Technical Report OSU-CISRC-8/15-TR13
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftpsite: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2015**
File: **TR13.pdf**
Website: **http://www.cse.ohio-state.edu/research/techReport.shtml**

Deep Ensemble Learning for Monaural Speech Separation

Xiao-Lei Zhang
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
xiaolei.zhang9@gmail.com

DeLiang Wang
Department of Computer Science and Engineering & Center for Cognitive and Brain Sciences
The Ohio State University, Columbus, OH 43210, USA
dwang@cse.ohio-state.edu

Abstract – Monaural speech separation is a fundamental problem in robust speech processing. Recently, deep neural network (DNN) based speech separation methods, which predict either clean speech or an ideal time-frequency mask, have demonstrated remarkable performance improvement. However, a single DNN with a given window length does not leverage contextual information sufficiently, and the differences between the two optimization objectives are not well understood. In this paper, we propose to stack ensembles of DNNs, named multi-resolution stacking, to address monaural speech separation. Each DNN in a module of the stack takes the concatenation of original acoustic features and expansion of the soft output of the lower module as its input, and predicts the ideal ratio mask of the target speaker. The DNNs in the same module explore different contexts by employing different window lengths. We have conducted extensive experiments with three speech corpora. The results demonstrate the effectiveness of the proposed method. We have also compared the two optimization objectives systematically and found that predicting the ideal time-frequency mask is more efficient in utilizing clean training speech, while predicting clean speech is less sensitive to SNR variations.

Index Terms – Deep neural networks, ensemble learning, mapping-based separation, masking-based separation, monaural speech separation, multi-resolution stacking.

Deep Ensemble Learning for Monaural Speech Separation

Xiao-Lei Zhang, *Member, IEEE* and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Monaural speech separation is a fundamental problem in robust speech processing. Recently, deep neural network (DNN) based speech separation methods, which predict either clean speech or an ideal time-frequency mask, have demonstrated remarkable performance improvement. However, a single DNN with a given window length does not leverage contextual information sufficiently, and the differences between the two optimization objectives are not well understood. In this paper, we propose to stack ensembles of DNNs, named multi-resolution stacking, to address monaural speech separation. Each DNN in a module of the stack takes the concatenation of original acoustic features and expansion of the soft output of the lower module as its input, and predicts the ideal ratio mask of the target speaker. The DNNs in the same module explore different contexts by employing different window lengths. We have conducted extensive experiments with three speech corpora. The results demonstrate the effectiveness of the proposed method. We have also compared the two optimization objectives systematically and found that predicting the ideal time-frequency mask is more efficient in utilizing clean training speech, while predicting clean speech is less sensitive to SNR variations.**

***Index Terms*—Deep neural networks, ensemble learning, mapping-based separation, masking-based separation, monaural speech separation, multi-resolution stacking.**

I. Introduction

MONAURAL speech separation aims to separate the speech signal of a target speaker from background noise or interfering speech from a single-microphone recording. In this paper, we focus on the problem of separating a target speaker from an interfering speaker. This problem is challenging because the target and interfering speakers have similar spectral shapes. A solution is important for a wide range of applications, such as speech communication, speech coding, speaker recognition, and speech recognition. It is theoretically an ill-posed problem with a single microphone, and to solve this problem, various assumptions have to be made. Recently, supervised (data-driven) speech separation has received much attention [22]. Based on the definition of the training target, supervised separation methods can be categorized to (i) *masking-based methods* and (ii) *mapping-based methods*.

Masking-based methods learn a mapping function from a mixed signal to a time-frequency (T-F) mask, and then use the estimated mask to separate the mixed signal. These methods typically predict the ideal binary mask (IBM) or ideal ratio mask (IRM). For the IBM [21], a T-F unit is assigned 1, if the signal-to-noise ratio (SNR) within the unit exceeds a local criterion, indicating target dominance. Otherwise, it is assigned 0, indicating interference dominance. For the IRM [17], a T-F unit is assigned some ratio of target energy and mixture energy. Kim *et al.* [15] used Gaussian mixture models (GMM) to learn the distribution of target and interference dominant T-F units and then built a Bayesian classifier to estimate the IBM. Jin and Wang [14] employed multilayer perceptron with one hidden layer, to estimate the IBM, and their method demonstrates promising results in reverberant conditions. Han and Wang [9] used support vector machines (SVM) for mask estimation and produced more accurate classification than GMM-based classifiers. May and Dau [16] first used GMM to calculate the posterior probabilities of target dominance in T-F units and then trained SVM with the new features for IBM estimation. Their method can generalize to a wide range of SNR variation.

Recently, motivated by the success of deep neural networks (DNN) with more than one hidden layer, Wang and Wang [24] first introduced DNN to perform binary classification for speech separation. Their DNN-based method significantly outperforms earlier separation methods. Subsequently, Wang *et al.* [23] examined a number of training targets and suggested that the IRM should be preferred over the IBM in terms of speech quality. Huang *et al.* [11], [12] used DNN to predict the IRM, and demonstrated significant performance improvement over standard non-negative matrix factorization based methods.

Mapping-based methods learn a regression function from a mixed signal to clean speech directly, which differs from masking-based methods in optimization objectives. Xu *et al.* [25], [26] trained DNN as a regression model to perform speech separation and showed a significant improvement over conventional speech enhancement methods. Han *et al.* [8], [10] used DNN to learn a mapping from reverberant and reverberant-noisy speech to anechoic speech. Their spectral mapping approach substantially improves SNR and objective speech intelligibility. Du *et al.* [6] improved the method in [25] with global variance equailization, dropout training, and noise-aware training strategies. They demonstrated significant improvement over a GMM-based method and good generalization to unseen speakers in testing. Tu *et al.* [20] trained DNN to estimate not only the target speech but also the interfering speech. They showed that using dual outputs improves the quality of speech separation.

We investigate DNN-based speech separation by incorpo-

Xiao-Lei Zhang and DeLiang Wang are with the Department of Computer Science & Engineering and Center for Cognitive & Brain Sciences, The Ohio State University, Columbus, OH, USA, and the Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Xi'an, China (e-mail: xiaolei.zhang9@gmail.com, dwang@cse.ohio-state.edu).

rating DNN into the framework of ensemble learning [5], which integrates multiple weak learners to create a stronger one. Ensemble learning is a methodology applicable to various machine learning methods, including DNN. To our knowledge, ensemble methods have not been systematically explored for speech separation. There are two key elements for ensemble learning to succeed: (i) weak learners are at least stronger than random guess, and (ii) strong diversity exists among the weak learners [5]. For the former, DNN is a good choice; for the latter, there are a number of ways to enlarge the diversity by manipulating input features, output targets, training data, and hyperparameters of base learners [5].

In this paper, we propose a deep ensemble learning method, called *multi-resolution stacking*, which uses DNN as the base learner and manipulates the input features and output targets of DNNs for enlarging learner diversity and exploring complementary contexts. In addition, we analyze the differences between the two optimization objectives, i.e. ideal masking and spectral mapping, systematically. The contributions of this paper are summarized as follows:

- **Multi-resolution stacking (MRS) for speech separation.** MRS is a stack of DNN ensembles. Each DNN in a module of the stack uses the IRM as the training target. It first concatenates original acoustic features and the estimated ratio masks from the lower module as a new acoustic feature, and then takes the expansion of the new feature in a window (called a resolution) as its input. The DNNs in the same module have different resolutions. MRS improves the accuracy of DNN by ensembling and stacking, and enlarges the diversity between the DNNs with the multi-resolution scheme which manipulates the input features of DNNs.
- **Comparison of masking and mapping for DNN-based speech separation.** The methods in comparison use the same type of DNN in MRS. Our systematic comparison leads to the following conclusions. (i) The masking-based approach is more effective in utilizing the clean training speech of a target speaker. (ii) The mapping-based method is less sensitive to the SNR variation of a training corpus. (iii) Given a training corpus with a fixed mixture SNR and plenty of clean training speech from the target speaker, the mapping and masking-based methods tend to perform equally well.

We have conducted extensive experiments on the corpora of speech separation challenge [2], TIMIT [7], and IEEE [13], and found that the proposed MRS method outperforms previous mapping- and masking-based methods in all experiments.

This paper is organized as follows. In Section II, we present the MRS algorithm. In Section III, we analyze the differences between mapping and masking. In Section IV, we present the results. Finally, we conclude in Section V.

II. Multi-Resolution Stacking

Speech signal is highly structured, and leveraging temporal context is important for improving the performance of a speech processing method. Generally, a learning machine uses the concatenation of neighboring frames instead of a single frame as its input for predicting the output. A good choice of input expansion is to select a fixed window that performs the best among several candidate windows. For example, in [11], the masking-based method sets the window length to 3; in [6], the mapping-based method sets the window length to 7. However, different candidate windows may provide complementary information that can further improve the performance. Motivated by the recent success of the multi-resolution cochleagram feature [1] and the relationship between the feature and its components [27], we propose the multi-resolution stacking algorithm for speech separation, where the term "resolution" denotes a window of neighboring frames.

MRS is a stack of ensemble learning machines, as shown in Fig. 1. The learning machines in a module of the stack have different resolutions; they take the concatenation of the output predictions of their lower module and the original acoustic features as their input. MRS can be either mapping-based, masking-based, or a combination of mapping and masking. In this paper, we instantiate the learning machines by DNN and use the IRM as the optimization objective.

In the preprocessing stage of MRS training, given a mixed signal and the corresponding clean signals of a target speaker and an interfering speaker, we extract the magnitude spectra of their short time Fourier transform (STFT) features, denoted as $\{\mathbf{y}_m\}_{m=1}^{M}$, $\{\mathbf{x}_m^a\}_{m=1}^{M}$, and $\{\mathbf{x}_m^b\}_{m=1}^{M}$, respectively, where M is the number of frames for the mixed signal, and subscript a denotes the target speaker and subscript b the interfering speaker. We further calculate the ideal ratio mask of the target speaker, denoted as $\{IRM_m\}_{m=1}^{M}$, from the STFT features (see Section III for the definitions of the ideal ratio mask).

In the training stage, MRS learns a mapping function $IRM = f(\mathbf{y})$ given a training corpus of mixed signals. Suppose MRS trains S modules, and the sth module has P_s learning machines, denoted as $\{f_p^{(s)}(\cdot)\}_{p=1}^{P_s}$, each of which has a unique resolution $W_p^{(s)}$. The pth DNN learns the mapping function $IRM_m = f_p^{(s)}(\mathbf{v}_{m,p}^{(s)})$ where the input $\mathbf{v}_{m,p}^{(s)}$ is an expansion of the feature $\mathbf{u}_m^{(s)}$ at resolution $W_p^{(s)}$:

$$\mathbf{v}_{p,m}^{(s)}{=}\Bigg[\mathbf{u}_{m-W_p^{(s)}}^{(s)}{}^{T}, \mathbf{u}_{m-W_p^{(s)}+1}^{(s)}{}^{T} \ldots, \mathbf{u}_m^{(s)}{}^{T}, \\ \ldots, \mathbf{u}_{m+W_p^{(s)}-1}^{(s)}{}^{T}, \mathbf{u}_{m+W_p^{(s)}}^{(s)}{}^{T}\Bigg]^{T} \quad (1)$$

with $\{\mathbf{u}_n^{(s)}\}_{n=m-W_p^{(s)}}^{m+W_p^{(s)}}$ defined as:

$$\mathbf{u}_n^{(s)} = \begin{cases} \mathbf{y}_n & \text{if } s = 1 \\ \left[RM_{n,1}^{(s-1)}{}^{T}, \ldots, RM_{n,P_{s-1}}^{(s-1)}{}^{T}, \mathbf{y}_n^T\right]^T & \text{if } s > 1 \end{cases} \quad (2)$$

where $\{RM_{n,l}^{(s-1)}\}_{l=1}^{P_{s-1}}$ are the estimated IRMs of $\mathbf{y}_n$ produced by the $(s-1)$th module $\{f_l^{(s-1)}(\cdot)\}_{l=1}^{P_{s-1}}$. Note that we usually train only one model with an empirically optimal resolution at the top module, as illustrated in Fig. 1.

In the test stage of MRS, given a mixed signal of two speakers in the time domain, we first extract $\{\mathbf{y}_m \exp(j\boldsymbol{\theta}_m)\}_{m=1}^{M}$ by STFT, where $\mathbf{y}_m$ and $\boldsymbol{\theta}_m$ represent the magnitude vector and phase vector of the mth frame respectively. We use $\{\mathbf{y}_m\}_{m=1}^{M}$

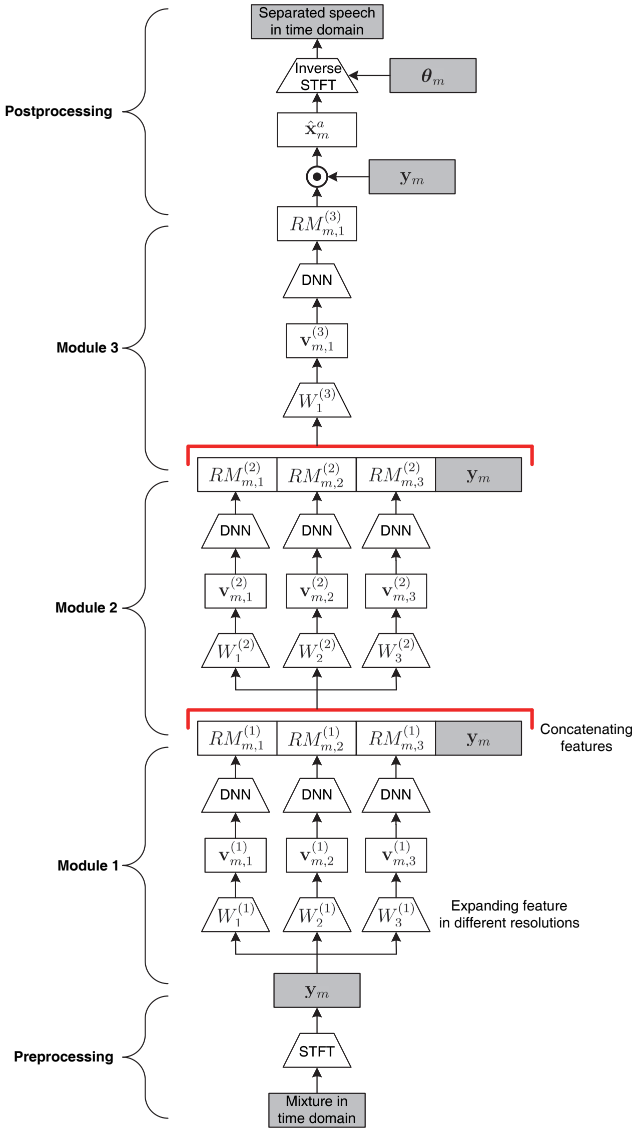
Separated speech in time domain
Inverse STFT
$\boldsymbol{\theta}_m$
$\hat{\mathbf{x}}^a_m$
$\odot$
$\mathbf{y}_m$
Postprocessing
$RM^{(3)}_{m,1}$
DNN
$\mathbf{v}^{(3)}_{m,1}$
$W^{(3)}_1$
Module 3
$RM^{(2)}_{m,1}$ $RM^{(2)}_{m,2}$ $RM^{(2)}_{m,3}$ $\mathbf{y}_m$
DNN DNN DNN
$\mathbf{v}^{(2)}_{m,1}$ $\mathbf{v}^{(2)}_{m,2}$ $\mathbf{v}^{(2)}_{m,3}$
$W^{(2)}_1$ $W^{(2)}_2$ $W^{(2)}_3$
Module 2
$RM^{(1)}_{m,1}$ $RM^{(1)}_{m,2}$ $RM^{(1)}_{m,3}$ $\mathbf{y}_m$
Concatenating features
DNN DNN DNN
$\mathbf{v}^{(1)}_{m,1}$ $\mathbf{v}^{(1)}_{m,2}$ $\mathbf{v}^{(1)}_{m,3}$
$W^{(1)}_1$ $W^{(1)}_2$ $W^{(1)}_3$
Expanding feature in different resolutions
Module 1
$\mathbf{y}_m$
STFT
Preprocessing
Mixture in time domain

Fig. 1. Diagram of multi-resolution stacking. The symbols in the figure are defined in Section II. Trapezoid modules represent resolutions or DNNs. Rectangle modules represent features.

as the input of MRS and get the estimated ratio masks in each module. After getting the estimated ratio masks $\{RM^{(S)}_m\}_{m=1}^M$ from the top module, we first get the estimated magnitude spectra $\{\hat{\mathbf{x}}^a_m\}_{m=1}^M$ by $\hat{\mathbf{x}}^a_m = RM^{(S)}_m \odot \mathbf{y}_m$ and then transform $\{\hat{\mathbf{x}}^a_m \exp(j\boldsymbol{\theta}_m)\}_{m=1}^M$ back to the time-domain signals via the inverse STFT, where the operator $\odot$ denotes the element-wise product. Note that we use the noisy phase to do resynthesis, and the Hamming window in STFT.

A DNN model has a number of nonlinear hidden layers plus an output layer. Each layer has a number of model neurons (or mapping functions). The model can be described as follows:

$$IRM = g\left(h_L\left(\ldots h_l\left(\ldots h_2\left(h_1\left(\mathbf{y}\right)\right)\right)\right)\right) \tag{3}$$

where $l = 1, \ldots, L$ denotes the lth hidden layer from the bottom, $h_l(\cdot)$ denotes nonlinear activation functions of the lth hidden layer, $g(\cdot)$ activation functions of the output layer, and $\mathbf{y}$ is the input feature vector. Common activation functions for the hidden layers include the sigmoid function $b = \frac{1}{1+e^{-a}}$, tanh function, and more recently rectified linear function $b = \max(0, a)$ where a is the input and b the output of a neuron. Common activation functions in the output layer include the linear function $b = a$, softmax function, and sigmoid function. Because the rectified linear function is shown to result in faster training and better learning of local patterns, we use it as the activation function for the hidden layers of DNN. As the training target is the IRM whose value varies between $[0, 1]$, we use the sigmoid function for the output layer.

Traditionally, DNN employs full connections between consecutive layers, which tends to overfit data and be sensitive to different hyperparameter settings. Dropout [3], which randomly deactivates a percentage of neurons, was proposed recently to alleviate the problem. It has been analyzed theoretically that dropout provides as a regularization term for DNN training. Due to this regularization, we are able to train much larger DNN model. Therefore, we use dropout for DNN training.

Although early research in deep learning uses pretraining to prevent poor local minima, recent experience shows that, when data sets are large enough, pretraining does not further improve the performance of DNN. Therefore, we do not pretrain DNN. In addition, we use the adaptive stochastic gradient descent algorithm [4] with a momentum term [18] to accelerate gradient descent and to facilitate parallel computing.

Note that the proposed MRS-based speech separation is different from our preliminary work in [28] which used MRS for separating speech from nonspeech noise, boosted DNN as the base weak learner, ideal binary mask as the optimization objective, and multi-resolution cochleagram [1] as the acoustic feature.

III. Mapping and Masking

A general training objective of DNN-based speech separation methods is as follows:

$$\min_{\boldsymbol{\alpha}} \sum_{m=1}^{M} \ell(\mathbf{x}^a_m, f_{\boldsymbol{\alpha}}(\mathbf{y}_m)) \tag{4}$$

where $\ell(\cdot)$ is a measurement of training loss and $\boldsymbol{\alpha}$ is the parameter of the speech separation algorithm $f(\cdot)$.

Mapping-based DNN methods learn a mapping function from the spectrum of the mixed signal to the spectrum of the clean speech of the target speaker directly, which can be formulated as the following *minimum mean squared error* problem:

$$\min_{\boldsymbol{\alpha}} \sum_{m=1}^{M} \|\mathbf{x}^a_m - f_{\boldsymbol{\alpha}}(\mathbf{y}_m)\|^2 \tag{5}$$

where $\|\cdot\|^2$ is the squared loss. In the test stage, mapping-based methods transform the prediction $\hat{\mathbf{x}}^a_m = f_{\boldsymbol{\alpha}}(\mathbf{y}_m)$ back to the time-domain signal by inverse STFT.

Masking-based DNN methods learn a mapping function from the spectrum of the mixed signal to the ideal time-frequency mask of the clean utterance of the target speaker:

$$\min_{\boldsymbol{\alpha}} \sum_{m=1}^{M} \|IRM_m - f_{\boldsymbol{\alpha}}(\mathbf{y}_m)\|^2 \tag{6}$$

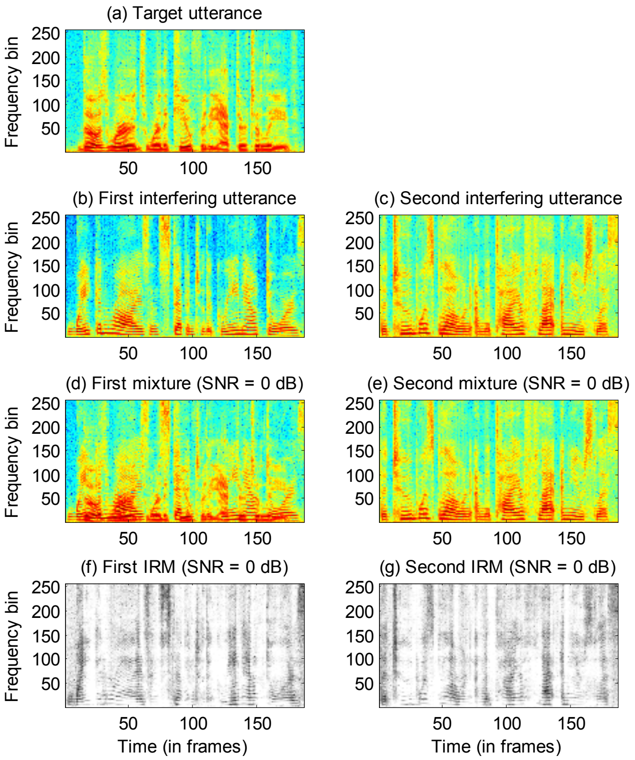

Fig. 2. Comparison of mapping and masking when the number of the utterances of the target speaker is limited. (a) The spectrum of the utterance of the target speaker. (b) The spectrum of the first utterance of the interfering speaker. (c) The spectrum of the second utterance of the interfering speaker. (d) The spectrum of the mixed signal produced from the target utterance (i.e. Fig. 2a) and the first interfering utterance (i.e. Fig. 2b). (e) The spectrum of the mixed signal produced from the target utterance and the second interfering utterance (i.e. Fig. 2c). (f) The IRM of the target utterance given the first interfering utterance. (g) The IRM of the target utterance given the second interfering utterance.

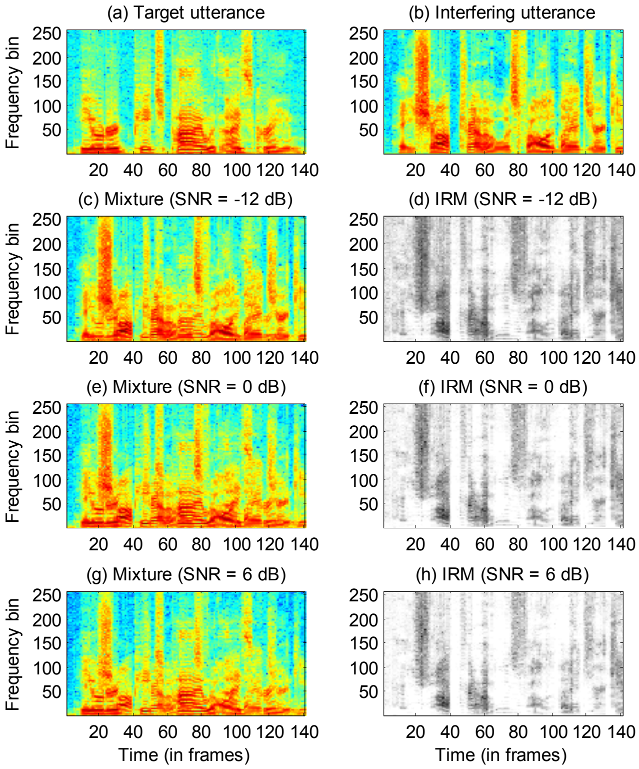

Fig. 3. Comparison of mapping and masking when the SNR of the mixed signal varies in a wide range. (a) The spectrum of an utterance of a target speaker. (b) The spectrum of an utterance of an interfering speaker. (c) The spectrum of the mixed signal with SNR = -12 dB. (d) The IRM of the target speaker with SNR = -12 dB. (e) The spectrum of the mixed signal with SNR = 0 dB. (f) The IRM of the target speaker with SNR = 0 dB. (g) The spectrum of the mixed signal with SNR = 6 dB. (h) The IRM of the target speaker with SNR = 6 dB.

where IRM_m is the ideal mask. In the test stage, we first apply the estimated mask RM_m to the spectrum of the mixed signal $\mathbf{y}_m$ by $\hat{\mathbf{x}}_m^a = RM_m \odot \mathbf{y}_m$ and then transform the estimated spectrum $\hat{\mathbf{x}}_m^a$ back to the time-domain signal by inverse STFT.

The ideal ratio mask in MRS is defined as:

$$IRM_{m,k} = \frac{x_{m,k}^a}{x_{m,k}^a + x_{m,k}^b + \epsilon} \tag{7}$$

where $x_{m,k}^a$ and $x_{m,k}^b$ denote $\mathbf{x}_m^a$ and $\mathbf{x}_m^b$ at frequency k respectively, and ϵ is a very small positive constant to prevent the denominator from being zero.

Here, we analyze the differences between mapping- and masking-based methods. Masking-based methods can explore the mutual information between target and interfering speakers better than mapping-based methods. Specifically, data-driven methods, such as DNN, need a large number of different patterns to train a good machine. When a target speaker has a limited number of utterances, we usually create a large training corpus by mixing each utterance of the target speaker with many utterances of interfering speakers. Fig. 2 illustrates such a process where one utterance of a target speaker (Fig. 2a) is mixed with two utterances of an interfering speaker (Figs. 2b and 2c), each at 0 dB, which produces two spectra from the two mixed signals (Figs. 2d and 2e) and two ideal ratio masks (Figs. 2f and 2g). In the IRM illustrations of Figs. 2f and 2g, white corresponds to 1 and black to 0. Mapping-based methods learn a mapping function from the spectra in Figs. 2d and 2e to the same output pattern in Fig. 2a. On the contrary, masking-based methods learn a mapping function that projects the spectrum in Fig. 2d to the ideal ratio mask in Fig. 2f, and the spectrum in Fig. 2e to the ideal ratio mask in Fig. 2g, respectively. In other words, training targets are different depending on interfering utterances (see also [23]). Therefore, masking-based methods can potentially utilize the training patterns better than mapping-based methods, and hence likely achieve better performance.

Mapping-based methods are less sensitive to the SNR variation of training data than masking-based methods. Specifically, the optimization objective $\min \sum \|\mathbf{x}^a - f(\mathbf{y})\|^2$ (or $\min \sum \|IRM - f(\mathbf{y})\|^2$) tends to recover the spectra $\mathbf{x}^a$ (or the ideal masks IRM) that have large energy and sacrifice those that have small energy, so that the overall loss is minimized. Fig. 3 illustrates such an example, where a target utterance (Fig. 3a) is mixed with an interfering utterance (Fig. 3b) at multiple SNR levels (Figs. 3c, 3e, and 3g). For mapping-based methods, no matter how the SNR changes, the reference $\mathbf{x}^a$ (Fig. 3a) is unchanged, which means that only the energy of $\mathbf{y}$ affects the optimization. On the contrary, for masking-based methods, the energy of the ideal masks IRM (Figs. 3d, 3f, and 3h) becomes small with the decrease of the SNR. One can imagine that when the SNR is low, the estimated ratio mask tends to suffer a larger loss than the estimated reference $\hat{\mathbf{x}}^a$ in mapping-based methods. As a result, when the SNR of a training corpus varies in a wide range, masking-based methods likely perform worse than mapping-based methods at low SNR levels.

Aside from these differences, Wang *et al.* [23] point out that masking as a form of normalization reduces the dynamic range of target values, leading to different training efficiency compared to mapping.

IV. Results and Comparisons

In this section, we compared the mapping-, masking-, and MRS-based speech separation methods in three different training conditions. We trained hundreds of DNN models and report the results of the comparison methods on each gender pair, i.e. male+male (M+M), female+male (F+M), female+female (F+F), and male+female (M+F), where the first speaker of a gender pair is the target speaker in all experiments.

A. Comparison with Single-SNR Speaker-pair Dependent Training

In this training condition, the target and interfering speakers of the training and test corpora are the same, and the training and test corpora are created at each SNR.

1) Datasets: In this experiment, we used the speech separation challenge (SSC) [2] and TIMIT datasets [7] as the separation corpora. SSC has predefined training and test corpora. The training corpus contains 34 speakers, each of which has 500 clean utterances. Each mixed signal in the test corpus is also produced from a pair of speakers in the training corpus. Because each pair of speakers contains at most 2 test mixtures, we did not use the test corpus. Instead, we randomly picked 2 pairs of speakers for each gender pair from the training corpus, and generated 8 separation tasks in total. Each task had 7 SNR levels ranging from $\{-12, -9, -6, -3, 0, 3, 6\}$ dB. For each SNR level of a task, we generated 1000 mixed signals as the training set, and 50 mixed signals as the test set. In other words, test results are reported from only one SNR that is matched to that in the training data. Each component of a mixture in the training set was a clean utterance randomly selected from the first 450 utterances of the corresponding speaker. Each component of a mixed signal in the test set was a clean utterance from the last 50 utterances of the corresponding speaker.

TIMIT contains 630 speakers, each of which has 10 clean utterances. We randomly picked 2 pairs of speakers for each gender pair, and formulated 8 tasks. Each task had 7 SNR levels ranging from $\{-12, -9, -6, -3, 0, 3, 6\}$ dB. For each SNR level of a task, we generated 600 mixed signals as the training set, and 2 mixed signals as the test set. Each mixture in the training set was constructed by randomly selecting 2 clean utterances, each from the first 8 utterances of a speaker, then shifting the interfering utterance randomly, wrapping the shifted utterance circularly, and finally mixing the two utterances together. For the test set, we mix the first target utterance with the first interfering utterance, and the second target utterance with the second interfering utterance. Note that the random shift operation was used to synthesize a large number of mixtures from a small number of clean utterances.

We resampled all corpora to 8 kHz, and extracted the STFT features with the frame length set to 25 ms and the frame shift set to 10 ms.

2) Evaluation Metrics: We used the short-time objective intelligibility (STOI) [19] as the evaluation metric. STOI evaluates the objective speech intelligibility of time-domain signals. It has been shown empirically that STOI scores are well correlated with human speech intelligibility scores. The higher the STOI value is, the better the predicted intelligibility is. STOI is a standard metric for evaluating speech separation performance [23], [6], [12].

3) Comparison Methods and Parameter Settings: We compared the mapping-, masking, and MRS-based speech separation methods. For all comparison methods, we used DFT to extract acoustic features. For the MRS-based method, we trained two modules (i.e. parameter $S = 2$). For the bottom module of MRS, we trained 3 DNNs with parameters $W_1^{(1)}$, $W_2^{(1)}$, $W_3^{(1)}$ set to 1, 2, and 3 respectively. For the top module of MRS, we trained 1 DNN with $W_1^{(2)}$ set to 1.

We searched for the optimal parameter settings of DNN using a development task, and used the optimal settings in all evaluation tasks. The development task was constructed from two male speakers of SSC. Its training set contained 1000 mixtures, and its test set contained 50 mixtures, both of which were at -12 dB.

The selected parameter settings are as follows. DNN was optimized by the minimum mean square error criterion. Each DNN has 2 hidden layers, each of which consists of 2048 rectified linear neurons. The output neurons of the DNN for the mapping-based method are the linear neurons. The output neurons of the DNNs for the masking- and MRS-based methods were the sigmoid functions. The number of epoches for backpropagation training was set to 50. The batch size was set to 128. The scaling factor for the adaptive stochastic gradient descent was set to 0.0015, and the learning rate decreased linearly from 0.08 to 0.001. The momentum of the first 5 epoches was set to 0.5, and the momentum of other epoches was set to 0.9. The dropout rate of the hidden neurons was set to 0.2. The half-window length W was set to 3 for the mapping-based method, and set to 1 for the masking-based method.

Note that we normalized data before training. For the mapping-based method, we first normalized the training data $\{\mathbf{y}_m\}_{m=1}^M$ to zero mean and unit standard deviation in each dimension, and then used the same normalization factor to normalize both the training references $\{\mathbf{x}_m^a\}_{m=1}^M$ and the test data. After getting the predictions in the test stage, we converted the predictions back to the original scale by the same normalization factor. For the masking-based method and MRS, we first normalized $\{\mathbf{y}_m\}_{m=1}^M$ and then used the same normalization factor to normalize the test data.

4) Results: We conducted a comparison at each SNR level of each separation task, and report the average results of the

TABLE I
STOI COMPARISON BETWEEN MAPPING-, MASKING-, AND MRS-BASED SPEECH SEPARATION METHODS WITH SINGLE-SNR SPEAKER-PAIR DEPENDENT TRAINING ON SSC CORPUS. "AVR" INDICATES THE AVERAGE PERFORMANCE. THE NUMBERS IN BOLD INDICATE THE BEST RESULTS.

SNR		−12 dB	−9 dB	−6 dB	−3 dB	0 dB	3 dB	6 dB
M+M	Noisy	0.41	0.47	0.55	0.63	0.71	0.78	0.84
	Mapping	0.65	0.71	0.77	0.81	0.85	0.89	0.92
	Masking	0.67	0.71	0.76	0.81	0.85	0.88	0.91
	MRS	0.68	0.74	0.78	0.83	0.87	0.90	0.93
F+M	Noisy	0.46	0.52	0.58	0.65	0.72	0.78	0.84
	Mapping	0.73	0.78	0.83	0.87	0.90	0.92	0.94
	Masking	0.73	0.78	0.82	0.87	0.90	0.93	0.94
	MRS	0.75	0.80	0.84	0.88	0.91	0.93	0.95
F+F	Noisy	0.51	0.57	0.64	0.70	0.77	0.83	0.89
	Mapping	0.70	0.75	0.79	0.83	0.87	0.91	0.94
	Masking	0.69	0.73	0.77	0.82	0.86	0.89	0.93
	MRS	0.71	0.75	0.80	0.84	0.87	0.91	0.94
M+F	Noisy	0.48	0.53	0.59	0.65	0.71	0.77	0.83
	Mapping	0.78	0.82	0.85	0.88	0.91	0.93	0.94
	Masking	0.80	0.83	0.86	0.89	0.91	0.93	0.95
	MRS	0.81	0.85	0.87	0.90	0.93	0.94	0.95
AVR	Noisy	0.46	0.52	0.59	0.66	0.73	0.79	0.85
	Mapping	0.71	0.77	0.81	0.85	0.88	0.91	**0.94**
	Masking	0.72	0.76	0.81	0.85	0.88	0.91	0.93
	MRS	**0.74**	**0.78**	**0.82**	**0.86**	**0.89**	**0.92**	**0.94**

TABLE II
STOI COMPARISON BETWEEN MAPPING-, MASKING-, AND MRS-BASED SPEECH SEPARATION METHODS WITH SINGLE-SNR SPEAKER-PAIR DEPENDENT TRAINING ON TIMIT CORPUS.

SNR		−12 dB	−9 dB	−6 dB	−3 dB	0 dB	3 dB	6 dB
M+M	Noisy	0.43	0.50	0.58	0.66	0.74	0.81	0.87
	Mapping	0.53	0.58	0.65	0.70	0.75	0.79	0.83
	Masking	0.58	0.62	0.68	0.74	0.79	0.84	0.87
	MRS	0.55	0.61	0.68	0.74	0.77	0.82	0.85
F+M	Noisy	0.54	0.59	0.64	0.70	0.76	0.82	0.87
	Mapping	0.59	0.64	0.68	0.72	0.74	0.77	0.79
	Masking	0.66	0.72	0.78	0.82	0.85	0.88	0.89
	MRS	0.67	0.72	0.79	0.82	0.84	0.87	0.88
F+F	Noisy	0.54	0.60	0.67	0.74	0.80	0.86	0.91
	Mapping	0.58	0.62	0.65	0.68	0.73	0.77	0.81
	Masking	0.59	0.65	0.70	0.73	0.76	0.78	0.84
	MRS	0.59	0.64	0.71	0.75	0.77	0.77	0.83
M+F	Noisy	0.48	0.54	0.60	0.67	0.74	0.80	0.86
	Mapping	0.63	0.67	0.72	0.77	0.80	0.83	0.86
	Masking	0.63	0.68	0.74	0.80	0.84	0.87	0.89
	MRS	0.62	0.68	0.74	0.80	0.84	0.87	0.88
AVR	Noisy	0.50	0.56	0.62	0.69	0.76	0.82	**0.88**
	Mapping	0.58	0.63	0.67	0.72	0.76	0.79	0.82
	Masking	**0.61**	**0.67**	0.72	0.77	**0.81**	**0.84**	0.87
	MRS	**0.61**	**0.67**	**0.73**	**0.78**	**0.81**	0.83	0.86

two tasks that belonged to the same gender pair.

Table I lists the comparison results on the SSC corpus. From the table, we observe that (i) all methods improve STOI scores over the original mixed signals significantly, particularly at low SNR levels; (ii) the MRS-based method slightly outperforms the mapping- and masking-based methods; (iii) the mapping- and masking-based methods perform equally well.

Table II lists the comparison results on the TIMIT corpus. From the table, we observe that (i) all methods improve the STOI scores at the low SNR levels, but the improvement becomes insignificant or nonexistent with the increase of the SNR. (ii) The masking- and MRS-based methods perform equivalently, and significantly outperform the mapping-based method in all cases. (iii) At positive SNR levels, the mapping-based method produces lower STOI scores than the original mixed signals.

Comparing Table I and Table II, we find that the mapping-based method works well on SSC but not on TIMIT, while the masking-based method works well on both corpora, consistent with our analysis in Section III. Note that STOI improvements are smaller on TIMIT than on SSC, reflecting the fact that the TIMIT dateset has much fewer utterances for each speaker.

B. Comparison with Multi-SNR Speaker-pair Dependent Training

In this training condition, the target and interfering speakers of the training and test corpora are the same, and the SNR of the training corpus varies in a wide range.

1) Experimental Settings: In this experiment, we followed the experimental settings in Section IV-A and made 16 speech separation tasks, each of which had 7 test sets. Different from Section IV-A where each task had 7 training sets, we had only 1 training set for each task encompassing various SNRs. Each training set of SSC contained 10,000 mixed signals. Each training set of TIMIT contained 6,000 mixed signals. Each training mixture had a random SNR level varying between −13 dB and 10 dB with the increment of 1 dB.

2) Results: For each speech separation task, we trained only one model for each comparison method, and tested the model on all 7 test sets at different SNRs. Then, we report the average results of the two tasks that belonged to the same gender pair.

Table III lists the comparison results on the SSC corpus. From the table, we observe that (i) all methods improve the STOI scores over the original mixed signals significantly; (ii) the MRS-based method performs overall the best across all SNR levels; (iii) the masking-based method underperforms the mapping-based method at low SNR levels, consistent with our analysis in Section III.

Table IV lists the comparison results on the TIMIT corpus. From the table, we observe a similar performance profile, albeit STOI improvements are lower in TIMIT compared to SSC.

Comparing Table III with Table I, we find that, when a training set is generated from a large number of clean utterances (each speaker in SSC has 450 clean utterances), enlarging the size of the training set from 1000 mixed signals in Table I to 10,000 mixed signals in Table III significantly elevates the performance. On the other hand, we find that, when a training set is constructed from limited clean utterances (each speaker in TIMIT has only 8 utterances), enlarging the size of the training set from 600 mixed signals in Table II to 6000 mixed signals in Table IV does not elevate the performance by as much. This can be seen from the fact that the results at low SNR levels in Table IV are worse than those in Table II.

C. Comparison with Target Dependent Training

In this condition, we compare the generalization ability of the mapping-, masking-, and MRS-based methods when

TABLE III
STOI COMPARISON BETWEEN MAPPING-, MASKING-, AND MRS-BASED SPEECH SEPARATION METHODS WITH MULTI-SNR SPEAKER-PAIR DEPENDENT TRAINING ON SSC CORPUS.

SNR		−12 dB	−9 dB	−6 dB	−3 dB	0 dB	3 dB	6 dB
M+M	Noisy	0.41	0.47	0.55	0.63	0.71	0.78	0.84
	Mapping	0.69	0.75	0.81	0.85	0.88	0.91	0.92
	Masking	0.66	0.72	0.78	0.83	0.87	0.90	0.93
	MRS	0.69	0.76	0.82	0.86	0.90	0.92	0.94
F+M	Noisy	0.46	0.52	0.58	0.65	0.72	0.78	0.84
	Mapping	0.77	0.82	0.86	0.89	0.91	0.93	0.94
	Masking	0.74	0.80	0.85	0.88	0.91	0.93	0.95
	MRS	0.77	0.83	0.87	0.90	0.93	0.95	0.96
F+F	Noisy	0.51	0.57	0.64	0.70	0.77	0.83	0.89
	Mapping	0.73	0.78	0.83	0.86	0.89	0.91	0.93
	Masking	0.69	0.75	0.80	0.84	0.88	0.91	0.93
	MRS	0.73	0.79	0.84	0.87	0.90	0.92	0.94
M+F	Noisy	0.48	0.53	0.59	0.65	0.71	0.77	0.83
	Mapping	0.81	0.85	0.89	0.91	0.93	0.94	0.95
	Masking	0.79	0.84	0.88	0.91	0.93	0.95	0.96
	MRS	0.81	0.86	0.90	0.92	0.94	0.96	0.97
AVR	Noisy	0.46	0.52	0.59	0.66	0.73	0.79	0.85
	Mapping	**0.75**	0.80	0.85	0.88	0.90	0.92	0.94
	Masking	0.72	0.78	0.83	0.87	0.90	0.92	0.94
	MRS	**0.75**	**0.81**	**0.86**	**0.89**	**0.92**	**0.94**	**0.95**

TABLE IV
STOI COMPARISON BETWEEN MAPPING-, MASKING-, AND MRS-BASED SPEECH SEPARATION METHODS WITH MULTI-SNR SPEAKER-PAIR DEPENDENT TRAINING ON TIMIT CORPUS.

SNR		−12 dB	−9 dB	−6 dB	−3 dB	0 dB	3 dB	6 dB
M+M	Noisy	0.43	0.50	0.58	0.66	0.74	0.81	0.87
	Mapping	0.50	0.57	0.64	0.71	0.76	0.79	0.81
	Masking	0.51	0.58	0.65	0.72	0.77	0.82	0.86
	MRS	0.50	0.59	0.67	0.74	0.80	0.84	0.87
F+M	Noisy	0.54	0.59	0.64	0.70	0.76	0.82	0.87
	Mapping	0.58	0.64	0.69	0.73	0.76	0.78	0.79
	Masking	0.66	0.72	0.77	0.81	0.84	0.86	0.88
	MRS	0.67	0.73	0.78	0.82	0.85	0.87	0.88
F+F	Noisy	0.54	0.60	0.67	0.74	0.80	0.86	0.91
	Mapping	0.52	0.57	0.63	0.68	0.72	0.76	0.77
	Masking	0.50	0.57	0.64	0.70	0.75	0.77	0.79
	MRS	0.51	0.58	0.66	0.72	0.75	0.78	0.79
M+F	Noisy	0.48	0.54	0.60	0.67	0.74	0.80	0.86
	Mapping	0.61	0.68	0.74	0.78	0.81	0.84	0.86
	Masking	0.59	0.66	0.72	0.78	0.83	0.87	0.90
	MRS	0.60	0.67	0.73	0.79	0.84	0.88	0.90
AVR	Noisy	0.50	0.56	0.62	0.69	0.76	0.82	**0.88**
	Mapping	0.55	0.61	0.68	0.73	0.76	0.79	0.81
	Masking	**0.57**	0.63	0.70	0.75	0.80	0.83	0.86
	MRS	**0.57**	**0.64**	**0.71**	**0.77**	**0.81**	**0.84**	0.86

TABLE V
STOI COMPARISON BETWEEN MAPPING-, MASKING-, AND MRS-BASED SPEECH SEPARATION METHODS WITH TARGET INDEPENDENT TRAINING ON IEEE-TIMIT CORPUS.

SNR		−12 dB	−9 dB	−6 dB	−3 dB	0 dB	3 dB	6 dB
M+F	Noisy	0.50	0.56	0.62	0.68	0.74	0.80	0.85
	Mapping	0.73	0.78	0.81	0.85	0.88	0.90	0.92
	Masking	0.73	0.78	0.82	0.86	0.89	0.92	0.94
	MRS	0.77	0.81	0.85	0.88	0.91	0.93	0.95
F+M	Noisy	0.48	0.54	0.61	0.68	0.75	0.81	0.86
	Mapping	0.72	0.76	0.80	0.84	0.87	0.90	0.93
	Masking	0.69	0.75	0.80	0.85	0.88	0.91	0.94
	MRS	0.74	0.79	0.84	0.88	0.91	0.93	0.95
AVR	Noisy	0.49	0.55	0.61	0.68	0.74	0.80	0.85
	Mapping	0.73	0.77	0.81	0.84	0.88	0.90	0.92
	Masking	0.71	0.77	0.81	0.85	0.89	0.92	0.94
	MRS	**0.75**	**0.80**	**0.84**	**0.88**	**0.91**	**0.93**	**0.95**

interfering speakers in the test set were different from those in the training set, but the target speakers of the training and test corpora are the same. Also, SNR levels of the test corpus are different from those of the training corpus.

1) Experimental Settings: We used the IEEE corpus as the source of target speakers [13] and TIMIT as the source of interfering speakers. We call this the IEEE-TIMIT corpus. The IEEE corpus has one male speaker and one female speaker. Each speaker utters 720 clean utterances. We formed two speech separation tasks: one task used the male speaker as the target speaker, and the other one used the female speaker as the target speaker.

Each task had one training set. The training set had 6000 mixed signals with the SNR in dB varying in the range of $[-13, -11, -10, -8, -7, -5, -4, -2, -1, 1, 2, 4, 5, 7, 8, 9, 10]$. The utterance of an target speaker in a mixed signal was randomly selected from the first 640 utterances of the speaker. The utterance of an interfering speaker in a mixed signal was randomly selected from the 6300 utterances of the entire TIMIT dataset.

Each task had 7 test sets with the SNR levels ranging at -12, -9, -6, -3, 0, 3, and 6 dB. Given the target speaker of a task, the interfering speaker in the test sets was the other speaker in the IEEE corpus. Each test set had 80 mixed signals, and each component of a mixture was a clean utterance selected from the last 80 clean utterances of its corresponding speaker.

The rest of the experimental settings follows that described in Section IV-A.

2) Results: Table V lists the comparison results on the IEEE-TIMIT corpus. From the table, we observe the following results. (i) All methods improve the STOI score over the original mixed signals significantly. (ii) The MRS-based method outperforms the mapping- and masking-based methods at all SNR levels. (iii) The mapping- and masking-based methods perform equivalently between -9 dB and 0 dB. But the mapping-based method outperforms the masking-based method at -12 dB, whereas the masking-based method outperforms the mapping-based method at 3 dB and 6 dB. The comparative performances of mapping and masking are consistent with our analysis in Section III.

Comparing Table V with Tables I and III, we find that even if the interfering speakers are unseen during training, target dependent training can still reach a similar performance to that of speaker-pair dependent training. This demonstrates the strong generalization of the DNN-based speech separation methods.

3) Effects of Number of Training Utterances of Target Speaker: From the experimental results on TIMIT, we see that when the clean utterances of the target speaker are limited, the performance improvement of all DNN-based methods is limited. In this subsection, we examine how this factor affects the separation performance.

We constructed 5 training sets for each target speaker of the IEEE-TIMIT corpus in the same way as described above,

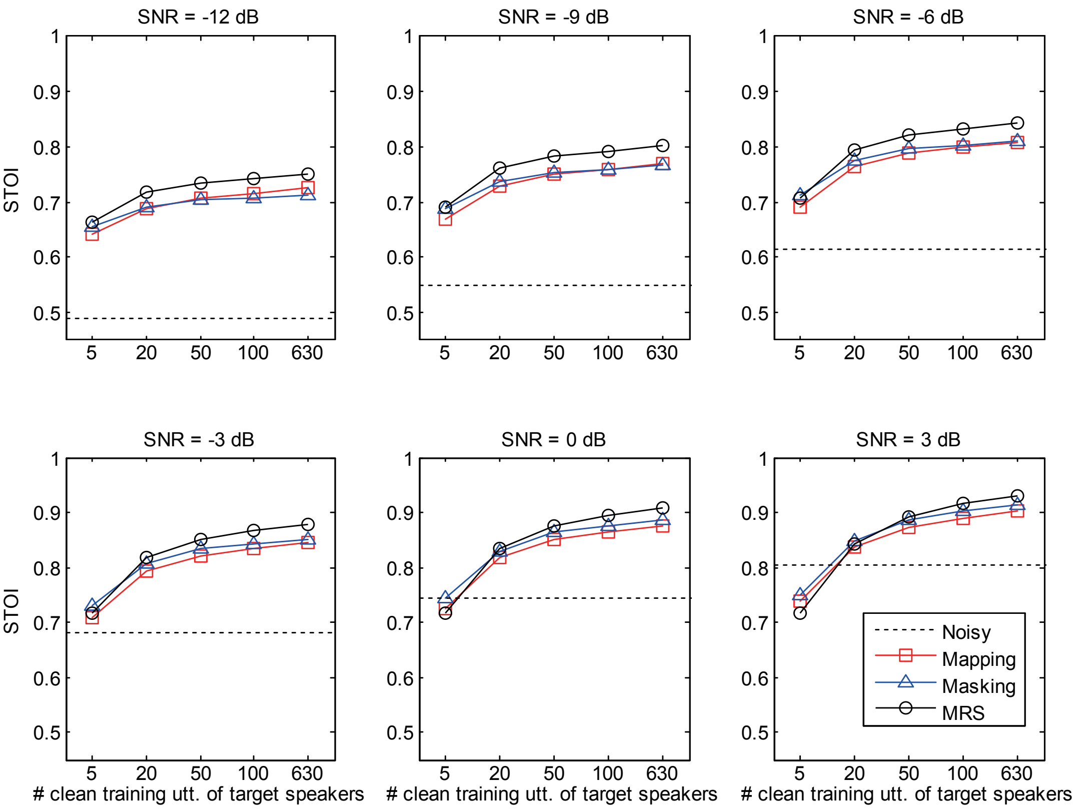

Fig. 4. Comparison of mapping-, masking-, and MRS-based methods with respect to the number of the utterances of the target speaker in training.

except for the only difference that the 6,000 mixed signals of each training set were generated from 5, 20, 50, 100, and 640 clean utterances of the target speaker. Fig. 4 shows the average results on the two separation tasks at various SNR levels. From the figure, we observed that (i) the MRS-based method outperforms the mapping- and masking-based methods, particularly at the low SNR levels; (ii) when the SNR is lower than -3 dB, the mapping- and masking-based methods perform about the same; (iii) when the SNR is higher than -3 dB, the masking-based method performs slightly better than the mapping-based method; (iv) consistent with our analysis, the masking-based method performs relatively better with fewer target training utterances; (v) the effects of the number of target training utterances weaken with the decrease of the SNR.

V. Concluding remarks

In this paper, we have proposed a deep ensemble learning algorithm—multi-resolution stacking—for speech separation. MRS is a stack of DNN ensembles. Each DNN model in a module of the stack takes the concatenation of original acoustic features and the estimated masks from its lower module as the input, and takes the ideal ratio mask as the training objective. The DNN models in the same module have different resolutions (i.e. window lengths), so as to capture different contextual information. MRS improves the accuracy of DNN-based mask estimation by ensembling and stacking multiple DNNs, and enlarges the diversity between the DNNs by expanding the training features.

We have compared the two commonly adopted training objectives for DNN-based speech separation—masking and mapping—systematically. We have found that (i) masking is more effective than mapping in utilizing clean training utterances of a target speaker, and therefore masking-based methods are more likely to achieve better performance when a target speaker has a limited number of training utterances. (ii) masking is more sensitive to the SNR variation of a training corpus than mapping, and masking-based methods are more likely to perform worse at low SNRs in the test stage when the SNR of the training corpus varies in a wide range.

To evaluate the proposed MRS and the differences between mapping and masking, we trained the mapping-, masking-, and MRS-based methods in three conditions, i.e. single-SNR speaker-pair dependent training, multi-SNR speaker-pair dependent training, and target dependent training. After testing hundreds of DNN models, we have observed that the MRS-based method outperforms the mapping- and masking-based methods uniformly, and the relative performances between the mapping- and masking-based methods are consistent with our analysis.

Acknowledgments

We thank Yuxuan Wang for providing his DNN code, Ke Hu for helping with the SSC, TIMIT, and IEEE corpora, and Jun Du, Yong Xu, and Yanhui Tu for assistance in using their code. We also thank the Ohio Supercomputing Center for providing computing resources. The research was supported in part by an AFOSR grant (FA9550-12-1-0130) and an NIDCD grant (R01 DC012048).

References

[1] J. Chen, Y. Wang, and D. L. Wang, "A feature study for classification-based speech separation at very low signal-to-noise ratio," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1993–2002, 2014.

[2] M. Cooke and T.-W. Lee, "Speech separation challenge," http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm, 2006.

[3] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8609–8613.

[4] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. W. Senior, P. A. Tucker *et al.*, "Large scale distributed deep networks." in *Adv. Neural Inform. Process. Sys.*, 2012, pp. 1232–1240.

[5] T. G. Dietterich, "Ensemble methods in machine learning," *Multiple Classifier Sys.*, pp. 1–15, 2000.

[6] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. Int. Conf. Signal Process.* IEEE, 2014, pp. 473–477.

[7] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," *NTIS order number PB91-100354*, 1993.

[8] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, 2015.

[9] K. Han and D. L. Wang, "A classification based approach to speech segregation," *J. Acoust. Soc. Am.*, vol. 132, no. 5, pp. 3475–3483, 2012.

[10] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2014, pp. 4628–4632.

[11] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2014, pp. 1562–1566.

[12] ——, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *arXiv preprint arXiv:1502.04149*, pp. 1–12, 2015.

[13] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust*, vol. 17, no. 3, pp. 225–246, 1969.

[14] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 625–638, 2009.

[15] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1486–1494, 2009.

[16] T. May and T. Dau, "Computational speech segregation based on an auditory-inspired modulation analysis," *J. Acoust. Soc. Am.*, vol. 136, no. 6, pp. 3350–3359, 2014.

[17] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, pp. 1486–1501, 2006.

[18] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Machine Learn.*, 2013, pp. 1–8.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[20] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *Proc. Int. Sym. Chinese Spoken Lang. Process.* IEEE, 2014, pp. 250–254.

[21] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech Separation by Humans and Machines*, vol. 60, pp. 63–64, 2005.

[22] Y. Wang, "Supervised speech separation using deep neural networks," Ph.D. dissertation, The Ohio State University, May 2015.

[23] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[24] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, 2013.

[25] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, 2014.

[26] ——, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, 2015.

[27] X.-L. Zhang and D. L. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Proc. Interspeech*, 2014, pp. 1534–1538.

[28] ——, "Multi-resolution stacking for speech separation based on boosted DNN," in *Proc. Interspeech*, 2015, in press.