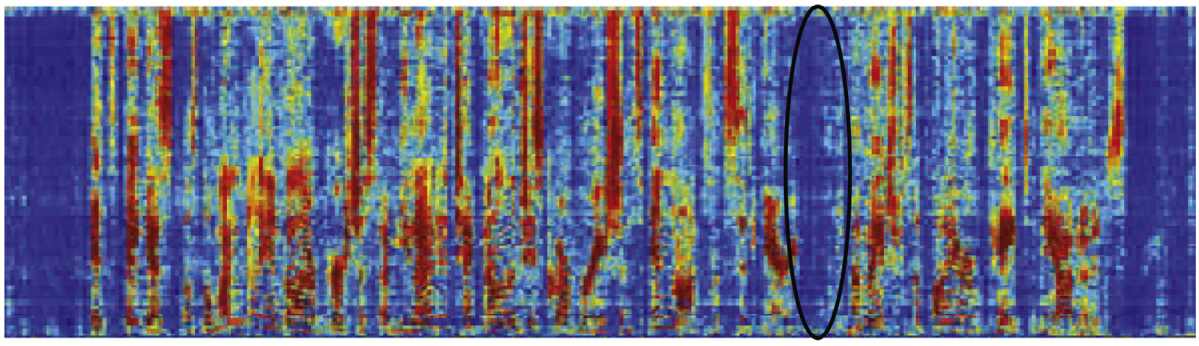
Fig. 5 plots an utterance example from the RealData test set of CHiME-4 to illustrate the motivation for using ASR-based VAD information. Fig. 5(a) and (b) show the spectrograms of channel 0 and channel 5, respectively. Fig. 5(c) is the combination of the CGMM-based mask and the DNN-based IRM, while Fig. 5(d) adds the ASR-based VAD information based on Eq. (15). Based on Eqs. (13)–(15), the VAD information only affects the mask combination in the non-speech segmentations. Accordingly, the non-speech segmentations could be cleaner af-



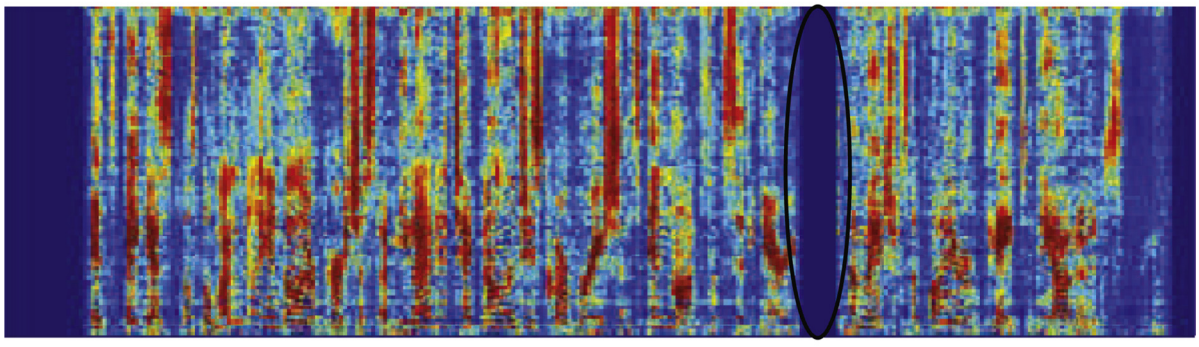
(a) The spectrogram of channel 0



(b) The spectrogram of channel 5



(c) The combination of CGMM-based mask and DNN-IRM



(d) +VAD information

Fig. 5. The comparison of mask estimation with/without VAD information for an utterance of the CHiME-4 RealData test set.

ter IME-based beamforming with VAD information, as shown in Fig. 5, which can significantly reduce insertion errors. On the other hands, for ASR systems, it is possible to misrecognize the speech segments as non-speech segments, thus yielding deletion errors. However, we observe that in most cases, these segmentations are with quite low SNRs, which appear as noise segmentations with very weak speech energy. Furthermore, because these segmentations are already misrecognized by the ASR system in the first-pass decoding, the recognition results on these segmentations would not be worse in the second-pass decoding.

Actually our proposed IME is one general framework to adopt both soft masks (CGMM-based and NN-based mask estimation between 0 and 1 in the T-F bin level) and hard masks (ASR-based VAD for the binary selection of speech and non-speech in the frame level). The motivation is for non-speech segmentations the hard/binary mask is exactly what we need to select noise frames while for speech segmentations the soft mask is more suitable due to the existence of both speech and noise in T-F bins.

Please note that for the mask combination, we empirically use different ways as in Eqs. (10), (12) and (14). For example, we use the arithmetic mean in Eq. (12) to combine the masks in the similar dynamic range from NNs with different architectures. However, to combine the CGMM-based and NN-based masks, the geometric mean is adopted as in Eq. (10) because NN-based mask values are often close to 0 at non-speech T-F bins while CGMM-based mask values are quite large at some non-speech T-F bins as shown in Fig. 4. Finally, to combine the ASR-based mask defined in Eq. (13), Eq. (14) is employed to perform as a hard selection mechanism.

5. Experimental evaluation

We now present the experimental evaluation of our framework in the CHiME-4 task (Vincent et al., 2016), which was designed to study real-world ASR scenarios where a person is talking to a mobile tablet device equipped with 6 microphones in a variety of adverse environments. Four conditions were selected: café (CAF), street junction (STR), public transport (BUS), and pedestrian area (PED). For each case, two types of noisy speech data were provided: RealData and SimData. RealData were collected from speakers reading the same sentences from the WSJ0 corpus (Garofalo et al., 2007) in the four conditions. SimData were constructed by mixing clean utterances with environmental noise recordings using the techniques described in Vincent et al. (2007). The SimData are used to train the DNN and LSTM for generating NN-based IRMs.

For the ASR evaluation, three data sets were designed, namely, the training set, development set and test set. The development and test data consist of 410 and 330 utterances, respectively, with the same sentence contents as the corresponding sets in the WSJ0 5k task, each read by four different speakers in one randomly selected condition. This resulted in 1640 (410 × 4) development and 1320 (330 × 4) test utterances of speech data in total. Similarly, simulated data were generated for the development and test sets. The training data include 1600 real noisy utterances from the combinations of four speakers each reading 100 utterances in four conditions (i.e., 4× 4 × 100) and 7138 simulated utterances from the WSJ0 training set. CHiME-4 offers three tasks (1-channel, 2-channel, and 6-channel) with different testing scenarios. In this paper, we focus only on the 6-channel case to make the paper concise, and more details about our back-end system, acoustic models and language model can be found in Tu et al. (2017). The readers are referred to Barker et al. (2017) and Vincent et al. (2016) for more detailed information regarding CHiME-4.

5.1. Experiments on CGMM-based beamformer

In this subsection, the baseline ASR system officially provided in Vincent et al. (2016) is used to evaluate the different beamformers on the test sets of real data. The acoustic model is a DNN-HMM discriminatively trained with the sMBR criterion (Vesely et al., 2013). The input of the DNN-HMM is a 440-dimensional feature vector extracted from channel 5, consisting of a 40-dimensional fMLLR (Gales, 1998) with an 11-frame expansion. The language models are 5-gram with Kneser-Ney (KN) smoothing (Kenser and Ney, 1995) for the first-pass decoding and the simple RNN-based language model (Mikolov et al., 2010) for rescoring.

For the CGMM-based beamforming, the multi-channel STFT coefficients are extracted from the test speech at a 16 kHz sampling frequency using a Hanning window of length 512 and shift of 256, resulting in 257 frequency bins. We studied the CHiME-4 baseline BeamformerIt for comparison in which the multichannel covariance matrix of noise is estimated from 400 ms to 800 ms of context immediately before the test utterance, and then the speech signal is estimated by time-varying minimum variance distortionless response (MVDR) beamforming with diagonal loading (Mestre and Lagunas, 2003).

Before moving to the next subsection, we would like to note that some microphones were found to be obstructed during the recording of the CHiME-4 data. Using signals from such microphones would degrade the beamforming performance. To circumvent this issue, the obstructed microphones are excluded from beamforming. The process to determine whether a microphone is obstructed is as follows. A set of cross-correlation functions (Vu et al., 2015) over the J microphone signals and the average cross-correlation are first computed. Then, a microphone is determined to be an obstructed microphone if its cross-correlation value (with other microphones) is lower than a threshold η. Empirically, the value of η is set to 0.2.

Table 1 presents the word error rate (WER) comparison of different beamformers averaged on the test sets of RealData. In this table, "CH5" denotes the recognition of original speech from channel 5. "IRM" is the single-channel speech enhancement approach using our DNN-IRM model with the channel 5 data as the input. "BeamformIt" is the CHiME-4 officially provided beamformer (Mestre and Lagunas, 2003). We list the result in Nugraha et al. (2016) as a reference by using DNN-based multi-channel speech enhancement approach mentioned above. "CGMM w/o CS" is our implemented version without channel selection mentioned above. "CGMM" is our implemented version with channel selection while Higuchi et al. (2016) denotes the published result of CGMM in the original paper with more powerful CNN-HMMs than CHiME-4 officially provided DNN-HMMs. Higuchi et al. (2016) does not give the corresponding performance using DNN-HMMs. "IME" is our proposed iterative mask estimation approach, and more detailed experimental results are shown in the following subsections. Please note that in Table 1, all systems use the same acoustic and language models except Nugraha et al. (2016) and Higuchi et al. (2016). Although Nugraha et al. (2016) uses different backend settings, CHiME-3 challenge results (Barker et al., 2017) confirmed that the CGMM approach outperformed all the other deep learning based multi-channel speech enhancement approaches in the submitted systems. So it is convincing to conclude that our proposed IME is the best beamformer among all the approaches in Table 1.

5.2. Experiments on IME-based beamformer

Next, we study our proposed iterative mask estimation approach for improving the CGMM-based method by leveraging NN-based IRM and ASR-based VAD.

5.2.1. NN-based IRM

We now investigate the impacts of both DNN-based and LSTM-based IRM estimation on the recognition performance. The DNN and LSTM architectures are shown in Fig. 3. The activation functions of the hidden and output layers are sigmoid units. For fine-tuning of the DNN, the learning rate is set to 0.01 for 50 epochs, and the mini-batch size is 128. For fine-tuning of the LSTM, the learning rate is set to 0.001 for 50

Table 1
WER (%) comparison of different beamformers averaged on the test sets of RealData.

	CH5	IRM	BeamformIt	Nugraha et al. (2016)	CGMM w/o CS	CGMM	Higuchi et al. (2016)	IME
AVG	23.47	22.96	11.82	10.14	9.56	8.54	8.37	**5.96**

Table 2
WER (%) comparison among the CGMM-based beamforming and the improved versions by incorporating the NN-based IRM on the test sets of RealData.

Test Data	BUS	CAF	PED	STR	AVG
CGMM	13.24	8.12	6.67	6.03	8.54
DNN-1	9.32	5.96	5.83	5.93	6.76
DNN-2	9.52	5.86	5.92	5.72	6.75
LSTM-1	9.58	5.87	5.96	5.54	6.73
LSTM-2	9.52	5.92	5.89	5.64	6.74
NN-based Ensemble	9.42	5.76	5.78	5.45	**6.60**

Table 3
WER (%) comparison among different approaches to combine the mask estimations from the CGMM and NN-based IRM on the test sets of RealData.

Test Data	BUS	CAF	PED	STR	AVG
DNN-CGMM	14.02	8.96	7.13	6.43	9.13
+DNN-1	10.03	6.22	6.14	5.96	7.09

epochs, and the mini-batch size is set to 128. The input features for both are globally normalized to zero mean and unit variance. In the training stage, only the simulation data are adopted with the input/output pairs of channel 5 speech and the corresponding IRMs.

Table 2 lists the WER comparison among the CGMM-based beamformer and its improved versions by incorporating the NN-based IRM on the test sets of RealData. “DNN-1” and “DNN-2” denote the iterative mask estimation in the first and second iterations, respectively. The results for the corresponding LSTM versions are denoted by “LSTM-1” and “LSTM-2”. Several observations can be made from the results. First, the DNN-based approach (DNN-1 and DNN-2) achieves consistent and significant improvements in recognition performance over the CGMM-based method, yielding an average relative WER reduction of 20.7% across all test sets for DNN-1. For BUS and CAF environments with lower WERs, the relative WER reductions are 27.3% and 26.4%, respectively, which demonstrates the effectiveness of NN-based mask estimation for ASR in adverse environments. Second, the performance is saturated after one iteration for both the DNN and LSTM (DNN-1 vs. DNN-2, LSTM-1 vs. LSTM-2). Finally, although the LSTM-based approach generates an average WER similar to that of the DNN-based approach, the ensemble of DNN-1 and LSTM-1 provides an additional 2% relative WER reduction. This result shows the complementarity of different architectures.

Table 3 lists the WER comparison among different approaches to combine the mask estimations from the CGMM and NN-based IRM on the test sets of RealData. “DNN-CGMM” denotes the CGMM-based beamforming with the initialized DNN-based IRM estimation from channel 5. “+DNN-1” denotes our IME approach using the DNN-based IRM based on the “DNN-CGMM” system in the first iteration. The recognition performance of “DNN-CGMM” shown in the first line of Table 3 is clearly worse than that of the conventional CGMM-based beamforming initialized as (Higuchi et al., 2016) shown in the first line of Table 2, which demonstrates that the DNN-IRM estimated from the original channel 5 is not accurate enough to improve the CGMM-based beamforming. Moreover, the significant improvements of “+DNN-1” over “DNN-CGMM” demonstrate the effectiveness of the proposed IME approach.

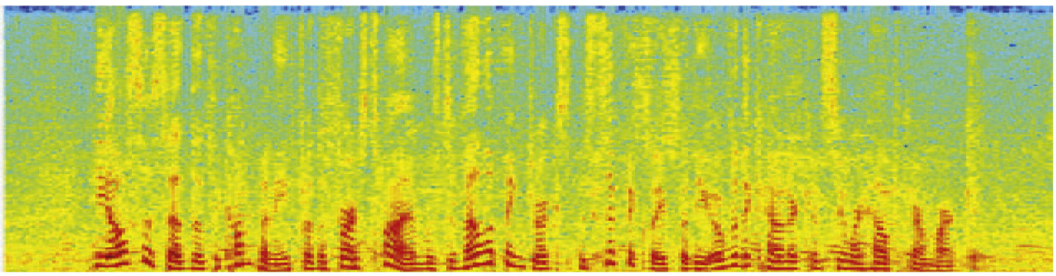
Table 4
WER (%) comparison of different beamformers by incorporating the ASR-based VAD on the test sets of RealData.

Test Data	BUS	CAF	PED	STR	AVG
CGMM	13.24	8.12	6.67	6.03	8.54
+VAD-1	10.94	8.29	6.11	5.42	7.69
+VAD-2	10.89	8.33	5.88	5.28	7.59
+Oracle VAD	10.66	8.19	5.72	5.20	7.44
DNN-1	9.63	5.98	5.85	5.62	6.77
+VAD-1	7.80	5.75	5.42	5.64	6.15
+VAD-2	7.72	5.68	5.39	5.56	6.08
NN-based Ensemble	9.42	5.76	5.78	5.45	6.60
+VAD-1	7.38	5.77	5.57	5.58	6.08
+VAD-2	7.37	5.83	5.33	5.32	**5.96**
CH0	5.07	5.02	5.27	6.48	5.46

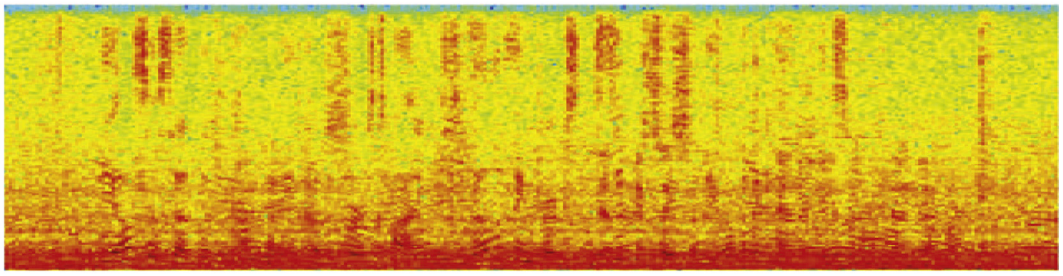
5.2.2. ASR-based VAD

In this subsection, the segmentation results of the speech recognizer based on the beamformed speech are used as the VAD information to improve the estimated mask. The recognizer is the same as in the previous experiment. Table 4 shows the WER comparison of different beamformers by incorporating the ASR-based VAD on the test sets of RealData. There are three blocks of results, with each consisting of three rows and denoting one system to incorporate the VAD information. The first block corresponds to the iterative mask estimation with the CGMM-based system via Eq. (14), while the second and third blocks represent the iterative mask estimation with the CGMM-based and NN-based systems via Eq. (15). For all three baseline systems (CGMM, DNN-1, and NN-based ensemble), the ASR-based VAD for iterative mask estimation yields significant and stable recognition performance gains across all test sets of RealData, with average relative WER reductions of 11.1%, 10.2%, and 9.7% after the second iterations (+VAD-2), which demonstrates the strong complementarity with both CGMM-based and NN-based mask estimation. One interesting observation is for the VAD-based iterative mask estimation: the second iteration (+VAD-2) can always slightly improve the recognition performance over the first iteration (+VAD-1), which is different from that of the NN-based mask estimation shown in Table 2. For example, an average relative WER reduction of 2% is achieved for the NN-based ensemble system. Overall, compared with the CGMM-based method (the first row of Table 4), our iterative mask estimation approach using both NN-based IRM and ASR-based VAD (the last row of Table 4) achieves an average relative WER reduction of 30.2%. Finally, “Oracle VAD” denotes that the ASR-based VAD information is obtained by force-alignment results of the channel 0 data. The results of “Oracle VAD” are slightly better than those of VAD-2, which demonstrates that the VAD information based on recognition results is sufficiently accurate for IME-based beamforming.

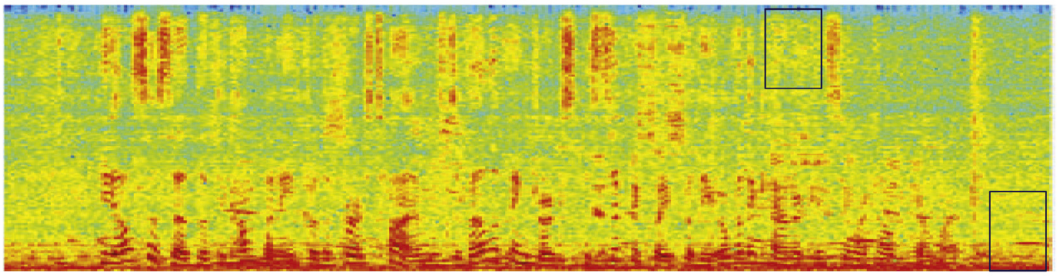
Fig. 6 plots the spectrograms with the recognition results of an utterance from the test set of RealData (F05_442C020Y_BUS_REAL) using different beamformers. Fig. 6(a) and (b) present the recognition of the original speech from channel 0 and channel 5, respectively. The recognition results of channel 0 as a close-talking microphone were 100% correct, whereas the recognition error rate of channel 5 is approximately 50% because of background BUS noises. The CGMM-based approach dramatically reduced the stationary noises, as shown by the spectrograms in Fig. 6(b) and (c). However, there are still three parts of substitution/insertion/deletion errors due to the existence of non-stationary and residual noises. With the NN-based IRM plus ASR-based VAD information [Fig. 6(d) and (e)], the recognition error is gradually reduced to 0.



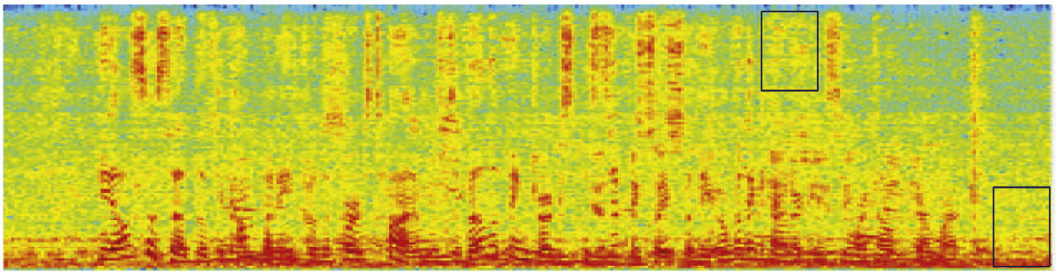
(a) CH0: HE ALSO SAID THAT THE COMPANY FOR THE FIRST TIME WAS DEVELOPING DRUGS SPECIFICALLY FOR THE OVER THE COUNTER CONSUMER HEALTH CARE MARKET



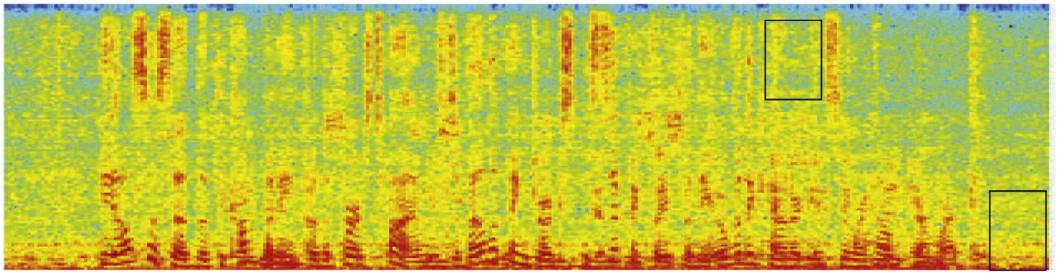
(b) CH5: HE ALSO SAID ~~THAT~~ THE COMPANY FOR THE FIRST TIME ~~WAS~~ IN DEVELOPING DRUGS ~~SPECIFICALLY FOR THE OVER THE COUNTER CONSUMER HEALTH CARE~~ BENEFITS A SURVEY OF ECONMISTS BY THE MARKET



(c) CH0: HE ALSO SAID THAT THE COMPANY FOR THE FIRST TIME ~~WAS~~ IN DEVELOPING DRUGS THE DIFFICULTIES OF ~~SPECIFICALLY FOR~~ THE OVER THE COUNTER CONSUMER HARDWARE ~~HEALTH CARE MARKET~~



(d) +IRM: HE ALSO SAID THAT THE COMPANY FOR THE FIRST TIME WAS DEVELOPING DRUGS THE DIFFICULTIES OF ~~SPECIFICALLY FOR~~ THE OVER THE COUNTER CONSUMER HEALTH CARE MARKET



(e) +VAD: HE ALSO SAID THAT THE COMPANY FOR THE FIRST TIME WAS DEVELOPING DRUGS SPECIFICALLY FOR THE OVER THE COUNTER CONSUMER HEALTH CARE MARKET

Fig. 6. Spectrograms with the recognition results of an utterance from the test set of RealData (F05_442C020Y_BUS_REAL) using different beamformers.

Table 5
WER (%) results of data augmentation for the DNN-HMM acoustic model on the test sets of RealData.

Training Data	Test Data	BUS	CAF	PED	STR	AVG
CH5	CGMM	13.24	8.12	6.67	6.03	8.54
	IME	7.37	5.83	5.33	5.32	5.96
CH13456	CGMM	8.39	6.56	5.36	5.30	6.40
	IME	5.76	4.46	3.81	4.80	4.71
CH13456	CGMM	9.18	5.77	4.35	4.58	5.97
+Beamformed	IME	5.12	3.90	3.50	4.61	**4.28**

Although the spectrograms in Fig. 6(c)–(e) are not "sounded" or "seen" quite different, the reduction of non-stationary and residual noises in the key regions plays an important role in the recognition process. Thus, although the spectrogram in Fig. 6(e) is not very "clean", the recognition result is correct due to the multi-condition training of the acoustic model.

5.3. Experiments on robustness of IME

In this section, we focus on the impacts of powerful acoustic and language models on the IME approach.

5.3.1. Training data augmentation

In contrast to the DNN-HMM in Sections 5.1 and 5.2 where only fMLLR features were used as input, the input feature vector for the data augmentation experiments consists of 42-dimensional LMFB, 40-dimensional fMLLR, and 20-dimensional i-vector (Tu et al., 2016). For both LMFB and fMLLR, the first-order and second-order derivatives with 9-frame expansion are adopted, yielding a 2234-dimensional ($2234 = 42 * 3 * 9 + 40 * 3 * 9 + 20$) feature vector fed to the input layer of the DNN. We use 7 hidden layers with 2048 nodes for each layer and 1965 nodes for the output layer. Other configurations follow the Kaldi setup officially provided in Vincent et al. (2016).

Table 5 lists the WER results of data augmentation for the DNN-HMM acoustic model on the test sets of RealData. "CH5" is the system using only channel-5 data for training which is the same setting as in Table 4. And the "IME" corresponds to the best setting in Table 4, namely "NN-based Ensemble + VAD-2". "CH13456" represents the data augmentation by using all channels of the original noisy speech except channel 2 while "CH13456 + Beamformed" includes additionally CGMM-based beamformed speech data. Several observations could be made. First, data augmentation can significantly improve the recognition performance when the training set is not large, e.g., a dozen hours of speech data. For CGMM-based beamformer, "CH13456" using 5-fold training data compared with "CH5" yields an average relative WER reduction of 25.1% while "CH13456 + Beamformed" with additional beamformed data achieves an average relative WER reduction of 6.7% over "CH13456". The complementarity between noisy speech and beamformed speech might be explained as that the beamforming can introduce some distortions although it can improve the SNR of speech signals. Second, for data augmentation, our proposed IME approach yields average relative WER reductions of 26.4% and 28.3% over the CGMM-based approach for "CH13456" and "CH13456 + Beamformed", respectively. These improvements are quite consistent with the average relative WER reduction of 30.2% in "CH5" system without data augmentation, which indicates that our proposed beamforming approach is still very effective when combined with enhanced acoustic modeling using data augmentation.

5.3.2. The ensemble of DNN-HMM and DCNN-HMMs

For DCNN-HMMs, four models are built with different settings of input features and kernel sizes, as shown in Table 6. The learning rate of DCNN training is set to 0.002, and the batch size is 2048. The model consists of the input layer, 4 blocks with different sizes of feature maps, one FC layer and the softmax output layer. For each block, there are four convolution layers, a ReLU layer, a batch normalization (BN) layer and a max-pooling layer. Table 7 shows the WER comparison with different acoustic models on the test sets of RealData.

Table 6
The settings of different DCNN-HMMs.

Acoustic Model	Input Feature	Kernel Size
DCNN1-HMM	LMFB	3 × 3
DCNN2-HMM	LMFB	3 × 5
DCNN3-HMM	fMLLR	3 × 3
DCNN4-HMM	fMLLR	3 × 5

Table 7
WER (%) comparison with different acoustic models (AMs) and language models (LMs) on the test sets of RealData.

Test	LM/AM	BUS	CAF	PED	STR	AVG
CGMM	RNN/DNN-HMM	9.18	5.77	4.35	4.58	5.97
	RNN/Ensemble	6.06	4.15	3.46	3.29	4.24
	LSTM/Ensemble	4.86	2.97	2.49	2.69	3.25
IME	RNN/DNN-HMM	5.12	3.90	3.50	4.61	4.28
	RNN/Ensemble	4.04	2.95	2.82	3.19	3.25
	LSTM/Ensemble	2.67	2.09	1.73	2.51	2.25

The complementarity of different input features and neural network architectures is well validated by the acoustic model fusion at the state level. With the best ensemble of one DNN-HMM and four DCNN-HMMs listed in Table 6 for acoustic modeling and RNN for language modeling, our proposed IME approach still achieved a significant WER reduction of 23.3% relative to the CGMM-based approach.

5.3.3. LSTM-based LMs

In this subsection, we examine the impact of LSTM-based language models. Table 7 presents the WER comparison between CGMM and IME with different language models on the test sets of RealData. Clearly, LSTM-based LMs generate much better results than the officially provided simple RNN-based LM. With the best configured ensemble acoustic model and LSTM-based language model, IME yields an average relative WER reduction of 30.8% over CGMM, which is a quite consistent improvement compared with previous experiments.

6. Conclusion

In this paper, we have proposed a simple and effective IME framework to precisely estimate the mask in an iterative manner from different pieces of complementary information sources with comprehensive and promising results on a state-of-the-art ASR challenge corpus. IME is a general framework to fully utilize the advantages of conventional beamforming (e.g., CGMM which can make use of the online spatial information), the purely deep learning based enhancement (e.g. DNN/LSTM IRM which can well learn the interactions between speech and noise from a large data set), and ASR feedbacks (e.g., VAD) to design a better beamformer. This is highly motivated from the experiences in previous CHiME challenges, namely both spatial beamforming (e.g., CGMM) and purely deep learning based multi-channel approach have obvious limitations. In the future, we can improve IME further by leveraging upon better spatial beamforming approaches, better deep learning algorithms for IRM estimation, and more informative feedback from the ASR systems. For example, in this study we only use DNN and unidirectional LSTM to estimate IRM as it is possible to develop the online version of IME (Tu et al., 2018). Without considering the latency and computational complexity, we will explore the BLSTM and CNN to further improve the performance. Furthermore, deep integration among different information sources will be investigated by designing the new objective functions for joint learning. And we will explore how to in-

corporate the CGMM-based approach into NN-training to simplify our system.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2017YFB1002202, in part by the National Natural Science Foundation of China under Grants 61671422 and U1613211, in part by the Key Science and Technology Project of Anhui Province under Grant 17030901005, and in part by Tencent.

References

Anguera, X., Wooters, C., Hernando, J., 2007. Acoustic beamforming for speaker diarization of meetings. IEEE Trans. Audio Speech Lang. Process. 15 (7), 2011–2022.
Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2017. The third chime speech separation and recognition challenge: analysis and outcomes. Comput. Speech Lang. 46, 605–626.
Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust., Speech, Signal Process. 27 (2), 113–120.
Buchner, H., Aichner, R., Kellermann, W., 2005. A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. IEEE Trans. Speech Audio Process. 13 (1), 120–134.
Capon, J., 1969. High-resolution frequency-wavenumber spectrum analysis. Proc. IEEE 57 (8), 1408–1418.
Cohen, I., Berdugo, B., 2001. Speech enhancement for non-stationary noise environments. Signal Process. 81 (11), 2403–2418.
Cox, H., Zeskind, R.M., Owen, M., 1987. Robust adaptive beamforming. IEEE Trans. Acoust., Speech, Signal Process. 35 (10), 1365–1376.
Du, J., Huo, Q., 2008. A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions.. In: Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH).
Du, J., Tu, Y.-H., Dai, L., Lee, C.-H., 2016. A regression approach to single-channel speech separation via high-resolution deep neural networks. IEEE Trans Audio Speech Lang Process 24 (8), 1424–1437.
Du, J., Tu, Y.-H., Sun, L., Ma, F., Wang, H.-K., Pan, J., Liu, C., Chen, J.-D., Lee, C.-H., 2016. The ustc-iflytek system for chime-4 challenge. In: Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Environments (CHiME 2016).
Du, J., Wang, Q., Tu, Y.-H., Bao, X., Dai, L.-R., Lee, C.-H., 2015. An information fusion approach to recognizing microphone array speech in the chime-3 challenge based on a deep learning framework. In: Proc. IEEE Automat. Speech Recognition and Understanding Workshop. (ASRU).
Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust., Speech Signal Process. 32 (6), 1109–1121.
Gales, M.J.F., 1998. Maximum likelihood linear transformations for hmm-based speech recognition. Comput. Speech Lang. Lang. 12 (2), 75–98.
Gao, T., Du, J., Dai, L.-R., Lee, C.-H., 2015. Joint training of front-end and back-end deep neural networks for robust speech recognition. In: Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP).
Gao, T., Du, J., Xu, Y., Liu, C., Dai, L.-R., Lee, C.-H., 2016. Joint training of dnns by incorporating an explicit dereverberation structure for distant speech recognition. EURASIP J. Adv. Signal Process. 2016 (1), 86.
Garofalo, J., Graff, D., Paul, D., Pallett, D., 2007. Csri (wsj0) complete. Linguistic Data Consortium, Philadelphia.
Gauvain, J.L., Lee, C.-H., 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Trans. Speech Audio Process. 2 (2), 291–298.
Heymann, J., Drude, L., Boeddeker, C., Hanebrink, P., Haeb-Umbach, R., 2017. Beamnet: end-to-end training of a beamformer-supported multi-channel asr system. In: Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP).
Heymann, J., Drude, L., Chinaev, A., Haeb-Umbach, R., 2015. Blstm supported gev beamformer front-end for the 3rd chime challenge. In: Proc. IEEE Automat. Speech Recognition and Understanding Workshop. (ASRU).
Higuchi, T., Ito, N., Yoshioka, T., Nakatani, T., 2016. Robust mvdr beamforming using time-frequency masks for online/offline asr in noise. In: IEEE International Conference on Acoustics, Speech and Signal Processing.
Hinton, G.E., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A.W., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al., 2012. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag. 29 (6), 82.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Hoshuyama, O., Sugiyama, A., Hirano, A., 1999. A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. IEEE Trans. Signal Process. 47 (10), 2677–2684.
Hummersone, C., Stokes, T., Brookes, T., 2014. On the ideal ratio mask as the goal of computational auditory scene analysis. Blind Source Separat. 349–368.
Jones, D.L., Ratnam, R., 2009. Blind location and separation of callers in a natural chorus using a microphone array. J. Acoust. Soc. Am. 126 (2), 895–910.
Jutten, C., Herault, J., 1991. Blind separation of sources, part i: an adaptive algorithm based on neuromimetic architecture. Signal Process. 24 (1), 1–10.
Kenser, R., Ney, H., 1995. Improved backing-off for m-gram language modeling. In: Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP), Vol. 1. Detroit, MI, USA.
Keyi, A.E., Kirubarajan, T., Gershman, A., 2005. Robust adaptive beamforming based on the kalman filter. IEEE Trans. Signal Process. 53 (8), 3032–3041.
Krueger, A., Warsitz, E., Haebumbach, R., 2011. Speech enhancement with a gsc-like structure employing eigenvector-based transfer function ratios estimation. IEEE Trans. Audio Speech Lang. Process. 19 (1), 206–219.
Lim, J.S., Oppenheim, A.V., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67 (12), 1586–1604.
Mcaulay, R.J., Quatieri, T.F., 1986. Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. Acoust., Speech, Signal Process. 34 (4), 744–754.
Menne, T., Heymann, J., Alexandridis, A., Irie, K., Zeyer, A., Kitza, M., Golik, P., Kulikov, I., Drude, L., Schlter, R., Ney, H., Haeb-Umbach, R., Mouchtaris, A., 2016. The rwth/upb/forth system combination for the 4th chime challenge evaluation. In: Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Environments (CHiME 2016).
Mestre, X., Lagunas, M.A., 2003. On diagonal loading for minimum variance beamformers. In: Proc. IEEE ISSPIT, pp. 459–462.
Meyer, J., Simmer, K.U., 1997. Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction. In: Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on, Vol. 2. IEEE, pp. 1167–1170.
Mikolov, T., Karafiat, M., Burget, L., 2010. Recurrent neural network based language model. In: Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH). Chiba, Japan
Mohamed, A., Dahl, G.E., Hinton, G.E., 2012. Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. 20 (1), 14–22.
Nakatani, T., Ito, N., Higuchi, T., Araki, S., Kinoshita, K., 2017. Integrating dnn-based and spatial clustering-based mask estimation for robust mvdr beamforming, 286–290.
Nugraha, A.A., Liutkus, A., Vincent, E., 2016. Multichannel audio source separation with deep neural networks. IEEE Trans. Audio Speech Lang. Process. 24 (9), 1652–1664.
Ochiai, T., Watanabe, S., Hori, T., Hershey, J.R., Xiao, X., 2017. Unified architecture for multichannel end-to-end speech recognition with neural beamforming. IEEE J. Sel. Top Signal Process. 11 (8), 1274–1288. doi:10.1109/JSTSP.2017.2764276.
Sainath, T.N., Weiss, R.J., Wilson, K.W., Li, B., Narayanan, A., Variani, E., Bacchiani, M., Shafran, I., Senior, A., Chin, K., Misra, A., Kim, C., 2017. Multichannel signal processing with deep neural networks for automatic speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (5), 965–979. doi:10.1109/TASLP.2017.2672401.
Sohn, J., Kim, N.S., Sung, W., 1999. A statistical model-based voice activity detection. IEEE Signal Process. Lett. 6 (1), 1–3.
Souden, M., Benesty, J., Affes, S., 2010. A study of the lcmv and mvdr noise reduction filters. IEEE Trans. Signal Process. 58 (9), 4925–4935.
Spriet, A., Moonen, M., Wouters, J., 2004. Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction. Signal Processing 84 (12), 2367–2387.
Talmon, R., Cohen, I., Gannot, S., 2009. Convolutive transfer function generalized sidelobe canceler. IEEE Trans. Audio Speech Lang. Process. 17 (7), 1420–1434.
Tu, Y.-H., Du, J., Dai, L.-R., Lee, C.-H., 2015. Speech separation based on signal-noise-dependent deep neural networks for robust speech recognition. In: Proc. IEEE Int'l Conf. Acoust. Speech Signal Process.(ICASSP).
Tu, Y.-H., Du, J., Sun, L., Ma, F., Lee, C.-H., 2017. On design of robust deep models for chime-4 multi-channel speech recognition with multiple configurations of array microphones. In: Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH), pp. 394–398.
Tu, Y.-H., Du, J., Wang, Q., Bao, X., Dai, L.-R., Lee, C.-H., 2016. An information fusion framework with multi-channel feature concatenation and multi-perspective system combination for the deep-learning-based robust recognition of microphone array speech. Comput. Speech Lang. doi:10.1016/j.csl.2016.12.004.
Tu, Y.-H., Du, J., Zhou, N., Lee, C.-H., 2018. Online lstm-based iterative mask estimation for multi-channel speech enhancement and asr. APSIPA.
Van Veen, B.D., Buckley, K.M., 1988. Beamforming: a versatile approach to spatial filtering. IEEE assp magazine 5 (2), 4–24.
Veen, B.D., Buckley, K.M., 1988. Beamforming: a versatile approach to spatial filtering. IEEE Signal Process. Mag. 10 (3), 4–24.
Vesely, K., Ghoshal, A., Burget, L., Povey, D., 2013. Sequence-discriminative training of deep neural networks. In: Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH), pp. 2345–2349.
Vincent, E., Watanabe, S., Nugraha, A.A., Barker, J., Marxer, R., 2007. Oracle estimators for the benchmarking of source separation algorithms. Signal Process. 87 (8), 1933—1950.
Vincent, E., Watanabe, S., Nugraha, A.A., Barker, J., Marxer, R., 2016. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. Comput. Speech Lang.
Vu, T.T., Bigot, B., Chng, E.S., 2015. Speech enhancement using beamforming and non negative matrix factorization for robust speech recognition in the chime-3 challenge. In: Proc. IEEE Automat. Speech Recognition and Understanding Workshop. (ASRU).
Wan, E.A., Nelson, A.T., 1998. Networks for speech enhancement. Handbook of Neural Networks for Speech Processing.
Wang, L., Ding, H., Yin, F., 2011. A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures. IEEE Trans. Audio Speech Lang. Process. 19 (3), 549–557.
Wang, Y., Wang, D., 2013. Towards scaling up classification-based speech separation. IEEE Trans. Audio Speech Lang. Process. 21 (7), 1381–1390.
Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Roux, J.L., Hershey, J.R., Schuller, B., 2015. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In: Latent Variable Analysis and Signal Separation, pp. 91–99.

Xiao, X., Zhao, S., Jones, D.L., Chng, E.S., Li, H., 2017. On time-frequency mask estimation for mvdr beamforming with application in robust speech recognition. In: Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP).

Xie, F., Van Compernolle, D., 1994. A family of mlp based nonlinear spectral estimators for noise reduction. In: Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP).

Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Yu, M.F.C., Fabian, W.J., Espi, M., Higuchi, T., Araki, S., Nakatani, T., 2015. The ntt chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices. In: Proc. IEEE Automat. Speech Recognition and Understanding Workshop. (ASRU).

Zhang, C., Florencio, D., Ba, D.E., Zhang, Z., 2008. Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. IEEE Trans. Signal Process. 10 (3), 538–548.

Zhao, S., Man, Z., Jones, D.L., Khoo, S., 2014. Robust adaptive beamforming based on the kalman filter. J. Acoust. Soc. Am. 136 (3).

Zhao, S., Xiao, X., Zhang, Z., Nguyen, T.N.T., Zhong, X., Ren, B., Wang, L., Jones, D.L., Chng, E.S., Li, H., 2015. Robust speech recognition using beamforming with adaptive microphone gains and multichannel noise reduction. In: Proc. IEEE Automat. Speech Recognition and Understanding Workshop. (ASRU).