Applied Acoustics 167 (2020) 107385

Contents lists available at ScienceDirect

Applied Acoustics

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

Deep neural networks based binary classification for single channel speaker independent multi-talker speech separation

Nasir Saleem *, Muhammad Irfan Khattak

Department of Electrical Engineering, University of Engineering and Technology (UET), Peshawar-25000, KPK, Pakistan

ARTICLE INFO

Article history:
Received 19 November 2019
Received in revised form 3 April 2020
Accepted 17 April 2020

Keywords:
Supervised speech separation
Binary classification
DNN
IBM
Variance equalization

ABSTRACT

Speech separation is an important task of separating a target speech from the mixture signals. Speaker-independent multi-talker speech separation is a challenging task due to unpredictability of the target and interfering speech in the target-interference mixtures. Conventionally, speech separation is used as a signal processing problem, but recently it is formulated as a deep learning problem and discriminative patterns of the speech are learned from the training data. In this paper, we consider the ideal binary mask (IBM) as a supervised binary classification training-target by using fully connected deep neural networks (DNN) for single-channel speaker-independent multi-talker speech separation. The train DNNs is used to estimate IBM training-target. The mean square error (MSE) is used as an objective cost function. Standard backpropagation and Monte-Carlo dropout regularization approaches are used for better generalization and overfitting during training. The estimated training-target is applied to the mixtures to obtain the separated target speech. We have addressed the over-smoothing problem and performed equalization of spectral variances to match the estimated and clean speech features. Our experimental results in various evaluating conditions report that the proposed method outperformed the competing methods in terms of the Perceptual Evaluation of Speech Quality (PESQ), Segmental SNR (SNRSeg), Short-time objective intelligibility (STOI), normalized Frequency weighted SNRSeg (*n*FwSNRSeg) and HIT-FA rates.

1. Introduction

Listening to the individuals in the crowded conditions frequently take place in presence of the interfering speakers. These conditions require the ability of an individual to separate a speech of interest from the mixture signals. Many proposed methods have demonstrated considerable performance gain on the separation task when prior knowledge of the speaker in mixtures is provided [1,2]. This, however, is still a challenging task when no prior knowledge about the speakers is given, and this particular problem is known as speaker-independent multi-talker speech separation. Humans are exceptionally adept at this task, however, this is a complicated task to model and emulate algorithmically. Even so, this challenge ought to be cracked to accomplish robust speech processing tasks. For instance, the performances of the existing automatic speech recognition (ASR) systems have achieved similar results to the humans in noise-free situations [3], these systems, however, are incapable to perform well in crowded conditions, and are less robust in the presence of interfering speakers. The task becomes even more challenging when the separation of all sources in a mixture is essential, such as meeting transcription. In multi-microphone conditions the beamforming algorithms can improve the performance [4,5]; however, the problem of the speech separation remains challenging when single-microphone is available.

Before emergence of the deep learning; approaches based on the statistical, clustering, and factorization have been used for the separation task. In statistical approaches, the target speech signals are modeled with probability distributions and the mixture signals are assumed to be statistically independent from target speech. Maximum likelihood estimation methods are usually applied based on some known statistical distributions of the target. In clustering approaches, the characteristics of the speakers are estimated from the observation and used to separate a target signal from the mixtures. Approaches such as computational auditory scene analysis (CASA) [6,7] fall into this set. Research in CASA as supervised speech separation has gained much attention [8-11]. IBM is a major computational goal of CASA [10], which indicates whether the target-speech dominates a T-F unit in time–frequency representation of mixture signals. With IBM as a major computational goal, the speech separation becomes a binary classification

* Corresponding author.
E-mail address: nasirsaleem@gu.edu.pk (N. Saleem).

https://doi.org/10.1016/j.apacoust.2020.107385
0003-682X/

approach, a basic framework of supervised learning. In [12], a classifier is trained to estimate IBM for binaural speech separation. A maximum *a posteriori* (MAP) classifier is trained based on the interaural-time and interaural-intensity differences binaural features to classify the T-F bins as speech-dominant or noise-dominant. This proposed system provided large gain in the speech intelligibility for matched training and testing situations. In [13], a Bayesian classifier is applied to estimate and eliminate the noise-dominant T-F bins for the robust ASR. Sub-band multilayer perceptrons (MLPs) are trained in [14] to classify T-F bins as speech-dominant or noise-dominant during grouping step of CASA-based speech separation. Gaussian mixture model (GMM) is applied to estimate the IBM in Mel-spectral domain [15]. At low SNRs and matched training–testing noise segments, these methods have been shown to improve the speech intelligibility for normal-hearing listeners. Factorization approaches, for example, NMF [16-18] formulated the separation as a matrix factorization problem in which the time–frequency representations of the mixtures are factorized into basis signals and activations. The activations learned for all basis signals are used to reconstruct target sources.

Recently, deep learning has made significant progress in source separation, specifically; deep neural networks have been successfully applied in the speech enhancement and speech separation [19-42] with considerably improved performance compared to the traditional approaches. The typical model for neural networks is to estimate time–frequency masks of the sources given the time–frequency representation of the mixtures (multiple speakers). Such model formulates the separation as a supervised regression problem and is important for supervised speech separation. Several types of masks and objective functions have been proposed, such as, phase-aware and binary masks that have been studied in [20,21]. DNNs have been shown remarkable success in the supervised tasks, such as, image classification, Automatic Speech Recognition (ASR) and speech enhancement [19-22].

1.1. Related work

A limited work relating supervised single-channel learning-based speaker-independent multi-talker speech separation exists in literature and [25-30] have addressed this task. To learn about the training labels, instant energy was used, which improved label permutation and facilitated unknown speaker separation [25]. Two-talker decoder is used to estimate and correlate the speaker and speech jointly. A penalty for speaker switching was estimated from the mixed-speech energy pattern change. The system achieved outperformed the state-of-the-art IBM by 2.8% absolute with fewer assumptions. Though, this approach worked well for mixtures of two-speakers, however, underperformed in many speakers' mixtures. DNN based Binary Mask estimation is proposed in [26] by considering audiovisual model for speaker independent separation. Hybrid DNN structure is exploited to leverage the complementary strengths of a stacked long short term memory (LSTM) and convolution LSTM network. The comparative simulation results in terms of speech quality and intelligibility demonstrate significant performance improvement for both speaker dependent and independent scenarios. An iterative DNN is proposed in [27] to perform the task of speaker-independent speech separation. Besides the commonly-used spectral features, the DNN also takes non-linearly wrapped spatial features as input, which is refined iteratively using parameter estimated from the DNN output via a feedback loop. DNN based method for attacking the single-channel multi-talker speech recognition problem is addressed in [28]. A Deep Attractor Network is proposed in [29], which creates attractor points in embedding spaces and attracts T-F bins related to the target speaker. The training process in this approach is similar to expectation–maximization (EM) principle. An attractor, a centroid of the speaker, in the embedding space was created to represent speakers. The time–frequency embeddings of speakers were then enforced to cluster around the attractor which is used to decide the time–frequency assignment of the speaker. The objective function for the network was standard signal reconstruction error which enables end-to-end maneuver during training and testing. Two deep learning approaches; deep clustering [34] and permutation invariant training (PIT) [30,35-37] have been proposed recently to resolve the multi-talker speech separation problem. PIT, however, solved the label permutation during training; but did not successfully solve the permutation during inference. Deep clustering is proposed in [38], deep Recurrent Neural Networks (DRNNs) are used to project the speech mixtures into embedding spaces, where T-F bins belonging to identical speakers form the clusters. Clustering algorithms are used then to classify the clusters in these embedding spaces. The T-F bins belonging to the identical clusters are grouped and a binary mask is estimated, which is used to separate the speakers from mixtures. The Utterance-level PIT (uPIT), a deep-learning approach, was proposed in [30] to solve speaker-independent multi-talker speech separation and extended the PIT approach with an utterance-level cost function. Recurrent neural networks (RNNs) were engaged to minimize the separation error at utterance-level. The uPIT was used for speaker independent multi-talker speech separation and denoising in [37]. Bi-directional LSTM RNNs were trained using uPIT. A similar approach using PIT was also proposed in [35]. A constrained uPIT (cuPIT) was proposed by computing a weighted MSE loss utilizing the dynamic information. The loss function ensured the temporal continuity of output frames with the identical speakers. The model was extended by adding an additional Grid LSTM layer to learn temporal and spectral patterns over input magnitude spectrum concurrently.

In this paper, the IBM is estimated by using DNN as a supervised binary classifier for the single-channel speaker-independent multi-talker speech separation. DNNs are trained which are based on the MSE cost function, standard backpropagation and Monte-Carlo dropout regularization. Hinton *et al.* [43,44] first suggested the dropout regularization concept to undermine overfitting in DNN training procedure. The dropout regularization discards inactive weights in training. Gal and Ghahramani [45] showed a theoretical relation between dropout regularization and estimate the inference in a Gaussian way and suggested the method of utilizing dropout during inference. Kendall *et al.* [46] showed that by enabling dropout during training and averaging the results of multiple stochastic forward passes, the testing show improvement and named as Monte-Carlo dropout regularization. In [47] also showed model uncertainty estimation from Monte-Carlo samples. Given these obvious improvements, we have used Monte-Carlo dropout regularization in our separation method. The over-smoothing problem is addressed and spectral variance equalization is performed to match the estimated and underlying clean speech to obtain good quality and intelligible speech. The main contributions of the proposed method are summarized as follows. (**i**): First, binary classification of the time–frequency units is achieved by using DNN structures for the single-channel speaker-independent multi-talker speech separation. (**ii**): Second, Monte-Carlo dropout regularization is used during training in order to achieve better generalization and to solve the over fitting of training data. (**iii**): Third, the over-smoothness problem is alleviated by adopting frequency-independent spectral variance equalization to match the input and output speech.

The remaining paper is organized as. The proposed speech separation method is presented in the Section 2. Experimental settings are presented in the Section 3. The results and analysis are presented in Section 4. Finally, the summary and conclusions are presented in Section 5.

2. Speech separation using deep neural networks

To explain the process of supervised deep learning to solve the problem of the speaker-independent multi-talker speech separation, we define the general problem of single-channel speech separation. The problem of speaker-independent multi-talker speech separation is defined by estimating all N speakers $z_1(t)$, $z_2(t)$, ..., $z_N(t)$, for a given the mixture $y(t)$ as:

$$y(t) = \sum_{j=1}^{N} z_j(t) \tag{1}$$

In time–frequency representation, the magnitude spectrums $Y(\omega,t)$ is equal to the sum of magnitude spectrums of all sources, and is defined as:

$$Y(\omega, t) = \sum_{j=1}^{N} Z_j(\omega, t) \tag{2}$$

The real-valued magnitude spectrums are used as input to the speech separation systems and time–frequency masks for target sources are estimated. The magnitude spectrum $|Y(\omega,t)|$ denotes the feature vectors $Y \in R^{1x\omega\tau}$, where ω is frequency and τ is time index. The magnitude spectrums and corresponding time–frequency masks for sources are vectors, representing as: $z_j \in R^{1x\omega\tau}$ and $m_j \in R^{1x\omega\tau}$. The estimated magnitude spectrums $\hat{z}_j \in R^{1x\omega\tau}$ are defined as:

$$\hat{z}_j = Y \otimes m_j, \; for \sum_{j=1}^{N} m_j \tag{3}$$

Where$\otimes$denotes the element-wise multiplication and $1 \in R^{1x\omega\tau}$ indicates an all-one vector. As shown in Fig. 1, the proposed method consists of the training phase and the separation phase. First, the clean and mixture speech are segmented, windowed and STFT is computed. Then complementary features are extracted from the utterances. These features are used to train the DNN. In the separation phase, the same method is used to extract the complementary features of the mixture speech. These features are fed to train DNN to estimate IBM. The estimated IBM is applied to magnitude spectrum of the mixture speech to obtain the target magnitude spectrum as:

$$\hat{Z}(\omega, t) = \hat{M}(\omega, t) \odot Y(\omega, t) \tag{4}$$

Where$\hat{M}(\omega,t)$and $\hat{Z}(\omega,t)$ denotes the estimated binary mask and spectral magnitude of the target speech. It is observed in the output estimated speech that the over-smoothing produces a muffled effect when compared to the clean version of the estimated speech. To mitigate this problem, frequency-independent spectral equalization is performed as a post-processing step in order to equalize spectral variances. Finally, during waveform reconstruction, the separated speech is reconstructed using phase of the mixture speech.

2.1. Binary classification based training target and training criterion

The mask estimation is a vital in the proposed method to estimate the magnitude spectrum of clean speech. We have trained DNN structures, which are formulated on binary classification training-target, comprised of IBM. IBM is a time–frequency representation constructs from clean speech and noise signals. For all time–frequency units, if the SNR ratio is larger than a local SNR criterion (LC), the time–frequency unit is called target-dominant and the resultant element is set to binary 1. If not, element is set to binary 0 and is called as the interference-dominant unit. IBM is defined as:

$$M(\omega, t) = \begin{cases} 1, & \text{if SNR}(\omega, t) \geqslant \text{LC(dB)} \\ 0, & \text{if SNR}(\omega, t) < \text{LC(dB)} \end{cases} \tag{5}$$

Where SNR shows signal-to-noise ratio, t, and ω denotes time and frequency index whereas LC denotes the local criterion. The LC is usually fixed at 0 dB. The mean square error (MSE) criterion is used as an objective cost function during training the deep network, defined as:

$$C_{MSE} = \frac{1}{TF} \sum_{t=1}^{T} \sum_{f=1}^{F} \left| \hat{M}(\omega, t) - M(\omega, t) \right|^2 \tag{6}$$

2.2. Proposed DNN framework

The proposed DNN framework is depicted in Fig. 1 which consists of the feature extraction, training, testing (separation), post-processing and waveform reconstruction, respectively. To reduce the complexity of the proposed separation method, we have adopted two fundamental approaches. (**a**) Same complementary features are utilized during training and testing phase. (**b**) Since same complementary features are used; the input and hidden layers have the same number of neurons. The neurons in output layer are different and fixed to 512 in the proposed method. The details of the simulation parameters are presented in Table 1.

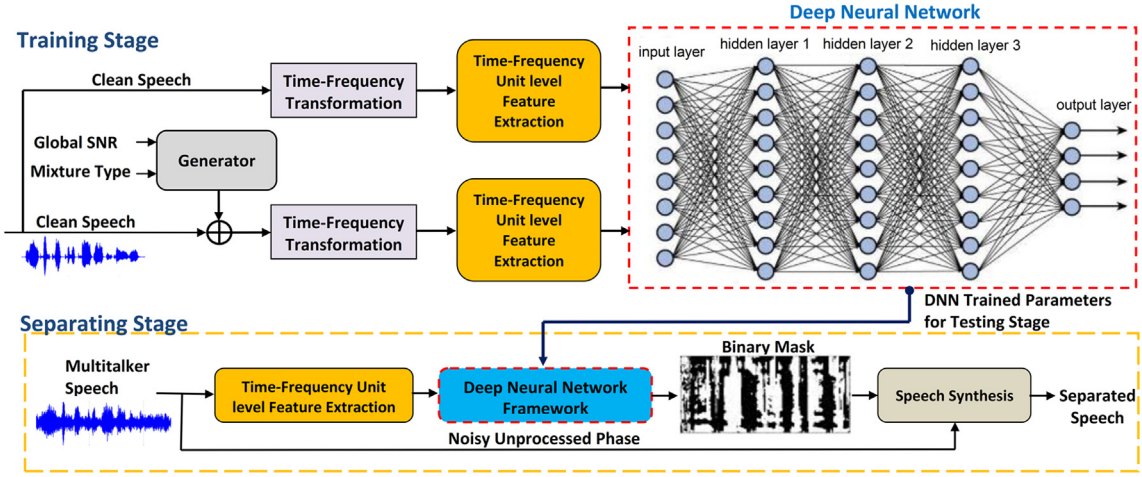

Fig. 1. Block Diagram speech separation method.

Table 1
Details of Simulation Parameters.

S. No.	Parameter Description	Parameter Value
1	Feature Set Dimension	256
2	Hidden Layers	3
3	Hidden Neurons	3072
4	Input Neurons	1024
5	Output Neurons	512
6	Momentum Term	0.4 and 0.8
7	Number of Epochs	100
8	Frame Length	20 msec
9	Frame Shift	10 msec
10	Scaling Factor	0.0010

2.2.1. Feature extraction

A set of complementary acoustic features is extracted from the input speech at frame level. Acoustic features are extracted with frame length set to 20 ms and the frame shift is set to 10 ms. The features set includes: Multi-resolution cochleagram (MRCG) [39], 13-dimension relative spectral transformed perceptual linear prediction coefficients (RASTA-PLP), 31-dimension Mel-frequency cepstral coefficients (MFCC), 64-dimension Gammatone filter energies (GFE) and 51-dimension amplitude modulation spectrogram (AMS). All the complementary features are concatenated with corresponding delta and double delta features. Finally, a set of 256-dimensional complementary features are obtained which is used to train and test the DNN. All the feature vectors are normalized to the zero-mean and unit-variance.

2.2.2. DNN architecture

DNN is a selective learning machine and has shown to perform well in the source separation [37,38]. The DNN training framework contains five layers; input layer, three hidden layers, and output layer, respectively, shown in Fig. 2. The size of the input layer is 1024 neurons, i.e., 256*4 = 1024, including 256 dimensional features and 4 frames window. The hidden layers contain 1024 neurons and used the rectified linear unit (ReLU) activation function. To select the activation function in the hidden layers based on the performance for the proposed method, preliminary experiments are conducted. Fig. 3 shows PESQ and STOI scores for DNN using two activation functions in the hidden layers, namely ReLU and sigmoid (Sig) with different global SNRs. We note that consistent high PESQ and STOI values are obtained for the ReLU activation function. As a result, the ReLU activation function is used in hidden layers. The output layer consists of 512 neurons and the sigmoid activation function. The standard backpropagation and Monte-Carlo (MC) dropout regularization [45,47] are used to train DNN. The adaptive gradient descent algorithm [49] with a momentum parameter β is used to optimize DNN. The batch size is fixed to 128. The scaling factor for the adaptive stochastic gradient descent is set to 0.0010 and the learning rate is minimized linearly from 0.06 to 0.002.Total of 100 epochs is used during the process. For the first few epochs, the β is fixed at 0.4 and the rate is increased to 0.8 for remaining epochs. The MSE cost function is used in training. The cost optimization curves at epochs for DNNs are demonstrated in Fig. 4.

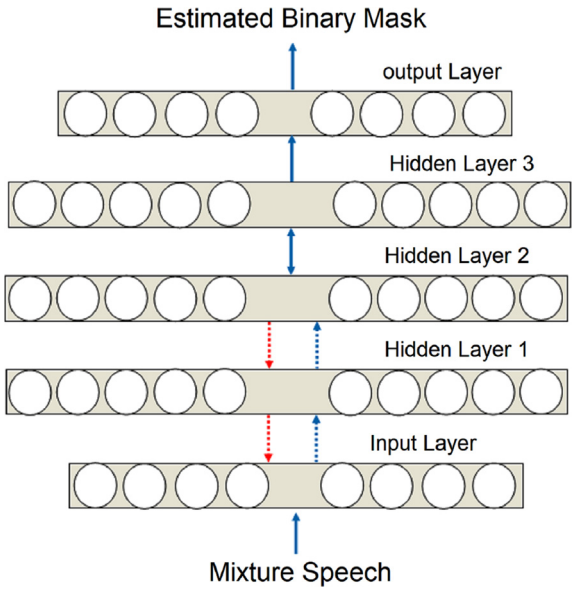

Fig. 2. Deep Neural Network Training.

2.2.3. DNN speech separation

During separation phase, the complementary features of the utterances mixed with different mixtures are used as the testing features and the IBM is used as a training-target. The IBM is the computational goal of computational auditory scene analysis (CASA) and has achieved good results [10]. The IBM is used as a binary classification approach and DNN is used to predict the labels of time–frequency units. The estimate of the target magnitude spectrum is achieved by multiplying the estimated mask with the mixture magnitude spectrums. The time-domain speech is recovered by computing an inverse STFT of the estimated magnitude spectrum $\hat{Z}_j$ using phase of the mixture.

2.2.4. Post processing

The over-smoothness in output speech generates a muffled effect. To mitigate this problem, frequency-independent spectral variance equalization is adopted as a post-processing step to match the features of output and clean speech for improving speech quality and intelligibility. According to study of Xu et al. [48], frequency-independent spectral variance equalization performs better that frequency-dependent spectral variance equalization. Also, it is verified that such variance equalization can greatly improve the subjective scores [50]. In frequency-independent spectral variance equalization, the variances of the estimated speech V_{EST} and the clean speech V_{CLEAN} are defined as:

$$V_{EST}(d) = \frac{1}{N}\sum_{n=1}^{N}\left(\hat{Z}(d) - \frac{1}{N}\sum_{n=1}^{N}\hat{Z}(d)\right)^2 \tag{7}$$

$$V_{CLEAN}(d) = \frac{1}{N}\sum_{n=1}^{N}\left(Z(d) - \frac{1}{N}\sum_{n=1}^{N}Z(d)\right)^2 \tag{8}$$

Where $\hat{Z}(d)$ is the d^{th} element of DNN output vectors at n^{th} frame and N is the total number of frames in training set. The variance of the estimated and clean speech in various frequency bands is illustrated in Fig. 5. The variances of the estimated speech utterance are smaller than clean speech; this indicates that the estimated speech spectra are smoothed. Furthermore, in low SNR conditions, the over-smoothing phenomenon becomes worst and the formant peaks are suppressed. The over-smoothing in the high-frequency bands leads to a muffling speech. Fig. 6 shows spectrograms of speech degraded by babble noise at 0 dB. A substantial over-smoothing can be noticed. The formant peaks between 2000 and 4000 Hz are suppressed. Based on (7) and (8), the equalization factor μ(d) is used to control the over-smoothing problem, defined as:

$$\mu(d) = \sqrt{\frac{V_{EST}(d)}{V_{CLEAN}(d)}} \tag{9}$$

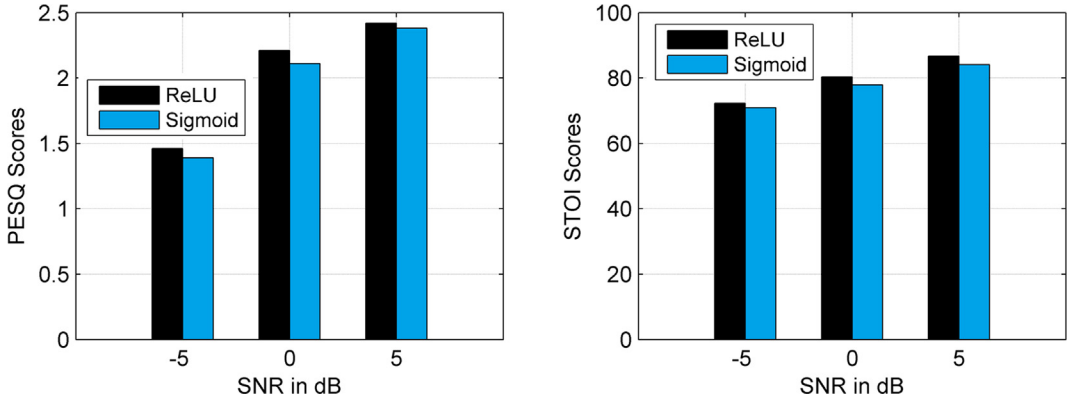

Fig. 3. PESQ and STOI scores using ReLU and Sigmoid activation functions.

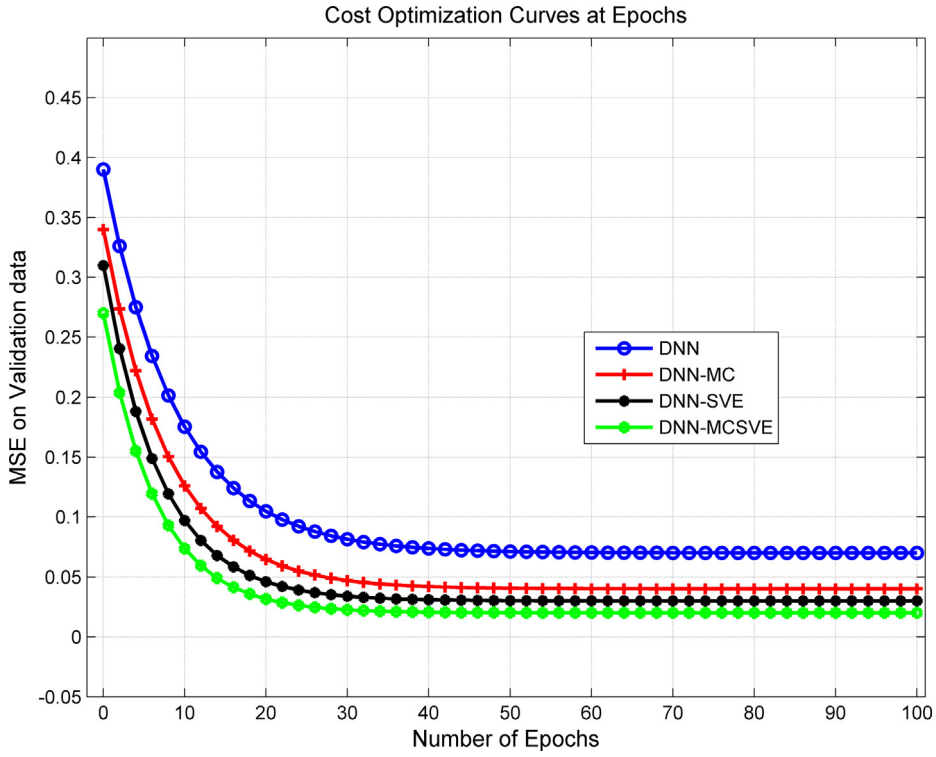

Fig. 4. MSE optimization curves for different DNN structures.

The equalization factor updates automatically during training process. The DNN outputs are transformed by using following expression given as:

$$\bar{Z}(\mathrm{d}) = \mu(\mathrm{d})\hat{Z}(\mathrm{d}) \otimes \sigma(\mathrm{d}) + m(\mathrm{d}) \tag{10}$$

Where σ(d) and m(d) denotes variance and mean of speech. The equalization factor elevates the variances of the speech according to (10). In the variance equalization post-filtering, the multiplication of the equalization factor to the network output features can be viewed as enforcing an exponential factor in linear spectral magnitude domain. Such post-filtering approach, can enlarge or diminish the variance of the spectral trajectories depending on the value of μ(d). Mostly μ(d) is bigger than 1 and the lack of dynamics in output features is alleviated. With variance equalization post filtering approach, the sharp formant peaks of the reconstructed speech are achieved and residual noise is minimized. Hence, it improves the quality and intelligibility of the separated speech.

3. Methods and materials

3.1. Datasets

We have used 720 IEEE speech utterances [51] in training and the testing set consists of 300 speech utterances from unknown speakers of both genders. The WSJ0-2mix dataset has also been used which was introduced in [52] and was derived from the WSJ0 corpus [53]. The 2000 speech utterances from the WSJ0 training set are selected in the experiments. Mixtures of two-talker, three-talker and four-talker are used in training and testing procedures. The sample spectrograms of the speech mixtures are demonstrated in Fig. 7. The duration of each mixture is about 6 min. To create the training sets, the first 3 min of each mixture is used and mixed with the training utterances at −5dB, 0 dB, and 5 dB SNR. The testing mixtures are created by mixing the last 3 min of mixtures. Two training and testing sets have been used in the experiments. First, a training set of 720 IEEE utterances × 3

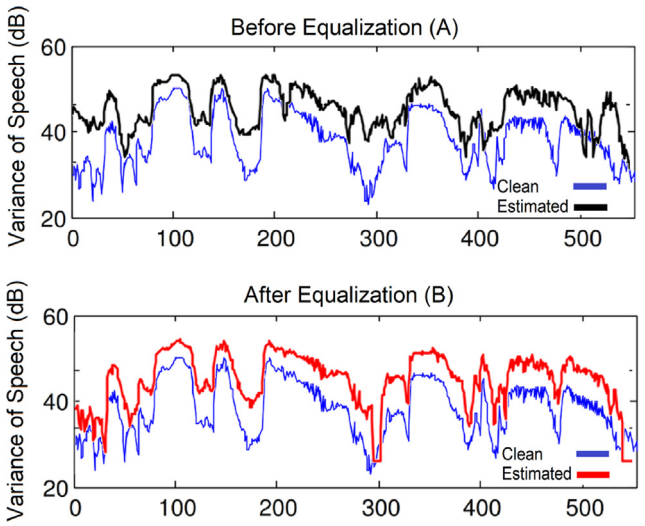

Fig. 5. Variance of the speech utterances before and after spectral equalization.

mixtures × 3 SNR and a testing set of 300 utterances × 3 mixtures × 3 SNR are obtained, respectively. The other training set of 2000 WSJ0-2mix utterances × 3 mixtures × 3 SNR and a testing set of 500 utterance × 3 mixtures × 3 SNR are obtained.

3.2. Evaluation metrics and competing methods

To measure the perceived speech intelligibility, we have used two objective metrics; Short-time objective intelligibility (STOI) [54] and normalized frequency weighted segmental SNR (*n*FwSNRSeg) [55]. Similarly, to evaluate the quality of the separated speech, we have used two objective metrics; Perceptual evaluation of speech quality (PESQ) [56] and segmental SNR (SNRSeg) [57]. The evaluation metrics with their mathematical expressions are given in Table 2. Two competing methods are considered in experiments for performance comparison. First, the IBM estimation based on CASA in [58], and second, the IBM estimation based on the DNN in [59]. The proposed method is denoted as $DNN_{MC\text{-}SVE}$. To measure accuracy of the supervised binary classification, average hit (HIT) and false-alarm (FA) rates are computed for three scenarios (two-talker, three-talker and four-talker) included in experiments. Each scenario comprised of 200 speech utterances. A total of 1000 speech utterances from IEEE and WSJ0-2mix correspond to three scenarios. If time–frequency units are present, we can decide the time–frequency units and these outcomes are called HIT and FA alarms. If the time–frequency unit is correctly decided, it has been categorized as HIT and FA in opposite case.

4. Results and analysis

We first examined the separation performance in terms of the speech quality by using PESQ for two-talker, three-talker and four-talker mixtures at −5dB, 0 dB and 5 dB SNRs. Table 3 provides the results of $DNN_{MC\text{-}SVE}$ in terms of the PESQ for IEEE and WSJ0-2mix datasets. All the PESQ scores are averaged over 200 utterances from IEEE dataset and 500 utterances from WSJ0-2mix dataset. The results demonstrate that $DNN_{MC\text{-}SVE}$ outperformed the CASA and DNN based separation methods at all SNRs consistently. For instance, the predicted PESQ scores with two-talkers mixture are improved from 1.28 with the mixture to 2.44 at −5dB SNR ($\Delta PESQ_{two\text{-}talkers} = 1.16$) with $DNN_{MC\text{-}SVE}$. Similarly, the predicted PESQ scores with three-talker mixture are improved from 1.87 with DNN to 2.32 at −5dB SNR ($\Delta PESQ_{three\text{-}talkers} = 0.45$) with $DNN_{MC\text{-}SVE}$. Whereas, in the case of a four-talker mixture, predicted PESQ scores are improved from 1.99 with CASA to 2.26 at −5dB

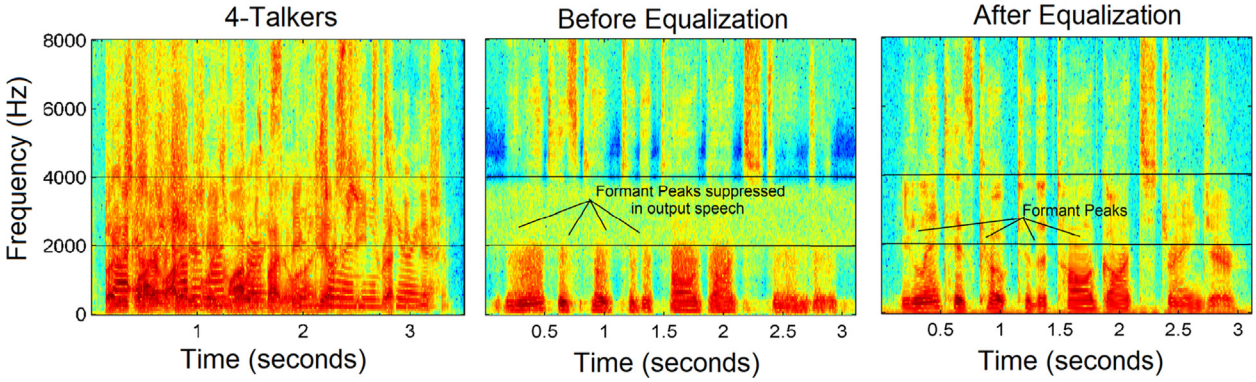

Fig. 6. Spectrogram analysis of over-smoothing before and after spectral equalization.

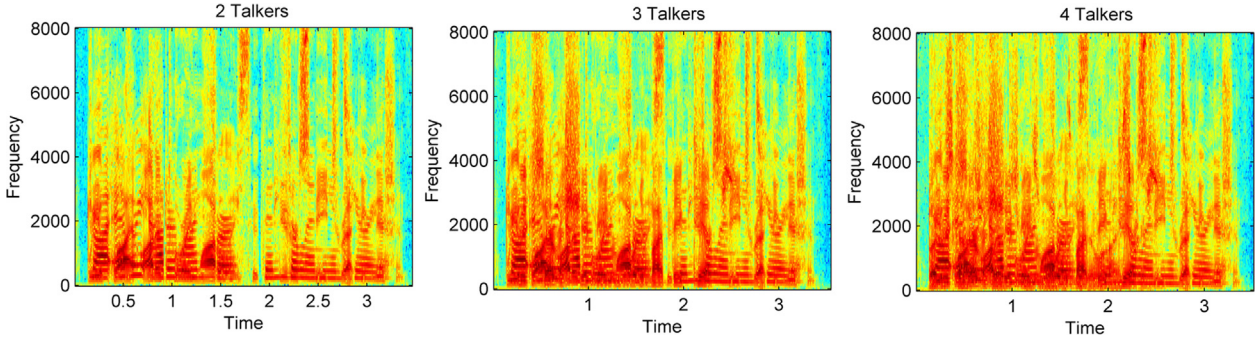

Fig. 7. Spectrograms of mixtures, 2-talker, 3-talker and 4-taker.

Table 2
Speech Enhancement Evaluations Measures.

S. No	Evaluation Metric	Mathematical Expression
1	PESQ : Perceptual Evaluation of Speech Quality	$PESQ = \alpha_0 - \alpha_1.A_{Asym} - \alpha_2 B_{Dsym},\ \alpha_0 = 4.5,\ \alpha_1 = -0.1,\ \alpha_2 = -0.039$
2	SNRSeg: Segmental Signal to Noise Ratio	$SNRSeg(m,\omega_m) = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\left(\frac{\lvert S(m,\omega_m)\rvert^2}{\lvert\lvert S(m,\omega_m)\rvert-\lvert S_{EST}(m,\omega_m)\rvert\rvert^2}\right)$
3	FwSNRSeg: Frequency Weighted Segmental Signal to Noise Ratio	$FWSNR_{SEG}(m,\omega_m) = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{k} B_j \log_{10}\left[\frac{F^2(m,j)}{\hat{F}(m,j)-F(m,j)}\right]}{\sum_{j=1}^{k} B_j}$
4	STOI: Short Time Objective Intelligibility	$f(STOI) = \frac{100}{1+\exp(\sigma STOI+\delta)}$

Table 3
Average PESQ analysis.

Processing Methods	2-Talkers				3-Talkers				4-Talkers			
	−5 dB	0 dB	5 dB	Avg	−5 dB	0 dB	5 dB	Avg	−5 dB	0 dB	5 dB	Avg
IEEE Database												
Mixture	1.28	1.81	2.15	1.74	1.17	1.63	2.01	1.61	1.12	1.51	1.91	1.52
CASA	2.17	2.49	2.76	2.47	2.01	2.41	2.70	2.37	1.99	2.33	2.62	2.31
DNN	1.97	2.23	2.49	2.23	1.87	2.12	2.42	2.13	1.77	2.09	2.38	2.08
$DNN_{MC\text{-}SVE}$	2.44	2.68	2.92	2.68	2.32	2.55	2.87	2.57	2.26	2.44	2.75	2.49
WJS0-2mix Database												
Mixture	1.33	1.86	2.20	1.79	1.2	1.66	2.04	1.63	1.14	1.55	1.92	1.53
CASA	2.26	2.58	2.85	2.56	2.08	2.48	2.77	2.44	2.04	2.38	2.67	2.36
DNN	2.10	2.36	2.62	2.36	1.96	2.21	2.51	2.22	1.84	2.16	2.45	2.15
$DNN_{MC\text{-}SVE}$	2.60	2.84	3.08	2.84	2.44	2.67	2.99	2.7	2.35	2.53	2.84	2.57

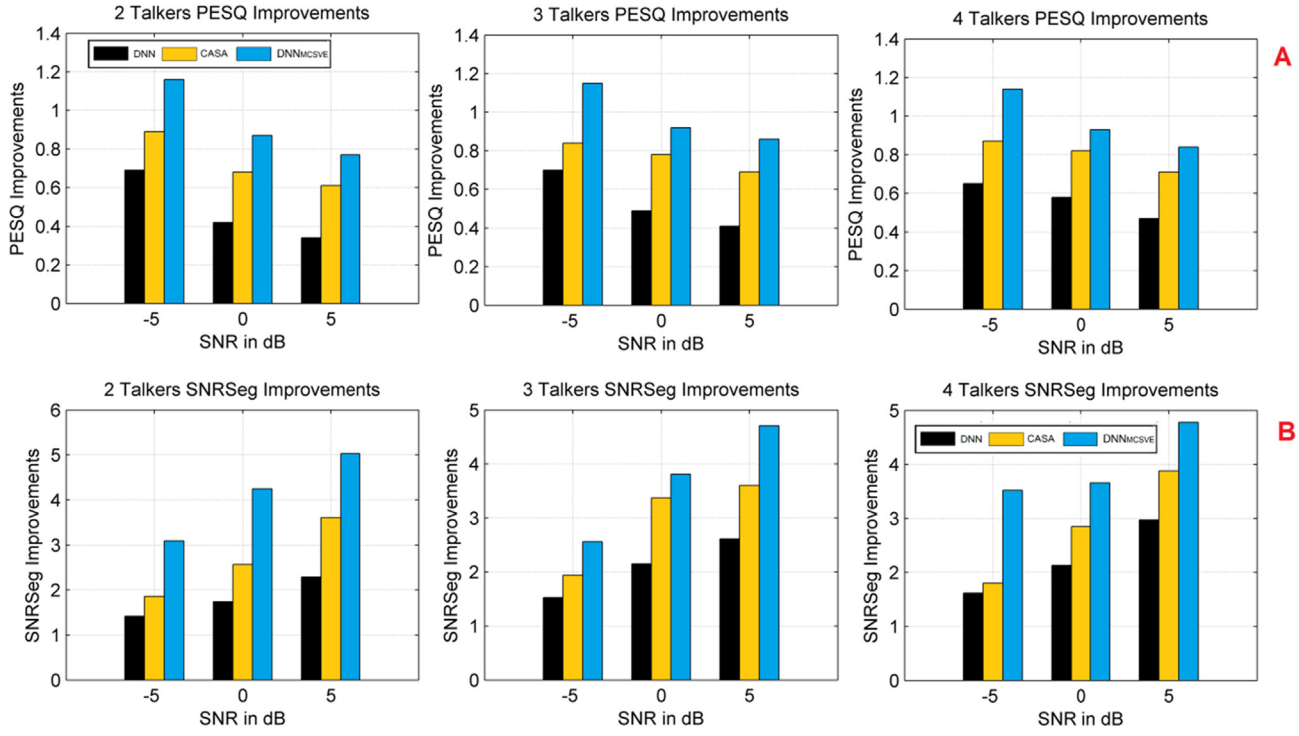

Fig. 8. PESQ and SNRSeg improvement analysis.

SNR ($\Delta PESQ_{four\text{-}talkers} = 0.27$) with $DNN_{MC\text{-}SVE}$. The overall PESQ improvements in all mixture types for $DNN_{MC\text{-}SVE}$ are depicted in Fig. 8 (A). Secondly, the separation performance is evaluated in terms of the SNRSeg for two-talker, three-talker and four-talker mixtures at −5dB, 0 dB and 5 dB SNRs. Table 4 provides results of $DNN_{MC\text{-}SVE}$ in terms of the SNRSeg for IEEE and WSJ0-2mix datasets, respectively. The predicted SNRSeg scores with two-talker mixture are improved from 2.46 with the mixture to 6.71 at 0 dB SNR ($\Delta SNRSeg_{two\text{-}talkers} = 4.25$) with $DNN_{MC\text{-}SVE}$. Similarly, the predicted SNRSeg scores with three-talker mixture are improved from 4.19 with DNN to 5.85 at 0 dB SNR ($\Delta SNRSeg_{three\text{-}talkers} = 1.66$) with $DNN_{MC\text{-}SVE}$. For the four-talker mixture, the predicted SNRSeg scores are improved from 4.81 with CASA to 5.62 at 0 dB SNR ($\Delta SNRSeg_{four\text{-}talkers} = 0.81$) with $DNN_{MC\text{-}SVE}$. The overall SNRSeg

Table 4
Average SNRSeg analysis.

Processing Methods	2-Talkers				3-Talkers				4-Talkers			
	−5 dB	0 dB	5 dB	Avg	−5 dB	0 dB	5 dB	Avg	−5 dB	0 dB	5 dB	Avg
IEEE Database												
Mixture	1.96	2.46	3.30	2.57	1.85	2.04	2.72	2.21	1.72	1.96	2.25	1.97
CASA	3.82	5.03	6.91	5.25	3.79	5.41	6.32	5.17	3.52	4.81	6.13	4.82
DNN	3.24	4.23	4.59	4.02	3.18	4.19	4.33	3.92	3.04	4.09	4.22	3.78
$DNN_{MC\text{-}SVE}$	5.05	6.71	8.33	6.69	4.41	5.85	7.43	5.89	4.01	5.62	6.89	5.51
WJS0-2mix Database												
Mixture	2.01	2.51	3.35	2.62	1.88	2.07	2.75	1.63	3.06	1.55	1.92	1.53
CASA	3.91	5.12	7.00	5.34	3.86	5.48	6.39	2.44	4.06	4.86	6.18	2.36
DNN	3.37	4.36	4.72	4.15	3.27	4.28	4.42	2.22	3.11	4.16	4.29	2.15
$DNN_{MC\text{-}SVE}$	5.21	6.87	8.49	6.85	4.53	5.97	7.55	2.7	4.10	5.71	6.98	2.57

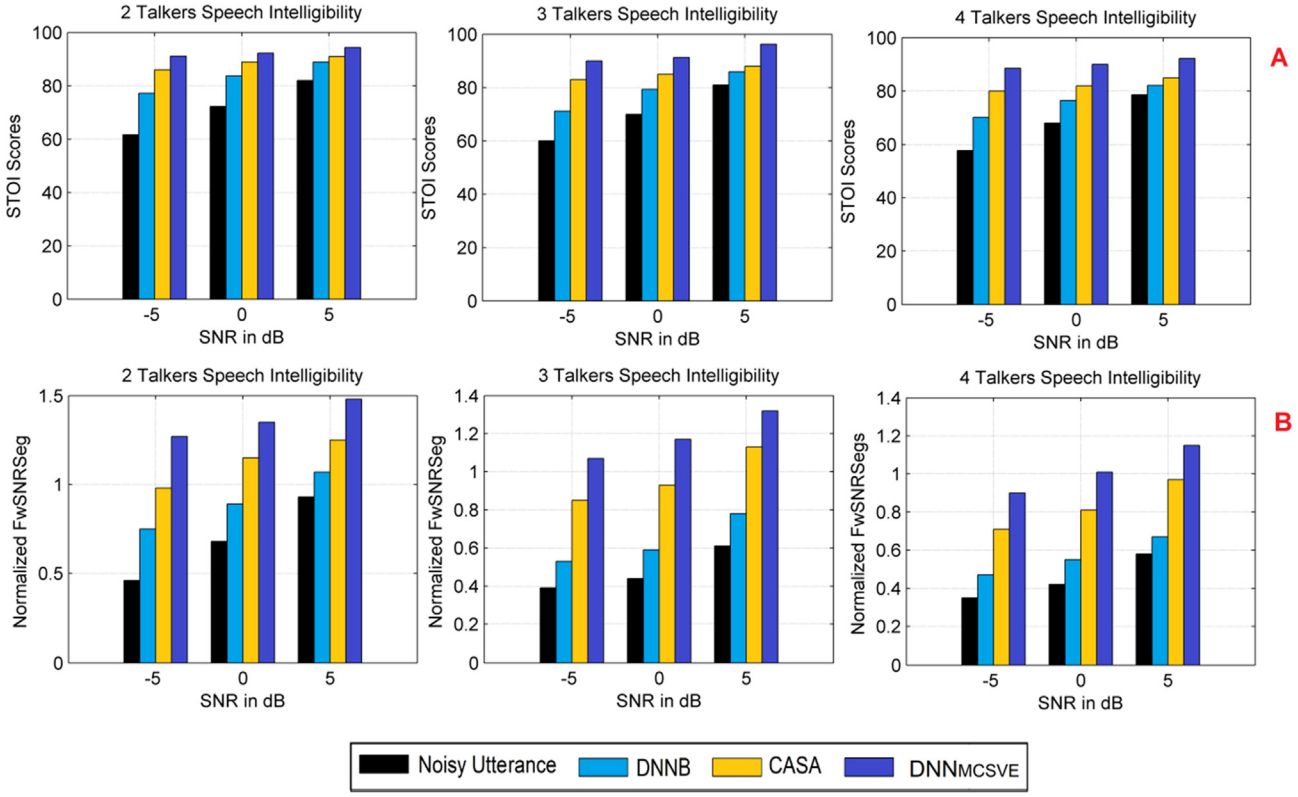

Fig. 9. Objective Speech Intelligibility scores using STOI and FwSNRSeg Measure.

Table 5
HIT and FALSE Rates (in %age).

Model	Metric	2-Talkers			3-Talkers			4-Talkers		
		−5dB	0 dB	5 dB	−5dB	0 dB	5 dB	−5dB	0 dB	5 dB
IEEE Database										
DNN	HIT	80.97	80.33	80.23	80.50	81.80	78.77	74.44	81.09	78.88
	FA	18.53	19.76	19.88	17.89	16.67	15.14	13.28	13.52	16.74
$DNN_{MC\text{-}SVE}$	HIT	84.61	82.42	84.59	83.71	84.29	85.20	79.36	83.91	85.62
	FA	15.85	12.12	12.50	17.36	10.77	10.96	10.49	12.24	13.66
WJS0-2mix Database										
DNN	HIT	81.05	80.41	80.31	80.55	81.87	78.86	74.48	81.14	78.93
	FA	18.61	19.84	19.96	17.94	16.74	15.23	13.33	13.58	16.79
$DNN_{MC\text{-}SVE}$	HIT	84.70	82.51	84.68	83.76	84.36	85.29	79.41	83.97	85.68
	FA	15.87	12.21	12.59	17.41	10.84	11.05	10.53	12.3	13.72

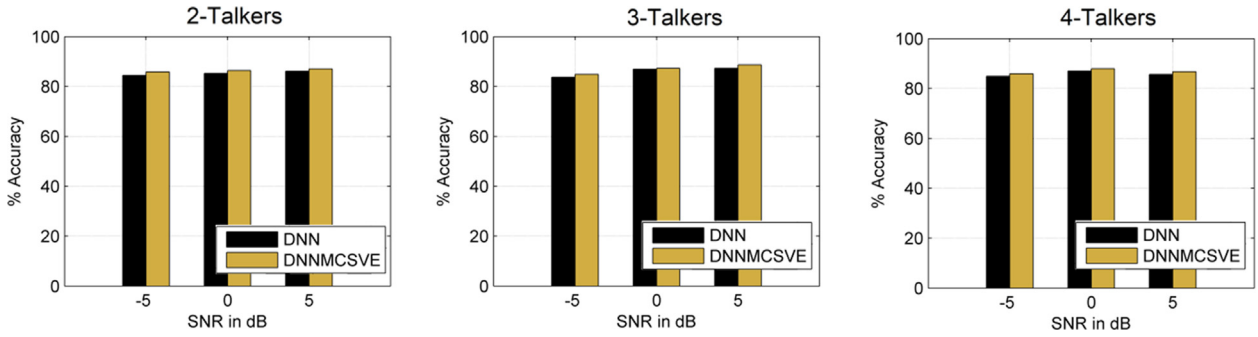

Fig. 10. Classification Accuracy of Time-Frequency units.

Table 6
HIT and FALSE rates (in %) for $DNN_{MC\text{-}SVE}$ with and without delta features.

Mixture Type	Metric	Features				Features + Delta			
		−5 dB	0 dB	5 dB	Avg	−5 dB	0 dB	5 dB	Avg
2-Talkers	HIT	80.97	80.33	80.23	80.51	84.61	82.42	84.59	83.87
	FA	18.53	19.76	19.88	19.39	15.85	12.12	12.50	13.49
3-Talkers	HIT	80.50	81.80	78.77	80.35	83.71	84.29	85.20	84.40
	FA	17.89	16.67	15.14	16.56	17.36	10.77	10.96	13.03
4-Talkers	HIT	74.44	81.09	78.88	78.13	79.36	83.91	85.62	82.92
	FA	13.28	13.52	16.74	14.51	10.49	12.24	13.66	12.13

Table 7
PESQ analysis. DNN_{MC}: With MC Dropout Regularization, DNN_{SVE}: With Spectral Variance Equalization, and $DNN_{MC\text{-}SVE}$: MC Dropout Regularization and Spectral Variance Equalization.

Mixture Type	DNN			DNN_{MC}			DNN_{SVE}			$DNN_{MC\text{-}SVE}$		
	−5 dB	0 dB	5 dB	−5dB	0 dB	5 dB	−5 dB	0 dB	5 dB	−5 dB	0 dB	5 dB
IEEE Database												
2-Talkers	1.97	2.23	2.49	2.23	2.51	2.78	2.31	2.60	2.86	2.42	2.67	2.91
3-Talkers	1.87	2.12	2.42	2.11	2.42	2.66	2.21	2.48	2.79	2.31	2.53	2.85
4-Talkers	1.77	2.09	2.38	2.01	2.34	2.59	2.12	2.38	2.68	2.23	2.43	2.73
Average	1.87	2.15	2.43	2.11	2.42	2.67	2.21	2.48	2.77	2.32	2.54	2.83
WJS0-2mix Database												
2-Talkers	2.04	2.34	2.62	2.3	2.62	2.91	2.38	2.71	2.99	2.49	2.78	3.04
3-Talkers	1.93	2.21	2.42	2.17	2.51	2.77	2.27	2.57	2.9	2.37	2.62	2.96
4-Talkers	1.77	2.09	2.44	2.05	2.4	2.67	2.16	2.44	2.76	2.27	2.49	2.91
Average	1.91	2.21	2.49	2.17	2.51	2.78	2.27	2.57	2.88	2.37	2.63	2.97

Table 8
HIT and FALSE Rates in %. DNN_{MC}: With MC dropout regularization, DNN_{SVE}: with spectral variance equalization, and $DNN_{MC\text{-}SVE}$: MC dropout regularization and spectral variance equalization.

Mixture Type	Metric	DNN_{MC}			DNN_{SVE}			$DNN_{MC\text{-}SVE}$		
		−5 dB	0 dB	5 dB	−5 dB	0 dB	5 dB	−5 dB	0 dB	5 dB
IEEE Database										
2-Talkers	HIT	83.31	84.01	86.53	84.61	85.42	87.59	85.53	86.77	88.31
	FA	16.81	12.16	13.69	15.85	11.12	10.50	13.76	10.33	09.76
3-Talkers	HIT	82.12	83.99	84.07	83.71	84.29	85.20	84.83	85.44	86.63
	FA	12.56	15.65	11.14	11.36	10.77	10.16	10.43	09.88	09.02
4-Talkers	HIT	81.36	83.91	85.62	83.39	85.24	87.36	84.21	86.56	88.42
	FA	12.49	11.94	11.06	11.36	11.03	10.16	10.75	09.23	09.01
WJS0-2mix Database										
2-Talkers	HIT	83.36	84.07	86.6	84.66	85.48	87.66	85.58	86.83	88.38
	FA	16.86	12.22	13.76	15.9	11.18	10.57	13.81	10.39	09.83
3-Talkers	HIT	82.15	84.03	84.12	83.74	84.33	85.25	84.86	85.48	86.68
	FA	12.59	15.69	11.19	11.39	10.81	10.21	10.46	9.92	09.07
4-Talkers	HIT	81.38	83.94	85.655	83.41	85.27	87.395	84.23	86.59	88.45
	FA	12.51	11.97	11.095	11.38	11.06	10.20	10.77	9.26	9.045

improvements in all mixture types for $DNN_{MC\text{-}SVE}$ are depicted in Fig. 8 (B). The results in Table 3 and 4 and Fig. 8 clearly show the superiority of $DNN_{MC\text{-}SVE}$ method over DNN and provide a high-quality speech with less residual noise.

We examined $DNN_{MC\text{-}SVE}$ in terms of the STOI and FwSNRSeg. STOI and FwSNRSeg provide the measures of the overall speech intelligibility of the separated speech utterance. Higher STOI and FwSNRSeg scores imply better performance. Fig. 9 shows the STOI and FwSNRSeg scores obtained with CASA, DNN and $DNN_{MC\text{-}SVE}$ respectively. All the STOI and FwSNRSeg outcomes are averaged over 200 utterances. Fig. 9(A) shows that $DNN_{MC\text{-}SVE}$ outperformed the CASA and DNN consistently in all conditions in terms of STOI. The only exceptions are: two-talker mixture at 0 dB and 5 dB SNRs, where we deem that CASA, DNN and $DNN_{MC\text{-}SVE}$ performed very well. However, $DNN_{MC\text{-}SVE}$ surpasses CASA and DNN at all SNRs. The $DNN_{MC\text{-}SVE}$ provides high intelligibility scores (STOI $\geq$ 85%) for all mixtures at SNR $\geq$ -5dB. The overall best prediction score in terms of STOI is 96.04%. For example, the average predicted STOI rates with three-talker mixture are improved from 74.22% with DNN and 82% with CASA to 89.9% with $DNN_{MC\text{-}SVE}$ at −5dB SNR. Similarly, the average predicted rates with four-talker mixture are improved from 70.02% with DNN and 80% with CASA to 86% with $DNN_{MC\text{-}SVE}$ at −5dB SNR. Fig. 9(B) shows the normalized FwSNRSeg (*n*FwSNRSeg) scores obtained with CASA, DNN and $DNN_{MC\text{-}SVE}$ respectively. The $DNN_{MC\text{-}SVE}$ provides high intelligibility scores (*n*FwSNRSeg $\geq$ 1.00) for all mixtures at SNR $\geq$ -5dB. For example, the normalized predicted *n*FwSNRSeg scores with three-talker mixture are improved from 0.7 with DNN and 0.88 with CASA to 1.04 with $DNN_{MC\text{-}SVE}$ at −5dB SNR. Similarly, the predicted normalized FwSNRSeg scores with four-talker mixture are improved from 0.44with DNN and 0.67 with CASA to 0.9 with $DNN_{MC\text{-}SVE}$ at −5dB SNR. The predicted STOI and *n*FwSNRSeg results confirmed the superiority of $DNN_{MC\text{-}SVE}$.

To measure the accuracy of the supervised binary classification, the average hit (HIT) and false-alarm (FA) rates are computed for three scenarios included in experiments. Each scenario comprised of 200 speech utterances. A total of 600 speech utterances correspond to three scenarios. The average HIT and FA rates are quantified by comparing the $DNN_{MC\text{-}SVE}$ based estimated binary mask against the oracle mask (IBM). Table 5 shows the results obtained using DNN and $DNN_{MC\text{-}SVE}$ models in the different talker conditions for IEEE and WSJ0-2mix datasets. High HIT rates (lowest with 4-talker at −5dB, 79.36%) and low FA rates (highest with 3-talker at −5dB, 17.36%) are obtained with $DNN_{MC\text{-}SVE}$ models. The average hit rate obtained with DNN models is about 5.2% lower than that of $DNN_{MC\text{-}SVE}$ models. The low average FA rates are required to achieve high speech intelligibility. According to the study [47], FA < 20% assumes high hit rates and ensures high speech intelligibility. The average FA rate in $DNN_{MC\text{-}SVE}$ is 12.88% which is significantly lower than DNN (16.82%). The percentage accuracy of the DNN and $DNN_{MC\text{-}SVE}$ is demonstrated in Fig. 10. HIT and FA rates obtained with and without delta features are used to quantify the gain in classification accuracy. Table 6 shows the evaluation of the HIT and FA rates obtained with and without delta features. As can be observed, the delta features improved the HIT rate significantly (07% in some cases) without increasing the FA rate.

In Table 7, we compare the PESQ scores among three DNN configurations, denoted as; DNN_{MC}: with MC Dropout Regularization, DNN_{SVE}: with Spectral Variance Equalization, and $DNN_{MC\text{-}SVE}$: with MC Dropout Regularization and Spectral Variance Equalization. A total of 400 speech utterances and three mixture types are used to train the DNNs. All DNN configurations have 3 hidden layers and 1024 neurons in all hidden layers. Compared with baseline DNN [47], DNN_{MC} showed improved performance, with PESQ score improved from 1.97 to 2.23 for 2-talker mixture at low SNR (-5dB). Similarly, DNN_{SVE} achieved significant improvements over baseline

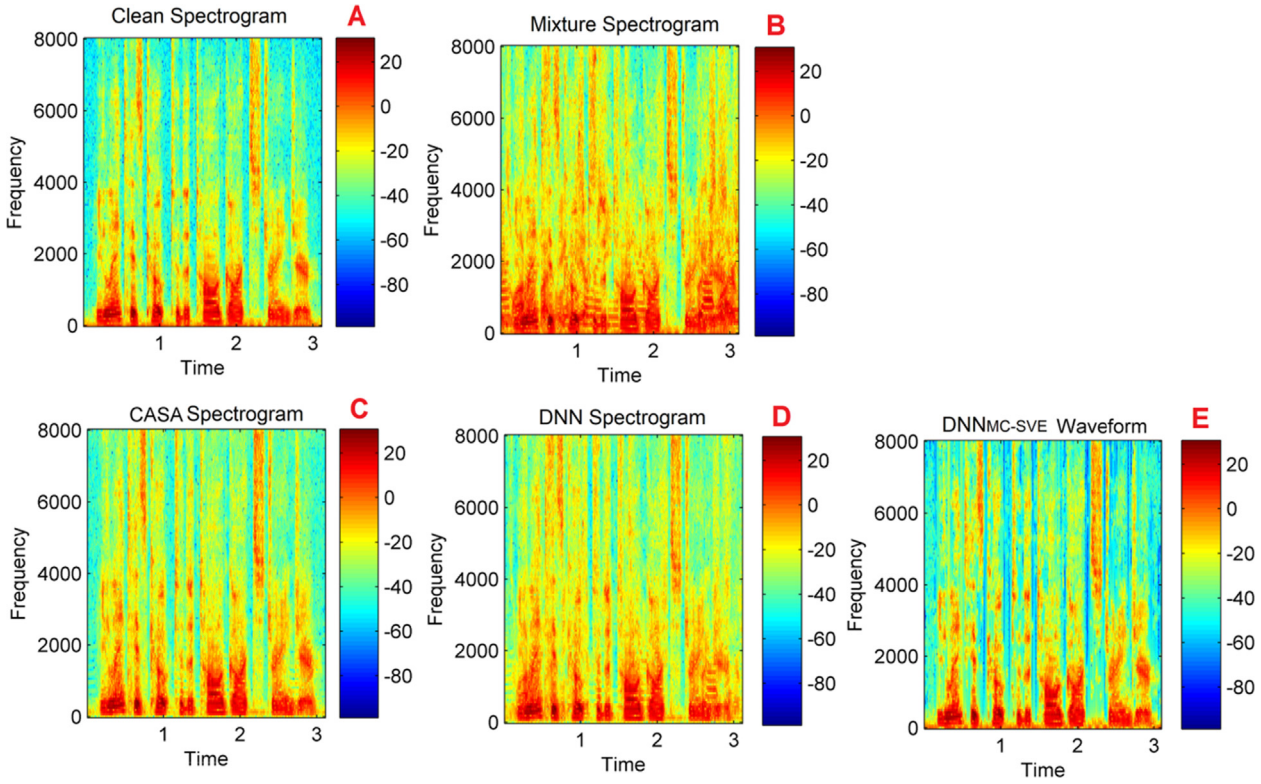

Fig. 11. Time-varying Spectral Analysis. (A) Clean speech, (B) 4-talker mixture at 5 dB SNR, (C) Separated by CASA, (D) Separate by DNN, and (E) Separated by $DNN_{MC\text{-}SVE}$.

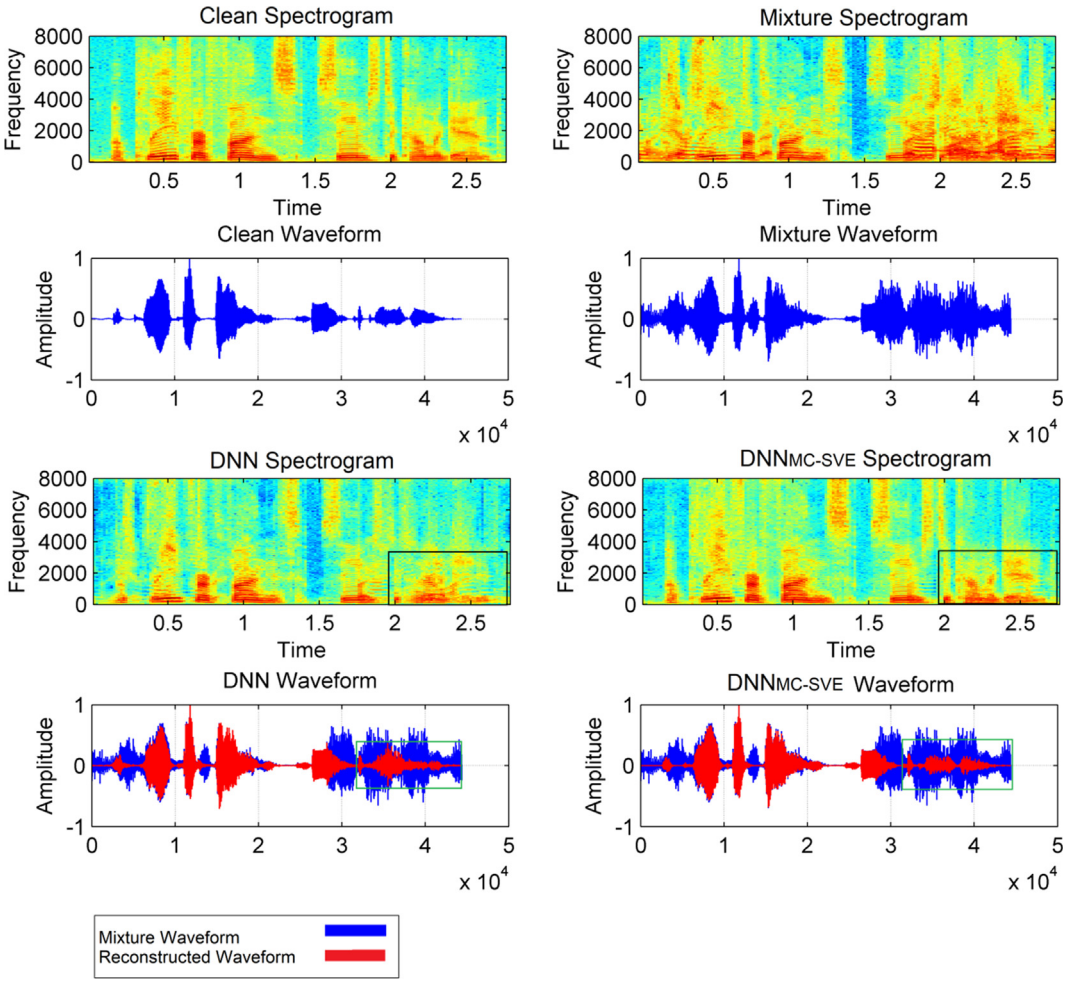

Fig. 12. Time-Waveform and Spectrogram Analysis of speech separated by DNN and $DNN_{MC\text{-}SVE}$.

DNN, with PESQ score improved from 1.87 to 2.21 for 3-talker mixture at low SNR (-5dB). Jointly, the $DNN_{MC\text{-}SVE}$ achieved significant improvements over baseline DNN, DNN_{MC}, and DNN_{SVE}. The PESQ further improved consistently, with PESQ improved from 1.77, 2.01, and 2.12 to 2.23 for 4-talker mixture at −5dB SNR. The average PESQ scores are improved from 2.15, 2.40, and 2.48 to 2.56 for 2-talker, 3-talker and 4-talker mixtures. In Table 8, we compare the HIT-FA rates for three DNN configurations; DNN_{MC}, DNN_{SVE}, and $DNN_{MC\text{-}SVE}$. For this experiment, again a set of 400 speech utterances and three mixture types are used. All three configurations have 3 hidden layers and 1024 neurons in all hidden layers. Compared with DNN_{MC}, and DNN_{SVE}, high HIT rates (lowest with 4-talker at −5dB, 84.36%) and low FA rates (highest with 2-talker at −5dB, 13.76%) are achieved with $DNN_{MC\text{-}SVE}$. The average hit rate achieved with baseline DNN is about 6.94% lower than that of $DNN_{MC\text{-}SVE}$. We have used two databases to evaluate the performance and it is examined in the experiments that proposed method performed better for WJS0-2mix dataset.

Finally, we performed time-varying spectral analysis to evaluate the performance gain of $DNN_{MC\text{-}SVE}$ over DNN and CASA. Fig. 11 shows a sample spectrogram analysis where a clean speech utterance is mixed with the 4-talker mixture at 5 dB SNR. The spectrogram of $DNN_{MC\text{-}SVE}$ is depicted in Fig. 11 (E), and it is clear that the harmonic spectrums of the vowel are sustained. The formant peaks are maintained because of spectral variance equalization. Moreover, the spectrogram also revealed a fine structure during speech activity areas. By analyzing the spectrum during speech-pause areas, $DNN_{MC\text{-}SVE}$ outperforms in removing the residual noise. The weak harmonic structures in high-frequency subbands are better preserved. Therefore, the perceptual quality of the enhanced speech provides by $DNN_{MC\text{-}SVE}$ is better. The utterance with weak energy is preserved and yielded less speech distortion; hence, the speech intelligibility is improved. The residual noise is evident in the spectrograms of CASA and DNN, shown in Fig. 11 (C)-(D). Fig. 12 shows the time waveform and time-varying spectral analysis of $DNN_{MC\text{-}SVE}$ and the competing DNN. A clean speech utterance is mixed with 3-talker mixture at 0 dB SNR. The areas highlighted with boxes indicate that the $DNN_{MC\text{-}SVE}$ successfully separated the speech from the mixture signal. Residuals are evident in the spectrogram and time waveform of the competing DNN.

5. Summary and conclusions

A supervised binary classification approach (IBM) is proposed in this paper for speaker-independent multi-talker speech separation based on the DNN. DNNs are trained to learn a mapping from the mixture features and estimate a binary time–frequency mask based on the mean square error (MSE) objective cost function, Monte-Carlo Dropout Regularization, and standard backpropagation. The over-smoothing problem is addressed and solved by performing spectral variance equalization. Four performance metrics and two competing methods are used in the experiments to eval-

uate the performance of the $DNN_{MC\text{-}SVE}$. The average objective quality analysis based on the PESQ and SNRSeg scores indicate that $DNN_{MC\text{-}SVE}$ outperformed the CASA and DNN at all SNRs. Similarly, the average predicted STOI and *n*FwSNRSeg scores confirmed superiority of $DNN_{MC\text{-}SVE}$ over CASA and DNN. Moreover, two DNN configurations, denoted as; DNN_{MC} and DNN_{SVE} are used in experiment to compare the performance of $DNN_{MC\text{-}SVE}$ in terms of PESQ and HIT-FA rates. The conclusions of the proposed method are summarized as:

i. The PESQ scores concluded that $DNN_{MC\text{-}SVE}$ outperformed the CASA and competing DNN based separation methods at all SNRs consistently. The predicted scores with three scenarios are improved significantly. $DNN_{MC\text{-}SVE}$ achieved substantial PESQ gains at low SNRs. The average ΔPESQ scores at −5dB and 0 dB for three mixtures are 1.15 and 0.91.
ii. The SNRSeg scores concluded that $DNN_{MC\text{-}SVE}$ outperformed the CASA and competing DNN based separation methods consistently. $DNN_{MC\text{-}SVE}$ achieved substantial SNRSeg gains at low SNRs. The average ΔSNRSeg scores at −5dB and 0 dB for three mixtures are 1.65 dB and 3.91 dB. Scores in the three scenarios suggest improved performance gain.
iii. $DNN_{MC\text{-}SVE}$ outperformed the CASA and competing DNN consistently in all conditions in terms of the STOI and *n*FwSNRSeg. $DNN_{MC\text{-}SVE}$ provided improved intelligibility scores (STOI $\geq$ 85%, *n*FwSNRSeg $\geq$ 1.00). The best predicted scores are: STOI = 96.04% and *n*FwSNRSeg = 1.48 dB.
iv. High HIT rates and low FA rates are obtained with $DNN_{MC\text{-}SVE}$ models. The average hit rate obtained with DNN is about 5.2% lower than $DNN_{MC\text{-}SVE}$. Similarly, the average FA rate for $DNN_{MC\text{-}SVE}$ is 12.88% which is lower than DNN (FA = 16.82% for DNN). The delta features improved the HIT rate significantly (7%) without increasing the FA rate.
v. The time-varying spectral analysis concluded that harmonic spectrums of the vowel and formant peaks are sustained because of the spectral variance equalization. Weak energies are preserved and yielded less speech distortion.

To conclude, the $DNN_{MC\text{-}SVE}$ performed significantly, providing high speech quality and the target speech utterances are excellently separated from three mixtures. Moreover, $DNN_{MC\text{-}SVE}$ showed high speech intelligibility in all mixture types. In the future work, we would be devoted in attempting further improvements in the performance of the proposed method by incorporating the phase information. Also, we will systematically examine the acoustic features set to find more robust acoustic features set in order to train the DNN structure more efficiently.

CRediT author statements

Nasir Saleem: Conceptualization, Data Creation, Methodology, Investigation, Software, Validation, Writing. **Muhammad Irfan Khattak:** Methodology, Investigation, Software, Validation, Writing, Review and Editing.

Acknowledgements

All persons who have made substantial contributions to the work reported in the manuscript (e.g., technical help, writing and editing assistance, general support), but who do not meet the criteria for authorship, are named in the Acknowledgements and have given us their written permission to be named. If we have not included an Acknowledgements, then that indicates that we have not received substantial contributions from non-authors.

References

[1] Zhang XL, Wang D. A deep ensemble learning method for monaural speech separation. IEEE/ACM Trans Audio, Speech and Language Processing (TASLP) 2016;24(5):967–77.
[2] Du J, Tu Y, Dai LR, Lee CH. A regression approach to single-channel speech separation via high-resolution deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 2016;24(8):1424–37.
[3] Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., ... & Zweig, G. (2016). Achieving human parity in conversational speech recognition. arXiv preprint arXiv:1610.05256.
[4] Cao Y, Sridharan S, Moody A. Multichannel speech separation by eigendecomposition and its application to co-talker interference removal. IEEE Trans Speech Audio Process 1997;5(3):209–19.
[5] Toroghi RM, Faubel F, Klakow D. Multi-channel speech separation with soft time-frequency masking. In SAPA-SCALE Conference, 2012.
[6] Hu G, Wang D. A tandem algorithm for pitch estimation and voiced speech segregation. IEEE Trans Audio Speech Lang Process 2010;18(8):2067–79.
[7] Hu K, Wang D. An unsupervised approach to cochannel speech separation. IEEE Trans Audio Speech Lang Process 2012;21(1):122–31.
[8] Jin Z, Wang D. A supervised learning approach to monaural segregation of reverberant speech. IEEE Trans Audio Speech Lang Process 2009;17(4):625–38.
[9] Wang D, Chen J. Supervised speech separation based on deep learning: An overview. IEEE/ACM Trans Audio Speech Lang Process 2018;26(10):1702–26.
[10] Wang D. Time-frequency masking for speech separation and its potential for hearing aid design. Trends in amplification 2008;12(4):332–53.
[11] Hu, G., & Wang, D. (2001). Speech segregation based on pitch tracking and amplitude modulation. In Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575) (pp. 79-82). IEEE.
[12] Roman N, Wang D, Brown GJ. Speech segregation based on sound localization. J Acous Soc Am 2003;114(4):2236–52.
[13] Seltzer ML, Raj B, Stern RM. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. Speech Commun 2004;43(4):379–93.
[14] Jensen J, Hendriks RC. Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions. IEEE Trans Audio Speech Lang Process 2011;20(1):92–102.
[15] Kim G, Lu Y, Hu Y, Loizou PC. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. The Journal of the Acoustical Society of America 2009;126(3):1486–94.
[16] Kang TG, Kwon K, Shin JW, Kim NS. NMF-based target source separation using deep neural network. IEEE Signal Process Lett 2014;22(2):229–33.
[17] Smaragdis P. Convolutive speech bases and their application to supervised speech separation. IEEE Trans Audio Speech Lang Process 2006;15(1):1–12.
[18] Wood SU, Rouat J, Dupont S, Pironkov G, Wood SU, Rouat & Pironkov G. Blind speech separation and enhancement with GCC-NMF. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 2017;25(4):745–55.
[19] Kolbk M, Tan ZH, Jensen J. Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 2017;25(1):153–67.
[20] Saleem N, Irfan Khattak M, Qazi AB. Supervised speech enhancement based on deep neural network. J Intell Fuzzy Syst 2019;Preprint:1–15.
[21] Saleem N, Irfan Khattak M, Ali MY, Shafi M. Deep neural network for supervised single-channel speech enhancement. Archives of Acoustics 2019;44.
[22] Zao L, Coelho R, Flandrin P. Speech enhancement with emd and hurst-based mode selection. IEEE/ACM Trans Audio Speech Lang Process 2014;22(5):899–911.
[23] Chen J, Wang D. Long short-term memory for speaker generalization in supervised speech separation. The Journal of the Acoustical Society of America 2017;141(6):4705–14.
[24] Seo H, Lee M, Chang JH. Integrated acoustic echo and background noise suppression based on stacked deep neural networks. Appl Acoust 2018;133:194–201.
[25] Wang Y, Wang D. Towards scaling up classification based speech separation. IEEE Trans Audio Speech Lang Process 2013;21(7):1381–90.
[26] Gogate, M., Adeel, A., Marxer, R., Barker, J., & Hussain, A. (2018). DNN driven speaker independent audio-visual mask estimation for speech separation. arXiv preprint arXiv:1808.00060.
[27] Liu, Q., Xu, Y., Jackson, P. J., Wang, W., & Coleman, P. (2018). Iterative deep neural networks for speaker-independent binaural blind speech separation. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 541-545). IEEE.
[28] Weng C, Yu D, Seltzer ML, Droppo J. Deep neural networks for single-channel multi-talker speech recognition. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 2015;23(10):1670–9.
[29] Luo Y, Chen Z, Mesgarani N. Speaker-independent speech separation with deep attractor network. IEEE/ACM Trans Audio Speech Lang Process 2018;26(4):787–96.
[30] Kolbæk M, Yu D, Tan ZH, Jensen J, Kolbaek M, Yu D, et al. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 2017;25(10):1901–13.

[31] Zheng N, Zhang XL. Phase-Aware Speech Enhancement Based on Deep Neural Networks. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 2019;27(1):63–76.
[32] Jiang W, Wen F, Liu P. Robust beamforming for speech recognition using DNN-based time-frequency masks estimation. IEEE Access 2018;6:52385–92.
[33] Mayer F, Williamson DS, Mowlaee P, Wang D. Impact of phase estimation on single-channel speech separation based on time-frequency masking. The Journal of the Acoustical Society of America 2017;141(6):4668–79.
[34] Min E, Guo X, Liu Q, Zhang G, Cui J, Long J. A survey of clustering with deep learning: From the perspective of network architecture. IEEE Access 2018;6:39501–14.
[35] Kolbæk, M., Yu, D., Tan, Z. H., & Jensen, J. (2017). Joint separation and denoising of noisy multi-talker speech using recurrent neural networks and permutation invariant training. In 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1-6). IEEE.
[36] Yu, D., Kolbæk, M., Tan, Z. H., & Jensen, J. (2017). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 241-245). IEEE.
[37] Xu, C., Rao, W., Xiao, X., Chng, E. S., & Li, H. (2018). Single channel speech separation with constrained utterance level permutation invariant training using grid lstm. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6-10). IEEE.
[38] Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 31-35). IEEE.
[39] Chen J, Wang Y, Wang D. A feature study for classification-based speech separation at low signal-to-noise ratios. IEEE/ACM Trans Audio Speech Lang Process 2014 Dec;22(12):1993–2002.
[40] Chen J, Wang D. DNN based mask estimation for supervised speech separation. In Audio source separation. Cham: Springer; 2018. p. 207–35.
[41] Saleem, N., & Khattak, M. I. (2019). Deep neural networks for speech enhancement in complex-noisy environments. International Journal of Interactive Multimedia and Artificial Intelligence, vol. In Press, issue In Press, no, 1-7.
[42] Wang Yannan, Jun Du, Dai Li-Rong, Lee Chin-Hui. A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks. IEEE/ACM Transactions on Audio: Speech, and Language Processing; 2017.
[43] Dahl GE, Sainath TN, Hinton GE. In: May). Improving deep neural networks for LVCSR using rectified linear units and dropout. IEEE; 2013. p. 8609–13.
[44] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 2014;15(1):1929–58.
[45] Gal Y, Ghahramani Z. June). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: In international conference on machine learning. p. 1050–9.
[46] Kendall A, Cipolla R. In: May). Modelling uncertainty in deep learning for camera relocalization. IEEE; 2016. p. 4762–9.
[47] Nazreen, P. M., & Ramakrishnan, A. G. (2018). DNN Based Speech Enhancement for Unseen Noises Using Monte Carlo Dropout. In 2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS) (pp. 1-6). IEEE.
[48] Xu Y, Du J, Dai LR, Lee CH. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 2014;23 (1):7–19.
[49] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research. 2011:2121–59.
[50] Toda, T., Black, A. W., & Tokuda, K. (2005). Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. (Vol. 1, pp. I-9). IEEE.
[51] Rothauser EH. IEEE recommended practice for speech quality measurements. IEEE Trans. on Audio and Electroacoustics. 1969;17:225–46.
[52] Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016, March). Deep clustering: Discriminative embeddings for segmentation and separation. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 31-35). IEEE.
[53] Garofolo J, Graff D, Paul D, Pallett D. CSR-I (WSJ0) complete LDC93S6A. Web Download. Philadelphia: Linguistic Data Consortium; 1993. p. 83.
[54] Taal CH, Hendriks RC, Heusdens R, Jensen J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans Audio Speech Lang Process 2011;19(7):2125–36.
[55] Ma J, Hu Y, Loizou PC. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J Acous Soc Am 2009;125(5):3387–405.
[56] Rix AW, Beerends JG, Hollier MP, Hekstra AP. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on 2001 (Vol. 2, pp. 749-752). IEEE.
[57] Loizou PC. Speech enhancement: theory and practice. CRC Press; 2013.
[58] Li N, Loizou PC. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. The Journal of the Acoustical Society of America. 2008 Mar;123(3):1673–82.
[59] Wang Y, Narayanan A, Wang D. On training targets for supervised speech separation. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 2014 Dec 1; 22(12):1849-58.