Speech Communication 95 (2017) 28–39

Contents lists available at ScienceDirect

Speech Communication

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

A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments

Tian Gao[a], Jun Du[*,a], Li-Rong Dai[a], Chin-Hui Lee[b]

[a] National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, Anhui, China
[b] Georgia Institute of Technology, Atlanta, Georgia, United States

ARTICLE INFO

Keywords:
Speaker-dependent speech processing
Speech enhancement
Speech separation
Deep neural network
Low SNR

ABSTRACT

We propose a unified speech enhancement framework to jointly handle both background noise and interfering speech in a speaker-dependent scenario based on deep neural networks (DNNs). We first explore speaker-dependent speech enhancement that can significantly improve system performance over speaker-independent systems. Next, we consider interfering speech as one noise type, thus a speaker-dependent DNN system can be adopted for both speech enhancement and separation. Experimental results demonstrate that the proposed unified system can achieve comparable performances to specific systems where only noise or speech interference is present. Furthermore, much better results can be obtained over individual enhancement or separation systems in mixed background noise and interfering speech scenarios. The training data for the two specific tasks are also found to be complementary. Finally, an ensemble learning-based framework is employed to further improve the system performance in low signal-to-noise ratio (SNR) environments. A voice activity detection (VAD) DNN and an ideal ratio mask (IRM) DNN are investigated to provide prior information to integrate two sub-modules at frame level and time-frequency level, respectively. The results demonstrate the effectiveness of the ensemble method in low SNR environments.

1. Introduction

Speech enhancement (Benesty et al., 2005) and speech separation (Wang and Brown, 1999a) are important front-ends of speech processing systems aimed at noise reduction and segregating speech from mixed speakers, respectively. Background noise and human voice interference can reduce both the quality and intelligibility of the speech signals and cause performance degradations in real-world applications, including speech communication, hearing aids and speech and speaker recognition. A key goal of speech enhancement (Loizou, 2013) is to improve quality and intelligibility in the presence of interfering noise. On the other hand, speech separation (Wang and Brown, 1999a; Roweis, 2000) aims to separate the voice of a target speaker when multiple speakers talk simultaneously.

Numerous methods were developed over the past several decades for speech enhancement and speech separation. For enhancement, the conventional methods include a wide range of approaches, such as spectral subtraction (Boll, 1979), Wiener filtering (Lim and Oppenheim, 1978) and statistical-model-based algorithms (McAulay and Malpass, 1980). Spectral subtraction is one of the first algorithms proposed for noise reduction. However, the resulting enhanced speech often suffers from an annoying artifact called musical noise (Kamath and Loizou, 2002). The Wiener algorithm, minimum mean squared error (MMSE) estimation (Ephraim and Malah, 1984, 1985) and optimally modified log-spectral amplitude (OM-LSA) speech estimator (Cohen and Berdugo, 2001) all exist in a statistical estimation framework that attempts to find a linear (or nonlinear) estimator of the parameters of interest. OM-LSA utilizes a minima controlled recursive averaging (MCRA) noise estimation approach to avoid the musical residual noise phenomena. One limitation of the conventional speech enhancement algorithms is that they can't improve speech intelligibility effectively. Supervised and unsupervised nonnegative matrix factorization (NMF) methods were investigated in Mohammadiha et al. (2013) and Fan et al. (2014). The basic idea is to decompose the training data into bases and weight matrices for speech and noise, respectively.

For separation, one broad class is the so-called computational auditory scene analysis (CASA) (Wang and Brown, 2006), usually in an unsupervised mode. CASA-based approaches (Wang and Brown, 1999b; Wu et al., 2003; Shao and Wang, 2006; Hu and Wang, 2010, 2013), use the psychoacoustic cues such as pitch, onset/offset, temporal continuity, harmonic structure and modulation correlation, and segregate a

* Corresponding author.
E-mail addresses: gtian09@mail.ustc.edu.cn (T. Gao), jundu@ustc.edu.cn (J. Du), lrdai@ustc.edu.cn (L.-R. Dai), chl@ece.gatech.edu (C.-H. Lee).

http://dx.doi.org/10.1016/j.specom.2017.10.003
Received 30 December 2016; Received in revised form 14 October 2017; Accepted 16 October 2017
Available online 18 October 2017

voice of interest by masking the interfering sources. For example, in Hu and Wang (2010), pitch and amplitude modulation are adopted to separate the voiced portions of cochannel speech. In Hu and Wang (2013), unsupervised clustering is used to separate speech regions into two speaker groups by maximizing the ratio of between-cluster distance and within-cluster distance. In the supervised approaches, speech separation is often formulated as an estimation problem based on

$$\boldsymbol{x}^m = \boldsymbol{x}^t + \boldsymbol{x}^i \tag{1}$$

where $\boldsymbol{x}^m$, $\boldsymbol{x}^t$, $\boldsymbol{x}^i$ are speech signals of the mixture, target speaker, and interfering speaker, respectively. To solve this under-determined equation, a general strategy is to represent the speakers by two models and use a certain criterion to reconstruct the sources given the signal mixtures. The training data can be modeled using probabilistic models, such as a Gaussian mixture model (GMM) (Kristjansson et al., 2004), hidden Markov model (HMM) or factorial HMM (Roweis, 2000) and NMF-based model (Schmidt and Olsson, 2006).

Recently, deep learning techniques became increasingly popular in many speech research areas, e.g., speech recognition (Dahl et al., 2012; Hinton et al., 2012), speech enhancement and speech separation. For enhancement, some data-driven methods attempt to make a binary or soft classification decision on time-frequency units, such as estimating the ideal binary mask (IBM) or smoothed ideal ratio mask (IRM) for monaural speech denoising (Wang and Wang, 2013; Wang et al., 2014). The hard targets IBM is effective to improve speech intelligibility, but predicting the soft targets IRM is especially beneficial for improving objective speech quality. IRM is in the range of [0,1], which can be seen as a suppression gain at each time-frequency unit. The final enhanced features are obtained as the element-wise product of estimated IRM and noisy features. The soft masking algorithms suppress noise in some degree, so the speech distortion will be less accordingly. In addition to the direct prediction of IRM, Huang et al. (2014, 2015) investigated joint optimization of masking functions and DNNs with an extra masking layer.

Apart from the prediction of time-frequency masks, deep learning approaches are also adopted to mapping speech spectral directly. Xu et al. (2014b, 2015a) proposed a DNN-based regression framework via training a deep and wide neural network architecture using a large collection of heterogeneous training data. Using the DNN-based regression approach has the advantage that it makes no assumptions about the statistical properties of the signals, and it can also handle non-linear and highly non-stationary noises effectively. One challenge is that some distortions are introduced to the estimated clean speech signal because the regression DNN removes the noise considerably from the noisy speech. To address this challenge, the regression DNN was further post-processed with variance equalization of features to alleviate the distortions in the estimated clean features (Xu et al., 2014c).

One key point of deep learning approaches is the generalization capacity to unseen noise conditions. To improve the generalization capability, dynamic noise aware training approach was used in Xu et al. (2014a), and DNN architecture was extended to a multi-objective framework in Xu et al. (2015b). Like the adaptation methods used in speech recognition, Kim *et, al.* aimed at a fine-tuning scheme at the test stage to improve the performance of a well-trained Denoising Auto-Encoder (DAE) (Kim and Smaragdis, 2015). Another challenge is the performance degradation in low SNR environments. In Gao et al. (2015b), a joint DNN framework combining speech enhancement with voice activity detection (VAD) was proposed to increase the speech intelligibility in serious noise situations.

More complex neural network structure is also a research point for speech enhancement. Weninger et al. explored long short-term memory (LSTM) network (Weninger et al., 2015). Convolutional neural network (CNN) was investigated in Fu et al. (2016).

For speech separation, Du et al. (2014) employed a regression DNN to solve the separation problem in Eq. (1). Tu et al. (2014) modified the architecture with dual outputs for learning both target source spectral and interfering source spectral. A semi-supervised mode to separate speech of the target speaker from an unknown interfering speaker was discussed in Du et al. (2016), where the performance in the semi-supervised mode can even surpass that of the GMM approach in the supervised mode. Another related work is the generative stochastic network (GSN) based method using a hybrid generative-discriminative training objective (Zöhrer and Pernkopf, 2014, 2015).

Background noise and interfering speech are the two interferences in real world. They usually appear together in the following scenarios: living room, restaurant, mess hall, cocktail party, etc. When speech products applied to such scenarios, both background noise and speech interference need to be removed. Based on the above review of the conventional methods for speech enhancement and separation, we find that the two tasks are often treated separately. Roweis (2003) investigated factorial HMM and refiltering to address noise and human interference. However, the interfering sources are needed in the testing stage, and its computational complexity is too involved.

From the perspective of model learning, we can see speech enhancement and separation as a task of dissociation aimed at removing different interference. Generally, it is hard to use one model to handle both noise reduction and speech separation because the existence of speech interference will influence the learning of the target speech signal. However, if the target speech signal to be separated is from a specific speaker, the speech interference is not as serious a problem. The speaker-dependent system is meaningful because personalized services are needed and feasible today. We also find speaker-dependent systems can significantly improve the system performances when compared with speaker-independent systems, especially in low signal-to-noise ratio (SNR) environments.

In this paper, we train a speaker-dependent DNN system to unify speech enhancement and speech separation. The signals of speech interference are considered as one noise type. The target clean speech is only interfered by a single speaker in the experiment. Experimental results demonstrate that the unified DNN can achieve comparable performance to each specific system when only noise or speech interference is present. Furthermore, we show that much better results can obtained in mixing noise and interfering speech scenarios than from individual enhancement and separation systems. The training data for the two specific tasks are also found to be complementary. The data corrupted by background noise become helpful for speech separation as the SNR increases, while the speech interference is useful to reduce speech-like noise.

One challenge of speech enhancement is the performance degradation in low SNR environments. Fig. 1 shows an utterance mixed with babble noise at 0 dB along with a corresponding clean speech and frame-level SNR sequence. Speech presence segments corrupted by high-energy noise, such as within the two red dashed rectangular boxes in Fig. 1(b), remain difficult to handle. Even when noise is reduced from those segments by DNN approaches, the speech quality is still severely degraded as it is not easy for a DNN to distinguish in those segments between speech and noise. They are very similar to pure noise segments in terms of frame-level SNR. In this study, we use two specific DNNs to address speech presence segments and speech absence segments separately. The DNN for speech is trained with only the speech presence segments of the training data, and it can preserve the speech quality for low frame-level SNR segments. Another DNN can remove pure noise segments. Finally, a joint DNN framework is employed to integrate the two DNNs. This method can be seen as an implementation of ensemble learning, which integrates multiple weak learners to create a stronger one. Zhang and Wang (2016b), proposed a deep ensemble network for monaural speech separation. They employ multi-context networks to integrate temporal information at different resolutions. Multiple modules are stacked to construct an ensemble, each performing multi-context masking or mapping. Differing from the multi-context networks, in this work, we focus on the composition of the training data. A voice

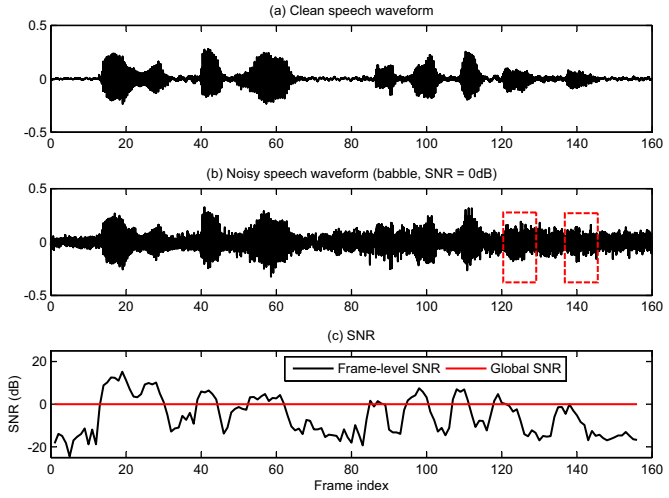

Fig. 1. Illustration of an utterance example in a babble noise environment at 0 dB along with corresponding clean speech and a frame-level SNR sequence.

activity detection (VAD) DNN and an IRM DNN are investigated to integrate two speech enhancement DNNs to construct an ensemble. The experimental results demonstrate the effectiveness of the ensemble method in low SNR environments.

When compared with our earlier conference paper (Gao et al., 2015a), the contributions of this paper are as follows: (1) The advantages of a speaker-dependent system is compared with speaker-independent systems when dealing with isolated noises. (2) A comprehensive set of experimental results showing noise reduction in general noise and speech-like noise scenarios, speech separation with only speech interference, and noise mixed with speech interference is presented. We discussed the complementarity between background noise and speech interference, which is helpful in reducing mixed noise. (3) A joint DNN framework proposed in Gao et al. (2015b) is employed to further improve the performance in serious noise environments. DNN-based IRM is first introduced into the framework at time-frequency unit level.

The remainder of the paper is organized as follows. In Section 2, we first give an overview of our proposed framework. In Section 3, DNN training for spectral mapping DNN, VAD DNN and IRM DNN are described in detail. In Section 4, the enhancement stage and DNN integration are presented. In Section 5, we report experimental results and analysis. Finally, we summarize our findings in Section 6.

2. Framework overview

The overall flowchart of the proposed speaker-dependent speech enhancement framework is illustrated in Fig. 2. In the training stage, we first use clean speech, noise data and speech interference to generate a large amount of noisy speech data. Then log-power spectral (LPS) features (Du and Huo, 2008) of both clean speech and synthesized noisy speech are extracted. First, a short-time Fourier analysis was applied to the input signal by computing the discrete Fourier transform (DFT) of each overlapping windowed frame:

$$y^f(d) = \sum_{l=0}^{L-1} y^t(l) w(l) e^{-2j\pi dl/L} \quad l = 0, 1, ..., L-1 \tag{2}$$

where d is the frequency bin index, $y^t(l)$ is the input signal at time domain, $w(l)$ denotes the window function (Hamming window here). Then LPS features are defined as

$$y(d) = \ln(|y^f(d)|^2) \quad d = 0, 1, ..., D-1 \tag{3}$$

where $D = L/2 - 1$. After feature extraction, four DNNs are trained including USE DNN, C-USE DNN, VAD DNN and IRM DNN explained as follows: (i) USE is a unified speaker-dependent DNN for speech enhancement and separation. (ii) C-USE is a conservative USE DNN trained with speech separation training data and the speech presence segments of speech enhancement training data to preserve the weak-energy speech presence segments in low SNR environments and conservatively remove the pure noise segments. (iii) VAD DNN is a binary classification DNN to detect speech presence at frame level. (iv) IRM DNN is used to predict ideal ratio mask at time-frequency unit level. In the enhancement stage, after feature extraction of the noisy utterances, the features are presented to the IRM or VAD DNN, USE DNN and C-USE DNN simultaneously. DNN integration is performed with speech presence probability estimated by IRM DNN or VAD DNN to obtain the final enhanced features as shown in Fig. 2. The additional phase information is calculated from the original noisy speech. Finally, an overlap-add method is used to reconstruct the waveform of enhanced speech. A detailed description of waveform reconstruction module can be found in Du and Huo (2008).

3. DNN training

3.1. Training for spectral mapping DNN

In Xu et al. (2014b), a DNN was adopted as a regression model to predict clean LPS features given the input noisy LPS features with acoustic context for speech enhancement. In Tu et al. (2014), the architecture was modified with dual outputs for learning both the target source and interfering source spectral for speech separation. In this paper, we improve the speech enhancement DNN to predict clean LPS and interference LPS features simultaneously in the output layer as shown in Fig. 3. The estimation of interference LPS can be considered as a regularization term, which leads to a better generalization capacity for estimating target speech. For DNN training, back-propagation with an MMSE-based object function of the differences between the LPS features of the estimated and reference LPS features is adopted to train the DNN. Another two techniques, namely dropout training and noise-aware training (NAT) (Xu et al., 2015a) are implemented to improve generalization capability. Dropout randomly omits a certain percentage (dropout rate is 0.1 in this work) of the neurons in the input and each hidden layer, which can be treated as model averaging to avoid the over-fitting problem. NAT is adopted to improve the generalization capability of the DNN to unseen noise conditions. The input of DNN is augmented with an estimate of noise. The noise estimate is obtained by averaging first T ($T = 6$) frames LPS feature of an utterance. In this paper, restricted Boltzmann machine (RBM) (Hinton et al., 2006)-based pre-training is not used for regression DNN training because the gains are small (Xu et al., 2015a).

A stochastic gradient descent algorithm is performed in a mini-batch mode with multiple epochs to improve learning convergence as follows:

$$Er = \frac{1}{N}\sum_{n=1}^{N} \{\beta \|\hat{\mathbf{X}}_n^t - \mathbf{X}_n^t\|_2^2 + (1-\beta)\|\hat{\mathbf{X}}_n^i - \mathbf{X}_n^i\|_2^2\} \tag{4}$$

where $\hat{\mathbf{X}}_n^t$ and $\mathbf{X}_n^t$ are the nth D-dimensional vectors of estimated and reference clean features of the target speaker, respectively. In the same way, $\hat{\mathbf{X}}_n^i$ and $\mathbf{X}_n^i$ are vectors of the estimated and reference interference features. For unified system, the interference contains both background noise and speech interference. β is used to tune the contribution from the target part and the interference part. Another benefit of the dual output DNN is that estimation of interference can be used by a post-processing module to be discussed in Section 4.3.

3.2. Training for voice activity detection (VAD) DNN

DNN for VAD is designed as a classification model where the output

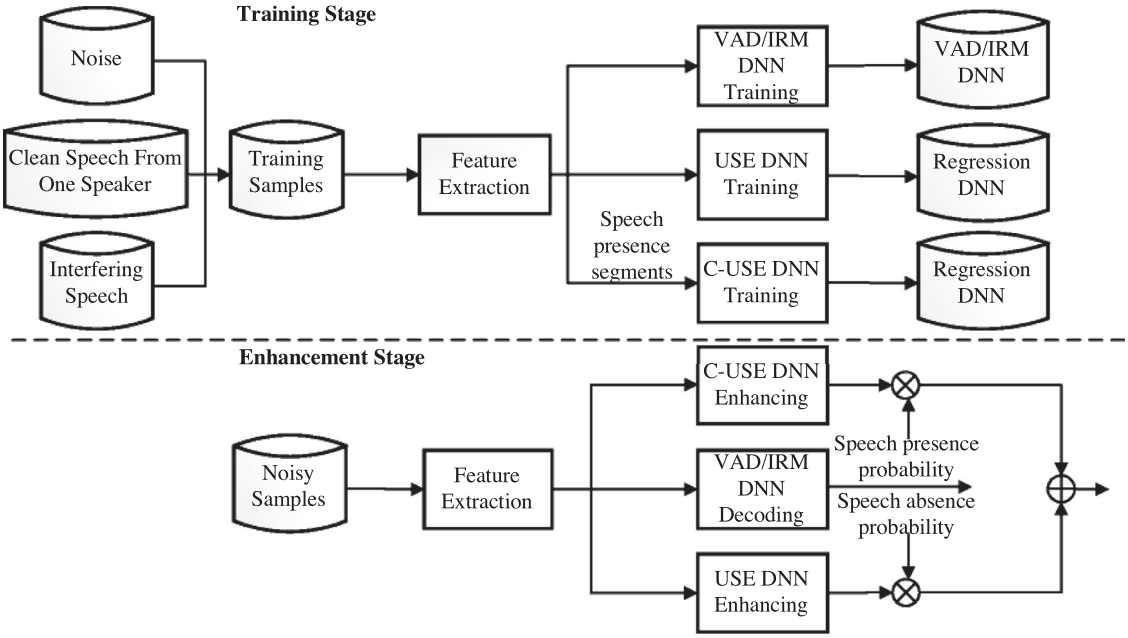

Fig. 2. Overall development flow and architecture.

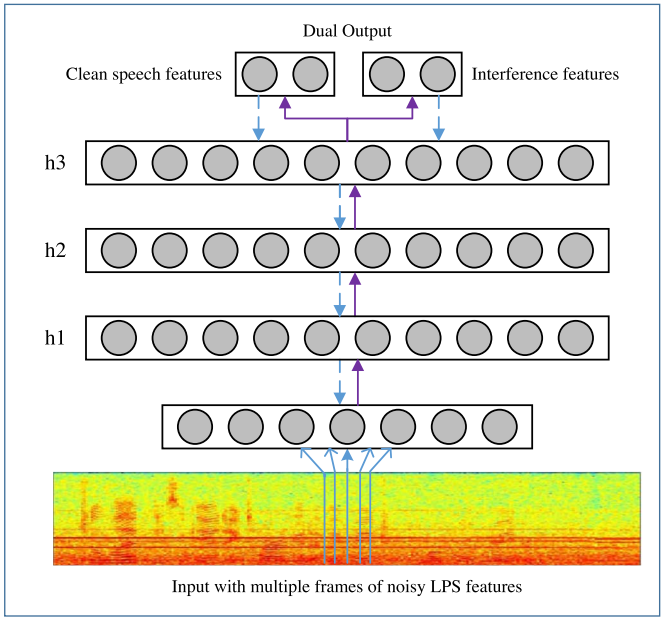

Fig. 3. DNN-based spectral mapping with dual output.

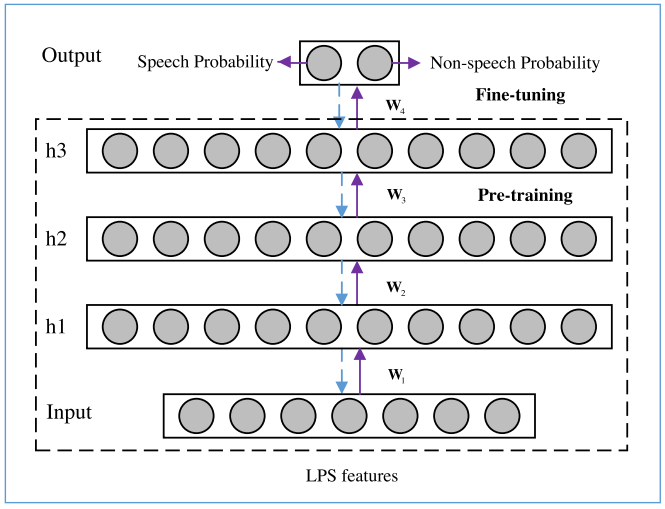

Fig. 4. DNN-based VAD.

refers to the probabilities of speech presence and absence classes (Zhang and Wu, 2013; Wang et al., 2015) as shown in Fig. 4. The inputs to DNN are the noisy LPS features from the current and neighboring frames. The learning targets are obtained as follows. We first apply a classical energy-based VAD algorithm provided by kaldi toolkit (Povey et al., 2011) on clean speech to detect speech presence segments. Then, the label information of clean speech are processed as the learning targets of corresponding noisy LPS features. Refer to the training procedure in Zhang and Wu (2013), training of this classification DNN consists of unsupervised pre-training and supervised fine-tuning. The former treats each consecutive pair of layers as a RBM while the parameters of the RBM are trained layer-by-layer with the approximate contrastive divergence algorithm (Hinton, 2002). After pre-training to initialize the weights of the first several layers, supervised fine-tuning of the parameters in the whole network is performed using a frame-level cross-entropy criterion:

$$C = -\sum_{j=1}^{Q} q_j \ln p_j \tag{5}$$

where C denotes cross-entropy cost function, p_j is the output of the softmax, q_j is the corresponding target, Q is the number of classes. In this paper, $Q = 2$.

3.3. Training for ideal ratio mask (IRM) DNN

The ideal ratio mask (IRM) (Wang et al., 2014) used as DNN learning target is defined as follow,

$$IRM_n^l(d) = \sqrt{\frac{\exp(X_n^t(d))}{\exp(X_n^t(d)) + \exp(X_n^i(d))}} \tag{6}$$

where $X_n^t(d)$ and $X_n^i(d)$ denote LPS features of target and interference, respectively. The $\exp(\cdot)$ operation restores the feature to denote energy. The notation l is used to make difference with calculated IRM in Section 4.3. DNN for IRM is designed as a regression model where the output can be considered as the probabilities of speech presence at each

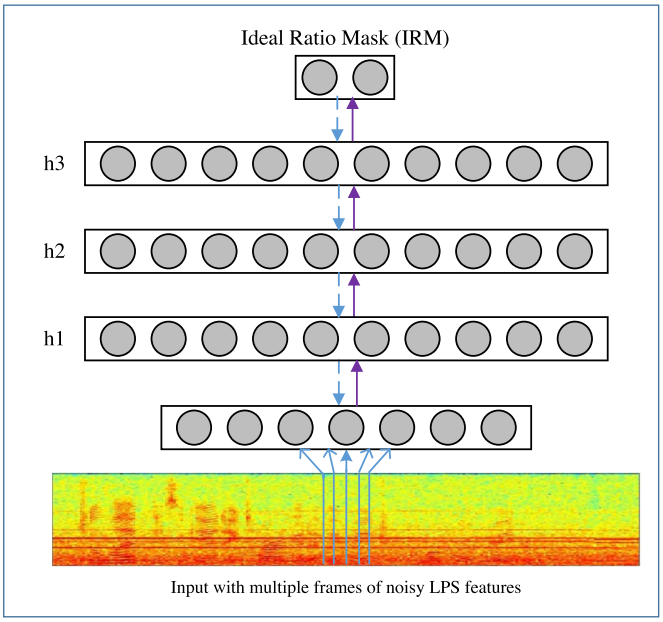

Fig. 5. DNN architecture for IRM prediction.

time-frequency unit as shown in Fig. 5. The inputs to DNN are the noisy LPS features from the current and neighboring frames. Dropout and NAT are also implemented for the IRM DNN, and the configurations are the same with spectral mapping DNN. The learning targets are IRM defined in Eq. (6). Then, supervised fine-tuning of the parameters in the whole network is performed using MMSE criterion,

$$Er_{irm} = \frac{1}{N}\sum_{n=1}^{N} \|\widehat{\mathbf{IRM}}_n^l - \mathbf{IRM}_n^l\|_2^2 \tag{7}$$

where $\widehat{\mathbf{IRM}}_n^l$ and $\mathbf{IRM}_n^l$ are the n^{th} D-dimensional vectors of estimated and reference IRM, respectively.

3.4. DNN configuration

As described in Section 2, there are four DNNs (USE, C-USE, VAD DNN and IRM DNN) needed. At first, USE is trained from a collection of stereo data, consisting of pairs of clean speech and noisy speech represented by their corresponding LPS features. The DNN architecture is 2056-2048-2048-2048-514, which denotes that the size is 2056 (257*7 + 257, including 3 left and 3 right context frames, and 1 frame for noise aware training (Xu et al., 2015a)) at the input layer, 2048 units for each of the three hidden layers, and 514 for the output layer (dual outputs). As the noise LPS variances are large and not stable, we mainly focus on the speech part. The regularization weighting coefficient β in Eq. (4) is set to 0.8. For fine-tuning, the learning rate is set at 0.1 for the first 10 epochs, and then decreased by 10% after every epoch. The total number of epochs is 30. The mini-batch size is set to $N = 128$. All DNN input features are normalized to zero mean and unit variance; the sigmoid activation functions are used in all hidden layers and the linear activation function is used in the output layer.

USE is a spectral mapping DNN which can remove noise considerably from the noisy speech. In low SNR environments, when the speech presence segments with weak-energy are corrupted by noises, the USE method tends to aggressively remove the noises resulting in a great possibility of triggering speech distortion. So a C-USE DNN is employed to preserve weak-energy speech presence segments and conservatively remove pure noise segments. The training data for USE contains two subsets. One subset is used to train specific speech enhancement system (SE). The other one is used for speech separation (SS). We use the speech presence segments of the training data for SE and the entire training data for SS to train C-USE. The entire SS training data are used for C-USE training because noises are more difficult to process than speech interference and has greater destructive power, especially in low SNR environments. The speech presence segments of SE training data are extracted as follows. We first apply classical energy-based VAD algorithm on clean speech to detect speech presence segments. Because the noisy speech and clean speech in the training data are one to one correspondence, the speech presence segments can be cut from the noisy speech by using the VAD information of clean speech. The other training configurations are the same with USE.

The architecture of the IRM DNN is 2056-2048-2048-2048-257, which denotes that the size is 2056 (257*7 + 257, including 3 left and 3 right context frames, and 1 frame for NAT) at the input layer, 2048 units for each of the three hidden layers, and 257 for the output layer. The sigmoid activation functions are used in the hidden layers and in the output layer.

The architecture of the VAD DNN is 1799-2048-2048-2048-2, which denotes that the size is 1799 (257*7, including 3 left and 3 right context frames) at the input layer, 2048 units for each of the three hidden layers, and 2 for the output layer. The sigmoid activation functions are used in the hidden layers and the softmax activation function is used in the output layer.

4. DNN-based speech enhancement

4.1. Masking DNN (SE-mask)

In the enhancement stage of IRM DNN, the enhanced DFT coefficients are obtained by multiplying the noisy speech DFT coefficients with the mask $\widehat{\mathbf{IRM}}_n^l$. Specific to this paper, the enhanced LPS features are obtained as follow,

$$\begin{aligned}\widehat{\mathbf{X}}_n &= \ln\left((\widehat{\mathbf{IRM}}_n^l)^2 \cdot \exp(\mathbf{Y}_n)\right) \\ &= 2\ln(\widehat{\mathbf{IRM}}_n^l) + \mathbf{Y}_n\end{aligned} \tag{8}$$

where, $\widehat{\mathbf{X}}_n$ and $\mathbf{Y}_n$ are enhanced and noisy LPS features, respectively. The masking-based speech enhancement system is denoted as SE-mask.

4.2. Mapping DNN (SE-mapp)

In the enhancement stage of spectral mapping DNN, the direct DNN output corresponding to the clean speech is chosen as the enhanced feature as follow,

$$\widehat{\mathbf{X}}_n = \widehat{\mathbf{X}}_n^t \tag{9}$$

The direct mapping-based speech enhancement system is denoted as SE-mapp.

4.3. Post-processing for SE-mapp (SE)

One challenge of spectral mapping DNN is that some distortions are introduced to the estimated clean speech signal because the mapping DNN removes the noise components considerably from the noisy speech. When compared with masking methods, mapping DNN can yield better PESQ scores but the performance of STOI is not so stable. So in the decoding stage of spectral mapping DNN, the estimation of target and interference are fully utilized by an IRM-based post-processing to improve speech intelligibility. Different from Section 3.3 where IRM is directly predicted by a well-trained IRM DNN, the IRM here is calculated by the DNN outputs for each dimension as follows,

$$\widehat{IRM}_n^c(d) = \sqrt{\frac{\exp(\widehat{X}_n^t(d))}{\exp(\widehat{X}_n^t(d)) + \exp(\widehat{X}_n^i(d))}} \tag{10}$$

Table 1
Average performance of speaker-dependent (SE-mask, SE-mapp, SE-Wiener and SE) and speaker-independent (SI-SE and OM-LSA) systems across four isolated noise situations (destroyer engine, factory, babble and mess hall) at different SNRs.

Metrics	SNR	Noisy	SE-mask	SE-mapp	SE-Wiener	SE	SI-SE	OM-LSA
PESQ	−5 dB	1.062	1.259	**1.570**	1.434	1.448	1.258	0.958
	0 dB	1.365	1.674	**2.143**	1.946	1.988	1.776	1.596
	5 dB	1.720	2.098	**2.651**	2.349	2.445	2.256	2.146
	10 dB	2.100	2.478	**3.050**	2.651	2.770	2.616	2.617
STOI	−5 dB	0.534	0.590	0.587	**0.611**	0.608	0.514	0.456
	0 dB	0.662	0.729	0.737	**0.768**	0.765	0.703	0.623
	5 dB	0.782	0.837	0.840	**0.869**	**0.869**	0.845	0.772
	10 dB	0.872	0.905	0.897	0.926	**0.927**	0.920	0.871
FWSegSNR	−5 dB	−3.305	−0.768	**5.160**	−0.181	4.239	3.340	−0.907
	0 dB	−1.644	1.600	**8.108**	3.045	7.609	6.263	2.208
	5 dB	1.312	5.384	9.852	6.492	**10.630**	9.361	5.764
	10 dB	5.437	9.506	11.152	10.130	**13.427**	12.635	9.869
SDR	−5 dB	−5.290	−1.741	**1.781**	0.694	1.259	0.134	−1.918
	0 dB	−0.460	3.749	5.709	5.472	**5.850**	5.153	3.910
	5 dB	4.483	8.547	8.617	9.415	**9.723**	9.286	8.662
	10 dB	9.464	12.724	10.665	13.128	**13.366**	13.068	12.690

where $\widehat{X}_n^i(d)$ is the estimation of interference (background noise or speech interference), $\widehat{IRM}_n^c(d)$ is the calculated IRM. d is feature dimension index. Then, the calculated IRM is used in the post-processing as follows,

$$\widehat{X}_n(d) = \begin{cases} Y_n(d), & \widehat{IRM}_n^c(d) > \gamma \\ \widehat{X}_n^t(d), & \widehat{IRM}_n^c(d) < \lambda \\ (\widehat{X}_n^t(d) + Y_n(d))/2, & \text{otherwise} \end{cases} \tag{11}$$

where $\widehat{X}_n(d)$ and $Y_n(d)$ are the features of enhanced speech and noisy speech, respectively. γ and λ are the thresholds to improve the overall performance. γ and λ in this work are set to 0.75 and 0.1, respectively. The resulting speech enhancement system is denoted as SE. Using the same procedure, different training data derive out SS, USE and C-USE.

4.4. Wiener filtering with mapping DNN (SE-Wiener)

Using the estimation of target and interference from mapping DNN, speech enhancement can also be conducted in Wiener filtering fashion as follow,

$$\begin{aligned} \widehat{\mathbf{X}}_n &= \ln\left(\frac{\exp(\widehat{\mathbf{X}}_n^t)}{\exp(\widehat{\mathbf{X}}_n^t) + \exp(\widehat{\mathbf{X}}_n^i)} \cdot \exp(\mathbf{Y}_n)\right) \\ &= \ln((\widehat{\mathbf{IRM}}_n^c)^2 \cdot \exp(\mathbf{Y}_n)) \\ &= 2\ln(\widehat{\mathbf{IRM}}_n^c) + \mathbf{Y}_n \end{aligned} \tag{12}$$

The results of using Wiener filtering with $\widehat{\mathbf{IRM}}_n^c$ is denoted as SE-Wiener.

4.5. DNN integration using VAD DNN (JDNN-SE-VAD)

After the enhancement of USE and C-USE, we use VAD DNN to integrate the enhanced features at frame level. The integration is performed with the VAD classification probabilities as follows,

$$\widehat{\mathbf{X}}_n = \alpha\widehat{\mathbf{X}}_n^1 + (1-\alpha)\widehat{\mathbf{X}}_n^2 \tag{13}$$

where α is the probability of the speech presence, and $(1-\alpha)$ is speech absence probability. α is taken directly from the VAD DNN. $\widehat{\mathbf{X}}$,$\widehat{\mathbf{X}}^1$ and $\widehat{\mathbf{X}}^2$ are the vectors of the final enhanced features, and features enhanced by C-USE and USE, respectively. The integration can utilize C-USE's good quality in speech presence segments and USE's pure noise reduction in the speech absence segments. The resulting joint DNN based speech enhancement system is denoted as JDNN-SE-VAD.

4.6. DNN integration using IRM

Compared with VAD DNN, IRM DNN can predict speech presence probability for each time-frequency unit. With the high-resolution prediction, the integration of USE and C-USE can be performed as follows,

$$\widehat{\mathbf{X}}_n = \widehat{\mathbf{IRM}}_n^l \cdot \widehat{\mathbf{X}}_n^1 + (1 - \widehat{\mathbf{IRM}}_n^l) \cdot \widehat{\mathbf{X}}_n^2 \tag{14}$$

where $\widehat{\mathbf{IRM}}_n^l$ is the output of IRM DNN. $\widehat{\mathbf{X}}$,$\widehat{\mathbf{X}}^1$ and $\widehat{\mathbf{X}}^2$ are the vectors of the final enhanced features, and features enhanced by C-USE and USE, respectively. The operation $\cdot$ denotes element-wise multiplication. The resulting joint DNN system is denoted as JDNN-SE-*IRM*^{l}. Both the trained $\widehat{\mathbf{IRM}}_n^l$ and the estimated $\widehat{\mathbf{IRM}}_n^c$ can be used to integrate the enhanced features. The results of using $\widehat{\mathbf{IRM}}_n^c$ is denoted as JDNN-SE-*IRM*^{c}.

5. Experimental results and analysis

5.1. Experimental configurations

In Xu et al. (2015a), 104 noise types[1] were used as the noise signals for synthesizing noisy training samples. In this study, we add another 200 hours of real-world noise[2] to handle a wide range of additive noise in the real-world situations. Six hours of speech interferences covering 30 males and 30 females were also used for speech separation. On the other hand, 2 h of Hi-Fi Mandarin data were recorded by a target female speaker as our clean data. They are added with the above-mentioned background noise and speech interferences at 5 levels of SNR (20 dB, 15 dB, 10 dB, 5 dB and 0 dB) to build a multi-condition stereo training set. This resulted in a collection of approximately 100 h of noisy training data (including two subsets, 80 h for speech enhancement and the remaining 20 h for speech separation) used to train DNN models. The enhancement data are more than separation data since there are more noise types covered.

The whole 100 h of training data were used to train a unified speech enhancement (USE) model. The two training subsets were also used for training the specific speech enhancement and speech separation system,

[1] The 104 noise types are N1-N17: crowd noise; N18-N29: machine noise; N30-N43: alarm and siren; N44-N46: traffic and car noise; N47-N55: animal sound; N56-N69: water sound; N70-N78: wind; N79-N82: bell; N83-N85: cough; N86: clap; N87: snore; N88: click; N88-N90: laugh; N91-N92: yawn; N93: cry; N94: shower; N95: tooth brushing; N96-N97: footsteps; N98: door moving; N99-N100: phone dialing; N101: AWGN; N102: babble; N103: restaurant; and N104: street.

[2] The noise types are vehicle: bus, train, plane and car; exhibition hall; meeting room; office; emporium; family living room; factory; bus station; and mess hall.

Table 2
Performance comparison of speaker-dependent systems in four isolated noise situations at different SNRs.

Metrics	SNR	Noisy	SE	SS	USE	C-USE	JDNN-SE-VAD	JDNN-SE-IRM^l	JDNN-SE-IRM^c
Destroyer engine noise									
PESQ	−5 dB	1.145	1.565	0.849	1.616	1.819	1.865	**1.868**	1.821
	0 dB	1.359	2.015	1.293	2.030	2.193	**2.211**	2.208	2.177
	5 dB	1.649	2.461	1.720	2.444	2.495	2.504	**2.540**	2.517
	10 dB	2.001	2.807	2.176	2.796	2.747	2.744	**2.822**	2.801
STOI	−5 dB	0.581	0.606	0.425	0.624	0.689	**0.699**	0.675	0.677
	0 dB	0.696	0.759	0.559	0.769	**0.806**	0.805	0.799	0.800
	5 dB	0.796	0.867	0.704	0.867	0.880	0.880	0.880	**0.882**
	10 dB	0.870	0.926	0.815	0.923	0.923	0.925	**0.928**	**0.928**
FWSegSNR	−5 dB	−4.337	6.379	−1.096	**7.165**	3.963	3.587	6.143	5.992
	0 dB	−2.690	8.818	2.154	9.243	7.630	8.204	**9.267**	8.914
	5 dB	0.257	11.395	5.904	11.511	10.220	10.502	**11.632**	11.301
	10 dB	4.326	13.715	8.944	**13.725**	12.175	12.301	13.679	13.315
SDR	−5 dB	−5.237	3.469	−0.014	3.683	3.109	3.540	**3.839**	3.625
	0 dB	−0.405	7.473	4.320	7.468	7.408	7.642	**7.654**	7.576
	5 dB	4.536	11.130	7.523	10.980	11.055	11.130	**11.132**	11.120
	10 dB	9.515	**14.484**	10.393	14.294	14.335	14.386	14.401	14.451
Factory noise									
PESQ	−5 dB	0.968	1.416	0.680	1.541	1.600	1.549	1.690	**1.698**
	0 dB	1.305	1.961	1.061	2.029	2.048	2.015	**2.124**	2.109
	5 dB	1.691	2.422	1.509	2.449	2.429	2.400	**2.501**	2.485
	10 dB	2.098	2.725	1.972	2.754	2.735	2.700	**2.784**	2.769
STOI	−5 dB	0.526	0.599	0.443	0.594	0.615	0.617	**0.628**	**0.628**
	0 dB	0.649	0.755	0.559	0.754	0.759	0.764	**0.769**	**0.769**
	5 dB	0.770	0.862	0.676	0.863	0.857	0.863	**0.868**	0.866
	10 dB	0.865	0.922	0.786	0.923	0.918	0.922	**0.925**	0.923
FWSegSNR	−5 dB	−4.728	3.106	−3.377	**4.701**	0.729	0.094	2.284	2.498
	0 dB	−3.231	6.588	−2.174	**7.685**	3.238	3.381	5.311	5.019
	5 dB	−0.502	9.761	0.219	**10.063**	6.006	6.378	8.419	7.634
	10 dB	3.365	12.032	3.718	**12.243**	8.951	9.289	11.293	10.351
SDR	−5 dB	−5.230	0.318	−7.492	**1.375**	-0.403	-0.085	0.749	0.559
	0 dB	−0.419	5.244	−1.632	**5.706**	4.281	4.779	5.310	4.935
	5 dB	4.514	9.046	3.655	**9.242**	8.361	8.652	9.190	8.714
	10 dB	9.490	12.552	7.664	12.654	12.259	12.420	**12.897**	12.457
Babble noise									
PESQ	−5 dB	1.029	1.396	1.153	1.397	1.608	1.578	**1.629**	1.600
	0 dB	1.412	1.992	1.702	1.977	2.109	2.078	**2.129**	2.113
	5 dB	1.793	2.446	2.150	2.437	2.467	2.446	**2.504**	2.483
	10 dB	2.188	2.783	2.485	2.750	2.727	2.715	**2.781**	2.764
STOI	−5 dB	0.514	0.618	0.556	0.601	0.646	**0.649**	0.648	0.646
	0 dB	0.646	0.773	0.714	0.761	0.781	**0.784**	0.782	0.783
	5 dB	0.776	0.871	0.823	0.865	0.871	**0.874**	**0.874**	0.875
	10 dB	0.873	0.927	0.887	0.923	0.924	**0.927**	**0.927**	0.928
FWSegSNR	-5 dB	−1.965	4.739	0.348	**5.801**	2.601	2.133	3.858	3.847
	0 dB	−0.209	8.286	3.736	**8.488**	5.636	5.939	7.261	6.867
	5 dB	2.865	11.224	7.533	**11.247**	8.846	9.041	10.559	10.098
	10 dB	7.175	**14.248**	11.205	14.172	12.376	12.345	13.813	13.520
SDR	-5 dB	−5.383	0.831	−1.939	**0.917**	−0.203	0.220	0.874	0.439
	0 dB	−0.525	5.378	3.491	5.277	4.672	5.225	**5.643**	5.066
	5 dB	4.435	9.238	7.939	9.075	8.778	9.070	**9.476**	9.066
	10 dB	9.425	**12.987**	11.521	12.673	12.548	12.744	12.798	12.800
Mess hall noise									
PESQ	-5 dB	1.106	1.416	1.123	1.432	1.562	1.540	**1.570**	1.536
	0 dB	1.384	1.982	1.587	1.982	2.026	2.007	**2.058**	2.041
	5 dB	1.745	2.452	2.016	2.432	2.425	2.391	**2.454**	2.436
	10 dB	2.112	2.765	2.399	2.741	2.695	2.674	**2.751**	2.735
STOI	−5 dB	0.516	0.610	0.553	0.597	0.619	0.622	**0.624**	0.622
	0 dB	0.655	0.771	0.706	0.768	0.769	0.772	**0.777**	0.776
	5 dB	0.785	0.876	0.820	0.871	0.868	0.870	**0.874**	0.873
	10 dB	0.878	0.932	0.890	**0.929**	0.922	0.925	**0.929**	**0.929**
FWSegSNR	−5 dB	−2.189	2.732	−1.018	**3.539**	1.769	1.288	2.358	2.495
	0 dB	−0.444	6.744	1.740	**7.330**	4.629	4.418	5.710	5.564
	5 dB	2.629	10.228	5.268	**10.496**	7.752	7.787	9.144	8.742
	10 dB	6.880	13.711	9.468	**13.765**	11.252	11.365	12.800	12.389
SDR	-5 dB	−5.310	0.416	−1.744	0.378	−0.558	−0.321	**0.506**	−0.058
	0 dB	−0.490	5.303	3.163	5.314	4.655	4.845	**5.715**	4.998
	5 dB	4.447	9.476	7.513	9.342	8.991	9.135	**9.839**	9.223
	10 dB	9.425	**13.439**	11.264	13.167	12.836	12.998	13.217	13.193

denoted as SE and SS, respectively. Then, we applied classical energy-based VAD algorithm on clean speech to detect speech presence segments. Because the noisy speech and clean speech in the training data are one to one correspondence, so speech presence segments can be cut from noisy speech by using the VAD information of clean speech. The speech presence segments of the 80-hours training data for SE and the entire 20 h training data for SS were used to train a conservative unified speech enhancement (C-USE) model. The whole 20 h SS training data

Table 3
Performance comparison of speaker-dependent systems in isolated speech interference situation at different SNRs.

Metrics	SNR	Noisy	SE	SS	USE	C-USE	JDNN-SE-VAD	JDNN-SE-IRM^I
Speech interference								
PESQ	−5 dB	1.360	1.409	**1.624**	1.551	1.559	1.613	1.599
	0 dB	1.728	1.827	1.973	1.988	1.982	1.993	**2.043**
	5 dB	2.084	2.247	2.342	2.378	2.382	2.402	**2.422**
	10 dB	2.429	2.609	2.656	2.736	2.716	2.735	**2.758**
STOI	−5 dB	0.571	0.601	**0.683**	0.674	0.666	0.666	0.675
	0 dB	0.677	0.715	0.767	0.768	0.767	0.768	**0.773**
	5 dB	0.779	0.813	0.842	0.845	0.849	0.851	**0.852**
	10 dB	0.863	0.887	0.896	0.905	0.906	**0.910**	0.909
FWSegSNR	−5 dB	0.880	2.619	**7.918**	6.444	5.334	4.739	5.434
	0 dB	3.529	5.813	**11.144**	10.462	9.757	9.295	9.787
	5 dB	7.936	10.735	14.656	**14.774**	14.699	14.155	14.627
	10 dB	13.781	16.552	18.052	18.655	**18.706**	18.453	18.541
SDR	−5 dB	−5.213	−4.301	**2.946**	0.215	−0.819	−1.168	−0.460
	0 dB	−0.390	1.053	**6.179**	4.732	4.394	4.293	4.782
	5 dB	4.547	6.233	**9.342**	8.785	8.837	8.852	9.161
	10 dB	9.524	11.188	12.490	12.738	12.759	**12.961**	12.846

were used for C-USE training because C-USE is mainly designed for serious noise environments. The entire 100 h of training data were also used to train a VAD DNN and an IRM DNN. Another 50 utterances recorded from the target speaker were used to construct the test set for each combination of noise types (isolated noise: babble, factory, destroyer engine (Varga and Steeneken, 1993), mess hall; speech interference: different gender speech interference; mixed noise: isolated noise overlapped with speech interference) and various SNR levels (−5 dB, 0 dB, 5 dB and 10 dB). The noise and speech text in the test set are different from those in the training set.

As for signal analysis, speech waveform was down-sampled to 16KHz, and the corresponding frame length was set to 512 samples (or 32 msec) with a 256 samples frame shift. Short-time Fourier analysis (Allen and Rabiner, 1977) was used to compute the DFT of each overlapping windowed frame. Then, the 257-dimensional LPS features were used to train the DNNs. The performance was evaluated using four measures, namely frequency-weighted segmental SNR (FWSegSNR) (Hu and Loizou, 2008) and source-to-distortion ratio (SDR) (Vincent et al., 2006), STOI and PESQ. STOI is shown to be highly correlated to human speech intelligibility while PESQ has a high correlation with subjective scores.

5.2. Advantages of a speaker-dependent system

Table 1 presents the average results of speaker-dependent and speaker-independent systems across four isolated noise types. We first focus on speaker-dependent scenario to reveal the advantage of IRM-based post-processing described in Section 4.3. SE-mapp is a direct mapping DNN described in Section 4.2. When compared with masking method SE-mask, we can find SE-mapp obtained better PESQ and FWSegSNR performance but the performances of STOI and SDR were not so consistent. SE-mapp removed more noise than SE-mask from the noisy speech, so some distortions were introduced, especially at high SNR. Because it is more desirable to enhance the intelligibility rather than the quality of speech, Wiener filtering and IRM-based post-processing were applied to spectral mapping DNN. The resulting SE-Wiener system and SE system sacrificed certain PESQ gains to attain the improvement of STOI. SE-Wiener and SE have almost the same performance of STOI, but SE performed better on PESQ, FWSegSNR and SDR. So, in the following experiments, IRM-based post-processing is applied to mapping DNN.

Speech enhancement systems are usually designed in a speaker-independent scenario (SI-SE, OM-LSA). Compared with DNN-based SI-SE, the classical speech enhancement algorithm OM-LSA (Cohen and Berdugo, 2001) has its limits to improve speech intelligibility. When speech enhancement system is adopted in a speaker-dependent scenario, it achieves reasonable performance that could not be achieved with the SI-SE system. In Table 1, SE is trained using speech from only one speaker. The noise data and other configurations are the same as with the SI-SE model. The result shows SE outperformed SI-SE at all SNRs for all metrics, especially at low SNRs, e.g., 0.094 STOI improvement at −5 dB SNR (from 0.514 to 0,608). This indicates that the speaker-dependent system is much more effective than the speaker-independent system.

5.3. Performance of a unified system in isolated interference situations

In this section, we analyze the performance of unified system (USE) when only noise or speech interference is present. We used two noise categories to discuss the performance of USE for speech enhancement. The first category is general noise: destroyer engine, factory. Another category is speech-like noise: babble, mess hall. The performance of different systems for the two noise categories at different SNRs is shown in Table 2.

In destroyer engine and factory noise conditions, with the comparison of SS and SE, we first observe that the performance of SS was dramatically degraded, even worse than unprocessed noisy speech at low SNR. Conversely, SE improved the PESQ score effectively at all SNRs. A comparison of SE and USE is a major focus. The results show that the unified system USE can achieve almost the same effect when compared with the specific system SE. SS had poor performances in the general noise situations. However, it is interesting to note something different when dealing with speech-like noise: babble, mess hall. In the lower part of Table 2, SS has obtained positive effects on PESQ, STOT, FWSegSNR and SDR when compared with unprocessed noisy speech at all SNRs. This can be explained by the speech-like noise having some similarities with the speech interference used in SS training. In the two noise situations, unified system USE still yielded a fairly good performance when compared with SE.

With USE as a benchmark, we analyze the performance of C-USE, JDNN-SE-VAD, JDNN-SE-IRM^I and JDNN-SE-IRM^c. C-USE yielded better PESQ and STOI performance, especially at low SNRs, e.g., 0.211 PESQ improvement (from 1.397 to 1.608) and 0.045 STOI improvement (from 0.601 to 0.646), at −5 dB SNR of babble noise. The FWSegSNR and SDR performance suffered degradation because C-USE is designed to preserve weak-energy speech presence segments but conservatively remove speech absence segments. When VAD DNN was used to integrate USE and C-USE at frame level, the resulting system JDNN-SE-VAD obtained inconsistent improvement on PESQ, STOI and limited improvement on FWSegSNR. With high-resolution IRM, USE and C-USE were effectively integrated at time-frequency unit level. The results of using IRM^I and IRM^c are almost the same on PESQ and STOI,

Table 4
Performance comparison of speaker-dependent systems in four mixed noise conditions at different SNRs.

Metrics	SNR	Noisy	SE	SS	USE	C-USE	JDNN-SE-VAD	JNN-SE-IRM^l
Destroyer engine noise + speech interference								
PESQ	−5 dB	0.990	1.258	1.258	1.394	1.455	1.471	**1.546**
	0 dB	1.496	1.741	1.689	1.845	1.894	1.896	**2.035**
	5 dB	1.883	2.217	2.111	2.302	2.321	2.333	**2.447**
	10 dB	2.268	2.625	2.475	2.674	2.645	2.672	**2.774**
STOI	−5 dB	0.540	0.565	0.570	0.600	0.610	0.610	**0.626**
	0 dB	0.656	0.706	0.700	0.736	0.742	0.743	**0.763**
	5 dB	0.765	0.816	0.803	0.837	0.838	0.841	**0.856**
	10 dB	0.855	0.892	0.875	0.902	0.901	0.906	**0.913**
FWSegSNR	−5 dB	−1.524	2.613	2.913	**3.681**	2.300	2.900	3.008
	0 dB	0.544	5.436	5.598	**6.775**	5.864	5.638	6.490
	5 dB	4.123	9.300	8.601	10.602	9.584	9.410	**10.712**
	10 dB	9.085	13.882	12.486	14.602	13.435	13.436	**14.644**
SDR	−5 dB	−5.212	1.201	**1.516**	1.361	−2.940	−3.005	−1.825
	0 dB	−0.391	**6.061**	4.960	5.968	2.953	2.802	3.892
	5 dB	4.546	**10.279**	7.802	10.101	8.230	8.206	9.001
	10 dB	9.524	**13.961**	10.899	13.809	12.613	12.805	13.175
Factory noise + speech interference								
PESQ	−5 dB	0.942	1.193	0.709	1.287	1.348	1.373	**1.375**
	0 dB	1.326	1.796	1.167	1.855	1.886	1.868	**1.932**
	5 dB	1.750	2.286	1.616	2.329	2.343	2.319	**2.373**
	10 dB	2.159	2.681	2.110	2.713	2.688	2.671	**2.729**
STOI	−5 dB	0.510	0.532	0.457	0.540	0.559	**0.563**	**0.563**
	0 dB	0.635	0.707	0.573	0.713	0.721	0.723	**0.726**
	5 dB	0.759	0.837	0.698	0.841	0.841	0.842	**0.846**
	10 dB	0.859	0.911	0.814	0.914	0.911	0.913	**0.916**
FWSegSNR	−5 dB	−3.979	1.641	−3.053	**2.731**	0.784	0.353	1.681
	0 dB	−2.379	4.675	−1.448	**5.374**	3.051	2.987	4.264
	5 dB	0.485	8.113	1.322	**8.457**	6.014	6.073	7.478
	10 dB	4.531	11.279	5.124	**11.512**	9.144	9.279	10.690
SDR	−5 dB	−5.211	−3.169	−5.288	**−2.675**	−2.948	−2.776	−2.698
	0 dB	−0.402	2.924	0.474	**3.176**	2.676	2.933	3.054
	5 dB	4.532	7.977	5.081	**8.095**	7.658	7.902	8.023
	10 dB	9.509	12.458	8.782	**12.462**	12.144	12.354	12.432
Babble noise + speech interference								
PESQ	−5 dB	0.949	1.098	1.093	1.138	1.352	**1.358**	1.331
	0 dB	1.394	1.725	1.604	1.720	1.827	1.846	**1.861**
	5 dB	1.782	2.264	2.063	2.248	2.281	2.291	**2.322**
	10 dB	2.179	2.683	2.422	2.676	2.621	2.632	**2.692**
STOI	−5 dB	0.492	0.530	0.531	0.535	0.571	**0.573**	0.567
	0 dB	0.622	0.709	0.680	0.706	0.724	**0.727**	0.725
	5 dB	0.752	0.832	0.801	0.830	0.835	0.838	**0.839**
	10 dB	0.856	0.908	0.875	0.907	0.903	0.907	**0.909**
FWSegSNR	−5 dB	−2.123	2.176	−0.016	**2.741**	1.559	1.217	2.100
	0 dB	−0.336	5.490	3.148	**5.722**	4.367	4.308	5.147
	5 dB	2.794	9.046	6.643	**9.146**	7.648	7.518	8.661
	10 dB	7.109	12.886	10.141	**12.921**	11.215	11.144	12.327
SDR	−5 dB	−5.217	−2.381	**−1.030**	−1.367	−1.923	−1.843	−1.482
	0 dB	−0.404	3.598	3.847	**4.090**	3.525	3.678	3.995
	5 dB	4.530	8.499	7.984	8.513	8.252	8.378	**8.596**
	10 dB	9.506	**12.895**	11.401	12.629	12.349	12.586	12.665
Mess hall noise + speech interference								
PESQ	−5 dB	0.917	1.183	1.131	1.241	1.363	**1.381**	1.366
	0 dB	1.432	1.777	1.585	1.793	1.857	1.858	**1.884**
	5 dB	1.800	2.284	2.022	2.271	2.317	2.308	**2.339**
	10 dB	2.185	2.694	2.400	2.675	2.635	2.637	**2.688**
STOI	−5 dB	0.507	0.547	0.533	0.556	0.576	0.577	**0.578**
	0 dB	0.643	0.722	0.685	0.724	0.733	0.733	**0.737**
	5 dB	0.770	0.842	0.802	0.841	0.842	0.844	**0.847**
	10 dB	0.867	0.914	0.880	0.912	0.909	0.911	**0.914**
FWSegSNR	-5 dB	−1.837	1.645	−0.717	**2.221**	1.454	1.128	1.721
	0 dB	−0.037	5.216	2.154	**5.518**	4.242	4.012	4.783
	5 dB	3.086	9.073	5.641	**9.031**	7.514	7.362	8.375
	10 dB	7.429	13.114	9.829	**12.979**	11.150	11.065	12.325
SDR	−5 dB	−5.325	−1.747	−1.283	**−1.210**	−1.853	−1.759	−1.377
	0 dB	−0.487	3.772	3.586	**4.089**	3.602	3.727	4.032
	5 dB	4.462	7.693	7.558	8.511	8.299	8.422	**8.647**
	10 dB	9.446	10.371	11.204	12.632	12.358	12.522	**12.701**

but JDNN-SE-IRM^l performed slightly better than JDNN-SE-IRM^c on FWSegSNR and SDR. In the following experiments, IRM^l is chosen for JDNN-SE-IRM method. When compared with C-USE, JDNN-SE-IRM^l could further improve PESQ, STOI and at the same time shorten the FWSegSNR gap between C-USE and USE. For example, 0.090 PESQ improvement, 0.013 STOI improvement and 1.555 FWSegSNR improvement were achieved at −5 dB SNR of factory noise. The results demonstrate the effectiveness of the ensemble framework in low SNR

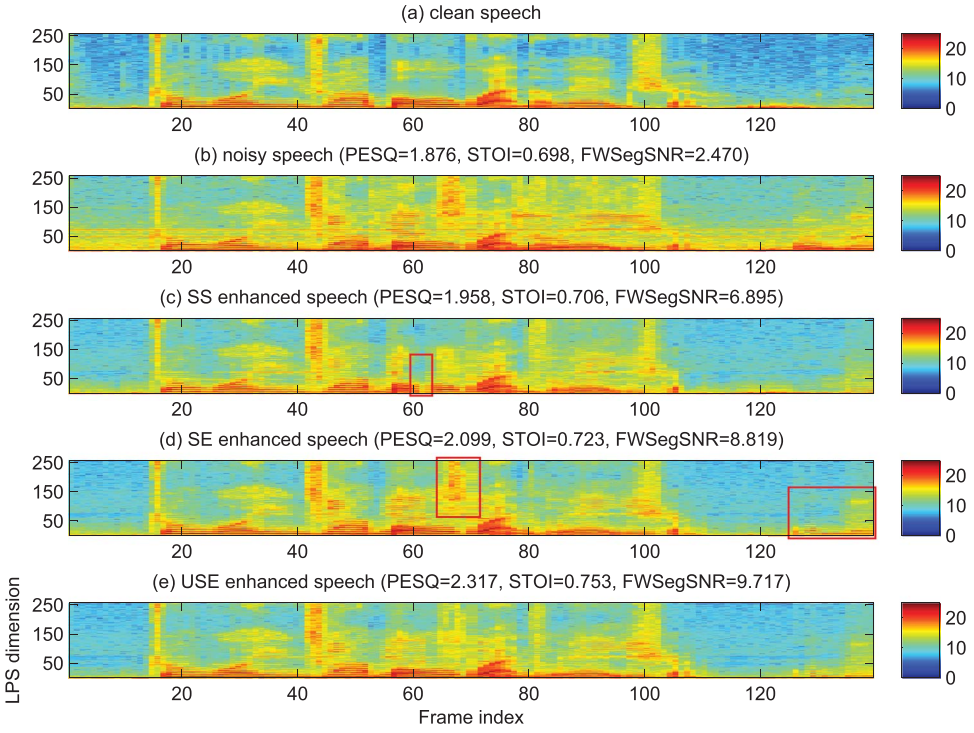

Fig. 6. Spectrograms of an utterance corrupted by mixed noise (destroyer engine + speech interference) at 5 dB and enhanced by speaker-dependent systems: (a) clean speech, (b) noisy speech, (c) SS enhanced speech, (d) SE enhanced speech and (e) USE enhanced speech.

environments. It should be noted that this work preliminarily investigated the implementation of an original DNN VAD in the ensemble framework. If more advanced DNN VAD methods, like the work in Zhang and Wang (2016a), Hwang et al. (2016) and Zazo et al. (2016), are used, the result of JDNN-SE-VAD may be improved.

For speech separation, the speech from different gender speakers were used as speech interference. In Table 3, we show the results of speech separation. The specific speech separation system SS effectively improved the PESQ, STOI and FWSegSNR. The system SE also achieved slight improvement compared with Noisy due to some speech-like noises being covered in SE training. The proposed system USE obtained a performance comparable with SS. Note, SS was better than USE at $^{-}5$ dB. Then, with the increase of SNR, USE achieved even better performance than SS. We try to explain this phenomenon as follows. In high SNR environments, the speech interference segments have weak energy and sound similar to babble noise. When speech interference segments are seen as noise, the existence of noisy data with background noise becomes helpful for speech separation. C-USE achieved almost the same results with USE because both the speech presence segments and speech absence segments of SS training data were used to train C-USE. The reason for treating SS and SE differently is that noises are more difficult to process than speech interference and has greater destructive power, especially in low SNR environments. So C-USE is only designed to address serious noise situations. Based on these two comparable results, JDNN-SE-IRM^l still yielded a little improvement on PESQ and STOI.

In summary, the unified system USE can maintain the system performance achieved by the SE and SS systems when isolated interference is present. The training data for the two specific tasks are also found to be complementary. The data corrupted by background noise become helpful for speech separation with increasing SNR, and the speech interference is useful for dealing with speech-like noise. In addition, JDNN-SE-IRM can take advantages of both USE and C-USE effectively.

5.4. Performance of a unified system in mixed noise situations

The performance of unified system USE in four mixed noise situations (by adding the speech interference to each isolated noise) is presented in Table 4. The results show that SS and SE have improved PESQ, STOI, FWSegSNR and SDR performance compared with unprocessed speech in three mixed noise situations. In the factory noise adding speech interference condition, SS caused performance degradation. Fig. 6 presents spectrograms of an utterance example corrupted by mixed noise at 5 dB and the enhanced results from the speaker-dependent system. As shown in the red rectangle of Fig. 6(c) for SS enhanced speech, SS degraded the continuity of target speech in the presence of noise. SE was able to protect target speech. However, SE was insufficient for removing speech interference that had a high energy, as shown in the red rectangle of Fig. 6(d). From the above analysis, we find the two isolated systems (SE, SS) have their own shortcomings when addressing mixed noises. The proposed USE significantly outperformed the individual SE and SS systems when mixed noises were present for three metrics at all SNRs, e.g., 0.136 PESQ improvement (from 1.258 for SE and SS to 1.394 for USE) and 0.035 STOI improvement (from 0.565 for SE and SS to 0.600 for USE) at $^{-}5$ dB of destroyer engine noise adding speech interference condition. The problems in the red rectangles of Fig. 6(c)(d), where SS degraded target speech and SE removed speech interference insufficiently, have largely been solved in Fig. 6(e) by utilizing the complementarity of SS and SE. More results can be found at the demo website[3]. In short, unified system could yield much better results over individual enhancement or separation systems in mixed noise scenarios.

In mixed noise situations, the performance of C-USE, JDNN-SE-VAD and JDNN-SE-IRM^l were consistent with isolated interference conditions. C-USE improved the PESQ and STOI results of USE except 10 dB case. JDNN-SE-VAD and JDNN-SE-IRM^l performed effectively to take advantages of both USE and C-USE.

In order to verify the generalization capacity of the proposed framework, another set of experiments for three additional speakers were added. Table 5 presents the average performance of speaker-dependent systems across four speakers (two females and two males) in the mixed noise conditions. Through the analysis of Table 5, the conclusion on the basis of more speakers is the same as before.

[3] http://home.ustc.edu.cn/~gtian09/demos/USE_DNN_Journal.html.

Table 5
Average performance of speaker-dependent systems across four speakers (two females and two males) in four mixed noise conditions.

Metrics	SNR	Noisy	SE	SS	USE	C-USE	JDNN-SE-VAD	JDNN-SE-*IRM*[1]
Destroyer engine noise + speech interference								
PESQ	−5 dB	1.120	1.454	1.252	1.554	1.592	1.624	**1.639**
	0 dB	1.413	2.017	1.706	2.090	2.111	2.135	**2.164**
	5 dB	1.748	2.481	2.157	2.525	2.526	2.547	**2.576**
	10 dB	2.109	2.840	2.531	2.861	2.826	2.846	**2.890**
STOI	−5 dB	0.580	0.649	0.596	0.670	0.676	0.677	**0.683**
	0 dB	0.690	0.786	0.728	0.797	0.799	0.800	**0.806**
	5 dB	0.786	0.864	0.824	0.870	0.869	0.871	**0.876**
	10 dB	0.859	0.910	0.887	0.912	0.911	0.913	**0.916**
FWSegSNR	−5 dB	−3.348	**3.856**	1.064	3.254	2.958	2.829	3.030
	0 dB	−1.626	**7.287**	4.019	6.591	6.584	6.086	6.488
	5 dB	1.374	**10.506**	7.088	9.882	9.773	9.241	9.841
	10 dB	5.513	**13.746**	10.686	13.258	12.873	12.411	13.140
SDR	−5 dB	−5.598	0.193	−0.457	**0.492**	−0.649	−0.588	−0.268
	0 dB	−0.699	5.399	4.316	**5.552**	4.770	4.766	5.060
	5 dB	4.268	9.865	8.400	**9.912**	9.340	9.356	9.619
	10 dB	9.257	13.933	12.384	**13.945**	13.538	13.605	13.778
Factory noise + speech interference								
PESQ	−5 dB	1.069	1.450	1.079	1.527	1.541	**1.572**	**1.572**
	0 dB	1.392	2.011	1.543	2.063	2.077	2.087	**2.101**
	5 dB	1.781	2.465	2.013	2.500	2.504	2.510	**2.525**
	10 dB	2.168	2.829	2.438	2.854	2.834	2.838	**2.864**
STOI	−5 dB	0.562	0.628	0.549	0.638	0.645	0.647	**0.649**
	0 dB	0.680	0.776	0.682	0.781	0.783	0.785	**0.787**
	5 dB	0.785	0.866	0.791	0.867	0.867	0.867	**0.869**
	10 dB	0.863	0.914	0.869	0.913	0.913	0.913	**0.914**
FWSegSNR	−5 dB	−3.966	**2.553**	−0.976	2.396	1.794	1.573	2.072
	0 dB	−2.424	**6.029**	1.320	5.501	4.627	4.415	5.096
	5 dB	0.325	**9.378**	4.364	8.629	7.535	7.321	8.211
	10 dB	4.160	**12.438**	7.953	11.833	10.585	10.461	11.421
SDR	−5 dB	−5.613	−1.354	−3.304	−1.116	−1.244	**−1.113**	−1.116
	0 dB	−0.716	4.289	2.460	**4.448**	4.288	4.386	4.427
	5 dB	4.251	8.982	7.243	**9.046**	8.871	8.913	9.024
	10 dB	9.240	13.306	11.515	**13.348**	13.162	13.215	13.328
Babble noise + speech interference								
PESQ	−5 dB	1.107	1.425	1.348	1.510	1.586	**1.604**	1.602
	0 dB	1.458	2.000	1.818	2.061	2.104	2.125	**2.126**
	5 dB	1.826	2.476	2.265	2.515	2.521	2.533	**2.547**
	10 dB	2.203	2.839	2.624	2.866	2.830	2.841	**2.873**
STOI	−5 dB	0.552	0.627	0.611	0.641	0.651	0.653	**0.655**
	0 dB	0.671	0.775	0.740	0.779	0.783	0.785	**0.786**
	5 dB	0.780	0.863	0.836	0.863	0.863	0.864	**0.866**
	10 dB	0.860	0.913	0.893	0.912	0.910	0.911	**0.913**
FWSegSNR	−5 dB	−0.851	2.386	0.095	**2.551**	2.062	1.952	2.239
	0 dB	−0.838	**6.072**	3.032	5.952	5.215	5.151	5.594
	5 dB	2.158	**9.847**	6.734	9.423	8.592	7.586	9.099
	10 dB	6.246	**13.450**	10.705	12.922	12.122	11.942	12.645
SDR	−5 dB	−5.609	−1.611	−1.793	**−0.274**	−1.243	−1.120	−1.014
	0 dB	−0.718	4.249	3.326	**4.546**	4.283	4.372	4.520
	5 dB	4.245	9.093	8.042	**9.138**	8.919	8.959	9.132
	10 dB	9.232	**13.420**	12.315	13.342	13.193	13.254	13.356
Mess hall noise + speech interference								
PESQ	−5 dB	1.110	1.510	1.384	1.594	1.629	1.657	**1.660**
	0 dB	1.503	2.078	1.834	2.125	2.154	2.158	**2.170**
	5 dB	1.846	2.528	2.265	2.544	2.549	2.549	**2.572**
	10 dB	2.211	2.868	2.621	2.875	2.833	2.842	**2.878**
STOI	−5 dB	0.574	0.669	0.640	0.681	0.685	0.681	**0.692**
	0 dB	0.692	0.800	0.765	0.803	0.805	0.801	**0.809**
	5 dB	0.794	0.877	0.851	0.876	0.877	0.874	**0.879**
	10 dB	0.868	0.920	0.903	0.918	0.918	0.917	**0.920**
FWSegSNR	−5 dB	−2.387	**4.028**	0.869	3.471	3.044	2.785	3.200
	0 dB	−0.593	**7.358**	4.098	6.679	6.175	5.984	6.368
	5 dB	2.505	**10.678**	7.707	9.978	9.386	9.122	9.648
	10 dB	6.719	**14.145**	11.607	13.551	12.860	12.452	13.220
SDR	−5 dB	−5.653	−0.343	−0.995	**0.029**	−0.207	−0.118	0.016
	0 dB	−0.756	4.924	3.974	**5.139**	4.952	4.997	5.130
	5 dB	4.211	9.252	8.369	9.529	9.374	9.384	**9.545**
	10 dB	9.201	13.045	12.499	**13.649**	13.478	13.516	13.645

6. Conclusion

We present a unified DNN approach to reduce both background noise and speech interference in a speaker-dependent scenario. A speaker-dependent system is much more robust than a speaker-independent system and can unify speech enhancement and speech separation. Empirical results demonstrate that the unified system USE can achieve fairly good results compared with specific systems where only

noise or speech interference is present, and it can achieve better performance for noise and speech interference mixed conditions. Moreover, we use a joint DNN based framework to improve the performance of USE in low SNR environments. In this ensemble learning-based method, speech presence segments and speech absence segments are presented to C-USE DNN and USE DNN separately. A VAD DNN and an IRM DNN are investigated to integrate the outputs of C-USE and USE. The resulting system can take advantages of both C-USE and USE to yield improvement in serious noise environments. In this paper, we mainly investigate the concept of speaker-dependent speech enhancement and employ an ensemble framework for low SNR environments. The upgrade of modules in the proposed framework, like advanced VAD method and speech enhancement architecture will be an important future research problem.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 61671422 and U1613211, in part by the National Key Research and Development Program of China under Grant 2017YFB1002200, in part by the MOE-Microsoft Key Laboratory of USTC.

References

Allen, J.B., Rabiner, L.R., 1977. A unified approach to short-time fourier analysis and synthesis. Proc. IEEE 65 (11), 1558–1564.
Benesty, J., Makino, S., Chen, J.D., 2005. Speech Enhancement. Springer.
Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. Acoust. Speech Signal Process. IEEE Transa. 27 (2), 113–120.
Cohen, I., Berdugo, B., 2001. Speech enhancement for non-stationary noise environments. Signal Process. 81 (11), 2403–2418.
Dahl, G.E., Yu, D., Deng, L., Acero, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. Audio Speech Lang. Process. IEEE Trans. 20 (1), 30–42.
Du, J., Huo, Q., 2008. A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions. Interspeech. pp. 569–572.
Du, J., Tu, Y., Dai, L.-R., Lee, C.-H., 2016. A regression approach to single-channel speech separation via high-resolution deep neural networks. Audio Speech Lang. Process. IEEE/ACM Trans. 24 (8), 1424–1437.
Du, J., Tu, Y., Xu, Y., Dai, L.-R., Lee, C.-H., 2014. Speech separation of a target speaker based on deep neural networks. ICSP. pp. 473–477.
Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. Acoustics Speech Signal Processing IEEE Trans 32 (6), 1109–1121.
Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. Acoustics Speech Signal Process. IEEE Trans. 33 (2), 443–445.
Fan, H.T., Hung, J., Lu, X., Wang, S.S., Tsao, Y., 2014. Speech enhancement using segmental nonnegative matrix factorization. ICASSP. IEEE, pp. 4483–4487.
Fu, S.-W., Tsao, Y., Lu, X., 2016. SNR-aware convolutional neural network modeling for speech enhancement. Interspeech. pp. 3768–3772.
Gao, T., Du, J., Xu, L., Liu, C., Dai, L.-R., Lee, C.-H., 2015. A unified speaker-dependent speech separation and enhancement system based on deep neural networks. ChinaSIP. pp. 687–691.
Gao, T., Du, J., Xu, Y., Liu, C., Dai, L.-R., Lee, C.-H., 2015. Improving deep neural network based speech enhancement in low SNR environments. International Conference on Latent Variable Analysis and Signal Separation. Springer, pp. 75–82.
Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., 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.
Hinton, G.E., 2002. Training products of experts by minimizing contrastive divergence. Neural Comput. 14 (8), 1771–1800.
Hinton, G.E., Osindero, S., Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18 (7), 1527–1554.
Hu, G., Wang, D.L., 2010. A tandem algorithm for pitch estimation and voiced speech segregation. Audio Speech Lang. Process. IEEE Trans. 18 (8), 2067–2079.
Hu, K., Wang, D.L., 2013. An unsupervised approach to cochannel speech separation. Audio SpeechLang. Process. IEEE Trans. 21 (1), 122–131.
Hu, Y., Loizou, P.C., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 16 (1), 229–238.
Huang, P.S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P., 2014. Deep learning for monaural speech separation. ICASSP. pp. 1562–1566.
Huang, P.S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P., 2015. Joint optimization of masks and deep recurrent neural networks for monaural source separation. Audio Speech Lang. Process. IEEE/ACM Trans. 23 (12), 2136–2147.
Hwang, I., Park, H.-M., Chang, J.-H., 2016. Ensemble of deep neural networks using acoustic environment classification for statistical model-based voice activity detection. Comput. Speech Lang. 38, 1–12.
Kamath, S., Loizou, P.C., 2002. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. ICASSP. 4. IEEE, pp. IV–4164.
Kim, M., Smaragdis, P., 2015. Adaptive denoising autoencoders: a fine-tuning scheme to learn from test mixtures. Latent Variable Analysis and Signal Separation. Springer, pp. 100–107.
Kristjansson, T., Attias, H., Hershey, J., 2004. Single microphone source separation using high resolution signal reconstruction. ICASSP. 2. IEEE, pp. ii–817.
Lim, J.S., Oppenheim, A.V., 1978. All-pole modeling of degraded speech. Acoustics Speech Signal Process. IEEE Trans. 26 (3), 197–210.
Loizou, P.C., 2013. Speech Enhancement: Theory and Practice. CRC press.
McAulay, R.J., Malpass, M.L., 1980. Speech enhancement using a soft-decision noise suppression filter. Acoustics Speech Signal Process. IEEE Trans. 28 (2), 137–145.
Mohammadiha, N., Smaragdis, P., Leijon, A., 2013. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. Audio Speech, Lang. Process. IEEE Trans. 21 (10), 2140–2151.
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society.
Roweis, S.T., 2000. One microphone source separation. NIPS. 13. pp. 793–799.
Roweis, S.T., 2003. Factorial models and refiltering for speech separation and denoising. nterspeech. pp. 1009–1012.
Schmidt, M.N., Olsson, R.K., 2006. Single-channel speech separation using sparse non-negative matrix factorization. Interspeech.
Shao, Y., Wang, D.L., 2006. Model-based sequential organization in cochannel speech. Audio Speech Lang. Process. IEEE Trans. 14 (1), 289–298.
Tu, Y.-H., Du, J., Xu, Y., Dai, L.-R., Lee, C.-H., 2014. Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers. ISCSLP. IEEE, pp. 250–254.
Varga, A., Steeneken, H.J., 1993. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12 (3), 247–251.
Vincent, E., Gribonval, R., Févotte, C., 2006. Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. 14 (4), 1462–1469.
Wang, D.L., Brown, G.J., 1999. Separation of speech from interfering sounds based on oscillatory correlation. Neural Netwo. IEEE Transa. 10 (3), 684–697.
Wang, D.L., Brown, G.J., 1999. Separation of speech from interfering sounds based on oscillatory correlation. Neural Netw. IEEE Trans. 10 (3), 684–697.
Wang, D.L., Brown, G.J., 2006. Computational auditory scene analysis: Principles, algorithms, and applications. Wiley-IEEE Press.
Wang, Q., Du, J., Bao, X., Wang, Z.-R., Dai, L.-R., Lee, C.-H., 2015. A universal VAD based on jointly trained deep neural networks. Interspeech. pp. 2282–2286.
Wang, Y., Narayanan, A., Wang, D., 2014. On training targets for supervised speech separation. Audio Speech Lang. Process. IEEE/ACM Trans. 22 (12), 1849–1858.
Wang, Y.X., Wang, D.L., 2013. Towards scaling up classification-based speech separation. Audio Speech Lang. Process. IEEE Trans. 21 (7), 1381–1390.
Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J.R., Schuller, B., 2015. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. International Conference on Latent Variable Analysis and Signal Separation. Springer, pp. 91–99.
Wu, M.Y., Wang, D.L., Brown, G.J., 2003. A multipitch tracking algorithm for noisy speech. Speech Audio Process. IEEE Trans. 11 (3), 229–241.
Xu, Y., Du, J., Dai, L.-R., Lee, C.-H., 2014. Dynamic noise aware training for speech enhancement based on deep neural networks. Interspeech. pp. 2670–2674.
Xu, Y., Du, J., Dai, L.-R., Lee, C.-H., 2014. An experimental study on speech enhancement based on deep neural networks. Signal Process. Lett. IEEE 21 (1), 65–68.
Xu, Y., Du, J., Dai, L.-R., Lee, C.-H., 2014. Global variance equalization for improving deep neural network based speech enhancement. Signal and Information Processing (ChinaSIP), 2014 IEEE China Summit & International Conference on. IEEE, pp. 71–75.
Xu, Y., Du, J., Dai, L.-R., Lee, C.-H., 2015. A regression approach to speech enhancement based on deep neural networks. Audio Speech Lang. Process. IEEE/ACM Trans. 23 (1), 7–19.
Xu, Y., Du, J., Huang, Z., Dai, L.-R., Lee, C.-H., 2015. Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement. Interspeech. pp. 1508–1512.
Zazo, R., Sainath, T.N., Simko, G., Parada, C., 2016. Feature learning with raw-waveform CLDNNs for voice activity detection. Interspeech. pp. 3668–3672.
Zhang, X.-L., Wang, D., 2016. Boosting contextual information for deep neural network based voice activity detection. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (2), 252–264.
Zhang, X.-L., Wang, D., 2016. A deep ensemble learning method for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) 24 (5), 967–977.
Zhang, X.-L., Wu, J., 2013. Deep belief networks based voice activity detection. IEEE Trans. Audio Speech Lang. Process. 21 (4), 697–710.
Zöhrer, M., Pernkopf, F., 2014. Single channel source separation with general stochastic networks. Interspeech. pp. 978–982.
Zöhrer, M., Pernkopf, F., 2015. Representation models in single channel source separation. ICASSP. IEEE, pp. 713–717.