Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation

Po-Sen Huang, Member, IEEE, Minje Kim, Member, IEEE, Mark Hasegawa-Johnson, Member, IEEE, and Paris Smaragdis, Fellow, IEEE

Abstract—Monaural source separation is important for many real world applications. It is challenging in that, given only single channel information is available, there is an infinite number of solutions without proper constraints. In this paper, we explore joint optimization of masking functions and deep recurrent neural networks for monaural source separation tasks, including the monaural speech separation task, monaural singing voice separation task, and speech denoising task. The joint optimization of the deep recurrent neural networks with an extra masking layer enforces a reconstruction constraint. Moreover, we explore a discriminative training criterion for the neural networks to further enhance the separation performance. We evaluate our proposed system on TSP, MIR-1K, and TIMIT dataset for speech separation, singing voice separation, and speech denoising tasks, respectively. Our approaches achieve 2.30–4.98 dB SDR gain compared to NMF models in the speech separation task, 2.48 dB GNSDR gain and 4.32–5.42 dB GSIR gain compared to previous models in the singing voice separation task, and outperform NMF and DNN baseline in the speech denoising task.

Index Terms—Monaural Source Separation, Time-Frequency Masking, Deep Recurrent Neural Network, Discriminative Training

I. INTRODUCTION

SOURCE separation are problems in which several signals have been mixed together and the objective is to recover the original signals from the combined signal. Source separation is important for several real-world applications. For example, the accuracy of chord recognition and pitch estimation can be improved by separating singing voice from music [1]. The accuracy of automatic speech recognition (ASR) can be improved by separating noise from speech signals [2]. Monaural source separation, i.e., source separation from monaural recordings, is more challenging in that, without prior knowledge, there are an infinite number of solutions given only single channel information is available. In this paper, we focus on source separation from monaural recordings for applications of speech separation, singing voice separation, and speech denoising tasks.

Several different approaches have been proposed to address the monaural source separation problem. We can categorize them into domain-specific and domain-agnostic approaches. For domain-specific approach, models are designed according to the prior knowledge and assumption of the tasks. For example, in the singing voice separation task, several approaches have been proposed to utilize the assumption of the low rank and sparsity of the music and speech signals, respectively [1], [3]–[5]. In the speech denoising task, spectral subtraction [6] subtracts a short-term noise spectrum estimate to generate the spectrum of clean speech. By assuming the underlying properties of speech and noise, statistical model-based methods infer speech spectral coefficients given noisy observations [7]. However, in real-world scenarios, these strong assumptions may not always be valid. For example, in the singing voice separation task, the drum sounds may lie in the sparse subspace instead of being low rank. In the speech denoising task, the models often fail to predict the acoustic environments due to the non-stationary nature of noise.

For domain-agnostic approach, models are learned from data directly and can be expected to apply equally well to different domains. Non-negative matrix factorization (NMF) [8] and probabilistic latent semantic indexing (PLSI) [9], [10] learn the non-negative reconstruction bases and weights of different sources and use them to factorize time-frequency spectral representations. NMF and PLSI can be viewed as a linear transformation of the given mixture features (e.g. magnitude spectra) during prediction time. Moreover, by the minimum mean square estimate (MMSE) criterion, $E[Y|X]$ is a linear model if $Y$ and $X$ are jointly Gaussian, where $X$ and $Y$ are the separated and mixture signals, respectively. In real-world scenarios, since signals might not always be Gaussian, we often consider the mapping relationship between mixture signals and different sources as a nonlinear transformation, and hence non-linear models such as neural networks are desirable.

Recently, deep learning based methods have been used in many applications, including automatic speech recognition [11], image classification [12], etc. Deep learning based methods have also started to draw attention from the source separation research community by modeling the nonlinear mapping relationship between input and output. Previous work on deep learning based source separation can be further categorized into three ways: (1) Given a mixture signal, deep neural networks predict one of the sources. Maas et al. proposed using a deep recurrent neural network (DRNN) for robust automatic speech recognition tasks [2]. Given noisy features, the authors apply a DRNN to predict clean speech features.
Xu et al. proposed a deep neural network (DNN)-based speech enhancement system, including global variance equalization and noise-aware training, to predict clean speech spectra for speech enhancement tasks [13]. Weninger et al. [14] trained two long-short term memory (LSTM) RNNs for predicting speech and noise, respectively. Final prediction is made by creating a mask out of the two source predictions, which eventually masks out the noise part from the noisy spectrum. Liu et al. explored using a deep neural network for predicting clean speech signals in various denoising settings [15]. These approaches, however, only model one of the mixture signals, which is less optimal compared to a framework that models all sources together. (2) Given a mixture signal, deep neural networks predict the time-frequency mask between the two sources. In the ideal binary mask estimation task, Nie et al. [16] utilized deep stacking networks with time series inputs and sources. In the ideal binary mask estimation task, Nie et al. [16] proposed a two-stage framework using deep neural networks. In the first stage, the authors use $d$ neural networks to predict each output dimension separately, where $d$ is the target feature dimension; in the second stage, a classifier (one layer perceptron or an SVM) is used for refining the prediction given the output from the first stage. The proposed framework is not scalable when the output dimension is high and there are redundancies between the neural networks in neighboring frequencies. Wang et al. [17] recently proposed using deep neural networks to train different targets, including ideal ratio mask and FFT-mask, for speech separation tasks. These mask-based approaches focus on predicting the masking results of clean speech, instead of considering multiple sources simultaneously. (3) Given mixture signals, deep neural networks predict two different sources. Tu et al. proposed modeling two sources as the output targets for a robust ASR task [20]. However, the constraint that the sum of two different sources is the original mixture is not considered. Grais et al. [21] proposed using a deep neural network to predict two scores corresponding to the probabilities of two different sources respectively for a given frame of normalized magnitude spectrum.

In this paper, we further extend our previous work in [22] and [23] and propose a general framework for the monaural source separation task for speech separation, singing voice separation, and speech denoising. Our proposed framework models two sources simultaneously and jointly optimizes time-frequency masking functions together with the deep recurrent networks. The proposed approach directly reconstructs the prediction of two sources. In addition, given that there are two competing sources, we further propose a discriminative training criterion for enhancing source to interference ratio.

The organization of this paper is as follows: Section II introduces the proposed methods, including the deep recurrent neural networks, joint optimization of deep learning models and a soft time-frequency masking function, and the training objectives. Section II presents the experimental setting and results using the TSP, MIR-1K, and TIMIT dataset for speech separation, singing voice separation, and speech denoising task, respectively. We conclude the paper in Section IV.

II. PROPOSED METHODS

A. Deep Recurrent Neural Networks

To capture the contextual information among audio signals, one way is to concatenate neighboring features together as input features to the deep neural network. However, the number of parameters increases proportionally to the input dimension and the number of neighbors in time. Hence, the size of the concatenating window is limited. A recurrent neural network (RNN) can be considered as a DNN with indefinitely many layers, which introduce the memory from previous time steps. The potential weakness for RNNs is that RNNs lack hierarchical processing of the input at the current time step. To further provide the hierarchical information through multiple time scales, deep recurrent neural networks (DRNNs) are explored [24], [25]. DRNNs can be explored in different schemes as shown in Figure 1. The left of Figure 1 is a standard RNN, folded out in time. The middle of Figure 1 is an $L$ intermediate layer DRNN with temporal connection at the $l$-th layer. The right of Figure 1 is an $L$ intermediate layer DRNN with full temporal connections (called stacked RNN (sRNN) in [25]). Formally, we can define different schemes of DRNNs as follows. Suppose there is an $L$ intermediate layer DRNN with the recurrent connection at the $l$-th layer, the $l$-th hidden activation at time $t$ is defined as:

$$h_t^l = f_h(x_t, h_{t-1}^l) = \phi_l(U^h h_{t-1}^l + W^h \phi_{l-1}(W^{l-1} (\ldots \phi_1(W^1 x_t))))$$ (1)

and the output, $y_t$, can be defined as:

$$y_t = f_o(h_t^L) = W^o \phi_{L-1}(W^{L-1} (\ldots \phi_1(W^1 h_t^L)))$$ (2)

where $x_t$ is the input to the network at time $t$, $\phi_l$ is an element-wise nonlinear function, $W^l$ is the weight matrix for the $l$-th layer, and $U^l$ is the weight matrix for the recurrent connection at the $l$-th layer. The output layer is a linear layer.

The stacked RNNs have multiple levels of transition functions, defined as:

$$h_t^l = f_h(h_{t-1}^l, h_{t-1}^l) = \phi_l(U^h h_{t-1}^l + W^h h_{t-1}^l)$$ (3)

where $h_t^l$ is the hidden state of the $l$-th layer at time $t$. $U^l$ and $W^l$ are the weight matrices for the hidden activation at time $t - 1$ and the lower level activation $h_{t-1}^l$, respectively. When $l = 1$, the hidden activation is computed using $h_0^1 = x_t$.

Function $\phi_l(\cdot)$ is a nonlinear function, and we empirically found that using the rectified linear unit $f(x) = \max(0, x)$ [26] performs better compared to using a sigmoid or tanh function. For a DNN, the temporal weight matrix $U^l$ is a zero matrix.

B. Model Architecture

At time $t$, the training input, $x_t$, of the network is the concatenation of features from a mixture within a window. We use magnitude spectra as features in this paper. The output targets, $y_{x_1}$, and $y_{x_2}$, and output predictions, $\hat{y}_{x_1}$, and $\hat{y}_{x_2}$, of the network are the magnitude spectra of different sources.

IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, FEBRUARY 2015 2
Since our goal is to separate one of the sources from a mixture, instead of learning one of the sources as the target, we adapt the framework from [27] to model all different sources simultaneously. Figure 2 shows an example of the architecture.

Moreover, we find it useful to further smooth the source separation results with a time-frequency masking technique, for example, binary time-frequency masking or soft time-frequency masking [1], [27]. The time-frequency masking function enforces the constraint that the sum of the prediction results is equal to the original mixture.

Given the input features, \( x_t \), from the mixture, we obtain the output predictions \( \hat{y}_1, \) and \( \hat{y}_2 \), through the network. The soft time-frequency mask \( m_t \) is defined as follows:

\[
m_t(f) = \frac{|\hat{y}_1(f)|}{|\hat{y}_1(f)| + |\hat{y}_2(f)|}
\]

where \( f \in \{1, \ldots, F\} \) represents different frequencies.

Once a time-frequency mask \( m_t \) is computed, it is applied to the magnitude spectra \( z_t \) of the mixture signals to obtain the estimated separation spectra \( \hat{s}_1, \) and \( \hat{s}_2, \) which correspond to sources 1 and 2, as follows:

\[
\hat{s}_1(f) = m_t(f)z_t(f) \\
\hat{s}_2(f) = (1 - m_t(f))z_t(f)
\]

where \( f \in \{1, \ldots, F\} \) represents different frequencies.

The time-frequency masking function can be viewed as a layer in the neural network as well. Instead of training the network and applying the time-frequency masking to the results separately, we can jointly train the deep learning models with the time-frequency masking functions. We add an extra layer to the original output of the neural network as follows:

\[
\tilde{y}_1 = \frac{|\hat{y}_1|}{|\hat{y}_1| + |\hat{y}_2|} \odot z_t \\
\tilde{y}_2 = \frac{|\hat{y}_2|}{|\hat{y}_1| + |\hat{y}_2|} \odot z_t
\]

where the operator \( \odot \) is the element-wise multiplication (Hadamard product).

In this way, we can integrate the constraints to the network and optimize the network with the masking function jointly. Note that although this extra layer is a deterministic layer, the network weights are optimized for the error metric between \( \tilde{y}_1, \tilde{y}_2, \) and \( y_1, y_2, \) using back-propagation. The time domain signals are reconstructed based on the inverse short time Fourier transform (ISTFT) of the estimated magnitude spectra along with the original mixture phase spectra.

### C. Training Objectives

Given the output predictions \( \tilde{y}_1, \) and \( \tilde{y}_2, \) (or \( y_1, \) and \( y_2, \)) of the original sources \( y_1, \) and \( y_2, \) we explore optimizing neural network parameters by minimizing the squared error,
as follows:

$$J_{MSE} = \|\tilde{y}_1 - y_1\|^2 + \|\tilde{y}_2 - y_2\|^2$$  \hspace{1cm} (7)

Eq. (7) measures the difference between predicted and actual targets. When target samples are similar, it is possible for the DNN to minimize Eq. (7) by being too conservative: when a feature could be attributed to either source 1 or source 2, the neural network attributes it to both. The conservative strategy is effective in training, but leads to reduced SIR (signal-to-interference ratio) in testing, as the network allows ambiguous spectral features to bleed through partially from one source to the other. Interference can be reduced, possibly at the cost of increased artifact, by the use of a discriminative network training criterion. For example, suppose that we define

$$J_{DIS} = -(1 - \gamma) \ln p_{12}(y) - \gamma D(p_{12}||p_{21})$$  \hspace{1cm} (8)

where $0 \leq \gamma \leq 1$ is a regularization constant. $p_{12}(y)$ is the likelihood of the training data under the assumption that the neural net computes the MSE estimate of each feature vector (i.e., its conditional expected value given knowledge of the mixture), and that all residual noise is Gaussian with unit covariance, thus

$$\ln p_{12}(y) = -\frac{1}{2} \sum_{t=1}^{T} \left( \|y_1 - \tilde{y}_1\|^2 + \|y_2 - \tilde{y}_2\|^2 \right)$$  \hspace{1cm} (9)

The discriminative term, $D(p_{12}||p_{21})$, is a point estimate of the KL divergence between the likelihood model $p_{12}(y)$ and the model $p_{21}(y)$, where the latter is computed by swapping affiliation of spectra to sources, thus

$$D(p_{12}||p_{21}) = -\frac{1}{2} \sum_{t=1}^{T} \left( \|y_1 - \tilde{y}_2 - \tilde{y}_1\|^2 - \|y_1 - \tilde{y}_1\|^2 - \|y_2 - \tilde{y}_2\|^2 \right)$$  \hspace{1cm} (10)

Combining Eqs. (8)–(10) gives a discriminative criterion with a simple and useful form:

$$J_{DIS} = \|y_1 - \tilde{y}_1\|^2 + \|y_2 - \tilde{y}_2\|^2 - \gamma \|y_1 - \tilde{y}_2\|^2 - \gamma \|y_2 - \tilde{y}_1\|^2$$  \hspace{1cm} (11)

III. **Experiments**

In this section, we evaluate our proposed models on a speech separation task, and a singing voice separation task, and a speech denoising task. The source separation evaluation is measured by using three quantitative values: Source to Interference Ratio (SIR), Source to Artifacts Ratio (SAR), and Source to Distortion Ratio (SDR), according to the BSS- Eval metrics \[28\]. Higher values of SDR, SAR, and SIR represent better separation quality. The suppression of interference is reflected in SIR. The artifacts introduced by the separation process are reflected in SAR. The overall performance is reflected in SDR. For speech denoising task, we additionally compute the short-time objective intelligibility measure (STOI) which is a quantitative estimate of the intelligibility of the denoised speech \[29\]. We use the abbreviations DRNN-k, sRNN to denote the DRNN with the recurrent connection at the $k$-th hidden layer, or at all hidden layers, respectively. We select the models based on the results on the development set. We optimize our models by back-propagating the gradients with respect to the training objectives. The limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm is used to train the models from random initialization. An example of the separation results is shown in Figure 5. The sound examples are available online.

A. **Speech Separation Setting**

We evaluate the performance of the proposed approaches for monaural speech separation using the TSP corpus \[31\]. In the TSP dataset, we choose four speakers, FA, FB, MC, and MD, from the TSP speech database. After concatenating all 60 sentences per each speaker, we use 90% of the signal for training and 10% for testing. Note that in the neural network experiments, we further divide the training set into 8:1 to set aside 10% of the data for validation. The signals are downsampled to 16kHz, and then transformed with 1024 point DFT with 50% overlap for generating spectra. The neural networks are trained on three different mixing cases: FA versus MC, FA versus FB, and MC versus MD. Since FA and FB are female speakers while MC and MD are male, the latter two cases are expected to be more difficult due to the similar frequency ranges of the same gender. After normalizing the signals to have 0 dB input SNR, the neural networks are trained to learn the mapping between an input mixture spectrum and the corresponding pair of clean spectra.

As for the NMF experiments, 10 to 100 speaker-specific basis vectors are trained from the training part of the signal. The NMF separation is done by fixing the the known speakers’ basis vectors during the test NMF procedure while learning the speaker-specific activation matrices.

In the experiments, we explore different ways to input features: spectral and log-mel filterbank features. The spectral representation is extracted using a 1024-point short time Fourier transform (STFT) with 50% overlap. In the speech recognition literature \[32\], the log-mel filterbank is found to provide better results compared to mel-frequency cepstral coefficients (MFCC) and log FFT bins. The 40-dimensional log-mel representation and the first and second order derivative features are used in the experiments.

For neural network training, in order to increase the variety of training samples, we circularly shift (in the time domain) the signals of one speaker and mix them with utterances from the other speaker.

B. **Speech Separation Results**

We use the standard NMF with the generalized KL-divergence metric using 1024-point STFT as our baselines. We report the best NMF results among models with different basis vectors, as shown in the first column of Figure 6. Note that NMF uses spectral features, and hence the results in the second row (log-mel features) of each figure are the same as the first row (spectral features).

\[1\]http://www.ifp.illinois.edu/~huang146/DNN_separation/demo.zip
Fig. 3. (a) The mixture (female (FA) and male (MC) speech) magnitude spectrogram (in log scale) for the test clip in TSP; (b) (d) The ground truth spectrograms for the two sources; (c) (e) The separation results from our proposed model (DRNN-1 + discrim).

Fig. 4. (a) The mixture (singing voice and music accompaniment) magnitude spectrogram (in log scale) for the clip Yifen-2_07 in MIR-1K; (b) (d) The ground truth spectrograms for the two sources; (c) (e) The separation results from our proposed model (DRNN-2 + discrim).

Fig. 5. (a) The mixture (speech and babble noise) magnitude spectrogram (in log scale) for a clip in TIMIT; (b) (d) The ground truth spectrograms for the two sources; (c) (e) The separation results from our proposed model (DNN).

The speech separation results of the cases, FA versus MC, FA versus FB, and MC versus MD, are shown in Figure 6, 7, and 8 respectively. We train models with two hidden layers of 300 hidden units, where the architecture and the hyperparameters are chosen based on the development set performance. We report the results of single frame spectra and log-mel features in the top and bottom rows of Figure 6, 7, and 8 respectively. To further understand the strength of the models, we compare the experimental results in several aspects. In the second and third columns of Figure 6, 7, and 8, we examine the effect of joint optimization of the masking layer using the DNN. Jointly optimizing masking
layer significantly outperforms the cases where a masking layer is applied separately (the second column). In the FA vs. FB case, DNN without joint masking achieves high SAR, but with low SDR and SIR. In the top and bottom rows of Figures 6, 7, and 8, we compare the results between spectral features and log-mel features. In the joint optimization case, (columns 3–10), log-mel features achieve better results compared to spectral features. On the other hand, spectral features achieve better results in the case where DNN is not jointly trained with a masking layer, as shown in the first column. In the FA vs. FB and MC vs. MD cases, the log-mel features outperform spectral features greatly.

Between columns 3, 5, 7, and 9, and columns 4, 6, 8, and 10 of Figures 6, 7, and 8, we make comparisons between different network architectures, including DNN, DRNN-1, DRNN-2, and sRNN. DRNN-2 and sRNN. In many cases, recurrent neural network models (DRNN-1, DRNN-2, or sRNN) outperform DNN. Between columns 3 and 4, columns 5 and 6, columns 7 and 8, and columns 9 and 10 of Figures 6, 7, and 8, we compare the effectiveness of using the discriminative training criterion, i.e., $\gamma > 0$ in Eq. (11). In most cases, SIRs are improved. The results match our expectation when we design the objective function. However, it also leads to some artifacts which result in slightly lower SARs in some cases. Empirically, the value $\gamma$ is in the range of 0.01~0.1 in order to achieve SIR improvements and maintain reasonable SAR and SDR.

Finally, we compare the NMF results with our proposed models with the best architecture using spectral and log-mel features in Figure 9. NMF models learn activation matrices from different speakers and hence perform poorly in the same sex speech separation cases, FA vs. FB and MC vs. MD. Our proposed models greatly outperform NMF models for all three cases. Especially for the FA vs. FB case, our proposed model...
achieve around 5 dB SDR gain compared to the NMF model while maintaining better SIR and SAR.

C. Singing Voice Separation Setting

Our proposed system can be applied to signing voice separation tasks, where one source is the singing voice and the other source is the background music. The goal of the task is to separate singing voice from music recordings.

We evaluate our proposed system using the MIR-1K dataset. A thousand song clips are encoded with a sample rate of 16 KHz, with a duration from 4 to 13 seconds. The clips were extracted from 110 Chinese karaoke songs performed by both male and female amateurs. There are manual annotations of the pitch contours, lyrics, indices and types for unvoiced frames, and the indices of the vocal and non-vocal frames; none of the annotations were used in our experiments. Each clip contains the singing voice and the background music in different channels.

Following the evaluation framework in [3], [4], we use 175 clips sung by one male and one female singer (‘abjones’ and ‘amy’) as the training and development set. The remaining 825 clips of 17 singers are used for testing. For each clip, we mixed the singing voice and the background music with equal energy (i.e., 0 dB SNR).

To quantitatively evaluate source separation results, we report the overall performance via Global NSDR (GNSDR), Global SIR (GSIR), and Global SAR (GSAR), which are the weighted means of the NSDRs, SIRs, SARs, respectively, over all test clips weighted by their length. Normalized SDR (NSDR) is defined as:

\[
\text{NSDR}(\hat{v}, v, x) = \text{SDR}(\hat{v}, v) - \text{SDR}(x, v), \quad (12)
\]

where \( \hat{v} \) is the resynthesized singing voice, \( v \) is the original clean singing voice, and \( x \) is the mixture. NSDR is for estimating the improvement of the SDR between the preprocessed mixture \( x \) and the separated singing voice \( \hat{v} \).

For the neural network training, in order to increase the variety of training samples, we circularly shift (in the time domain) the signals of the singing voice and mix them with the background music. In the experiments, we use magnitude spectra as input features to the neural network. The spectral representation is extracted using a 1024-point short time Fourier transform (STFT) with 50% overlap. Empirically, we found that using log-mel filterbank features or log power spectrum provide worse performance.

D. Singing Voice Separation Results

In this section, we compare different deep learning models from several aspects, including the effect of different input context sizes, the effect of different circular shift steps, the effect of different output formats, the effect of different deep recurrent neural network structures, and the effect of the discriminative training objectives.

For simplicity, unless mentioned explicitly, we report the results using 3 hidden layers of 1000 hidden units with the mean squared error criterion, joint masking training, and 10K samples as the circular shift step size using features with context window size 3.

Table I presents the results of different deep recurrent network with single source as a target and using two sources as targets (with and without joint mask training).

| Model (num. of output sources, joint mask) | GNSDR | GSIR | GSAR |
|------------------------------------------|-------|------|------|
| DNN (1, no)                              | 5.64  | 8.87 | 9.73 |
| DNN (2, no)                              | 6.44  | 9.08 | 11.26|
| DNN (2, yes)                             | 6.93  | 10.99| 10.15|

where the authors in [5] for providing their trained model for comparison.
Table II: MIR-1K Separation Result Comparison for the Effect of Discriminative Training Using Different Architectures. The “discrim” denotes the models with discriminative training.

| Model     | GNSDR | GSIR | GSAR |
|-----------|-------|------|------|
| DNN       | 6.93  | 10.99| 10.15|
| DRNN-1    | 7.11  | 11.74| 9.93 |
| DRNN-2    | 7.27  | 11.98| 9.99 |
| DRNN-3    | 7.14  | 11.48| 10.15|
| sRNN      | 7.09  | 11.72| 9.88 |

| Architecture Comparison |
|-------------------------|
| Model                  | GNSDR | GSIR | GSAR |
| Unsupervised           |       |      |      |
| RPCA [1]               | 3.15  | 4.43 | 11.09|
| RPCA + FASST [2]       | 3.25  | 4.52 | 11.10|
| RPCA + FASST [3]       | 3.84  | 6.22 | 9.19 |
| Supervised             |       |      |      |
| MLRR [3]               | 3.85  | 5.63 | 10.70|
| RNMF [3]               | 4.97  | 7.06 | 10.03|
| DRNN-2                 | 7.27  | 11.98| 9.99 |
| DRNN-2 + discrim       | 7.45  | 13.08| 9.68 |

Table III: MIR-1K Separation Result Comparison Between Our Models and Previous Proposed Approaches. The “discrim” denotes the models with discriminative training.

E. Speech Denoising Setting

Our proposed framework can be extended to a speech denoising task as well, where one source is the clean speech and the other source is the noise. The goal of the task is to separate clean speech from noisy speech. In the experiments, we use magnitude spectra as input features to the neural network. The spectral representation is extracted using a 1024-point short time Fourier transform (STFT) with 50% overlap. Empirically, we found that log-mel filterbank features provide worse performance. We use 2 hidden layers of 1000 hidden units neural networks with the mean squared error criterion, joint masking training, and 10K samples as the circular shift step size. The results of different architectures are shown in Figure 10. We can observe that deep recurrent networks achieve similar results compared to deep neural networks. With discriminative training, though SDRs and SIRs are improved, STOIs are similar and SARS are slightly worse.

To understand the effect of degradation in the mismatch condition, we set up the experimental recipe as follows. We use a hundred utterances,spanning ten different speakers, from the TIMIT database. We also use a set of five noises: Airport, Train, Subway, Babble, and Drill. We generate a number of noisy speech recordings by selecting random subsets of noises and overlaying them with speech signals. We also specify the signal to noise ratio when constructing the noisy mixtures. After we complete the generation of the noisy signals, we split them into a training set and a test set.

F. Speech Denoising Results

In the following experiments, we examine the effect of the proposed methods under different scenarios. We can observe that the recurrent neural network architectures (DRNN-1, DRNN-2, sRNN) achieve similar performance compared to the DNN model. Including discriminative training objectives improves SDR and SIR, but results in slightly degraded SAR and similar STOI values.

To further evaluate the robustness of the model, we examine our model under a variety of situations in which it is presented with unseen data, such as unseen SNRs, speakers and noise types. In Figure 11 we show the robustness of this model under...
various SNRs. The model is trained on 0dB SNR mixtures and it is evaluated on mixtures ranging from 20 dB SNR to -18dB SNR. We compare the results between NMF, DNN without joint optimization of a masking layer, and DNN with joint optimization of a masking layer. In most cases, DNN with joint optimization achieves the best results. For 20 dB SNR case, NMF achieves the best performance. DNN without joint optimization achieves highest SIR in high SNR cases, though SDR/SAR/STOI are lower.

Next we evaluate the robustness of the proposed methods for data which is unseen in the training stage. These tests provide a way of understanding the performance of the proposed approach to work when applied on unseen noise and speakers. We evaluate the models with three different cases: (1) the testing noise is unseen in training, (2) the testing speaker is unseen in training, and (3) both the testing noise and testing speaker are unseen in training stage. For the unseen noise case, we train the model on mixtures with Babble, Airport, Train and Subway noises, and evaluate it on mixtures that include a Drill noise (which is significantly different from the training noises in both spectral and temporal structure). For the unknown speaker case, we hold out some of the speakers from the training data. For the case where both the noise and speaker are unseen, we use the combination of the above.

We compare our proposed approach with NMF model and DNN without joint optimizing the masking layer [15]. These experimental results are shown in Figure 12. For the case where the speaker is unknown, we observe that there is only a mild degradation in performance for all models, which means that the approaches can be easily used in speaker variant situations. With the unseen noise we observe a larger degradation in results, which is expected due to the drastically different nature of the noise type. The result is still good enough compared to other NMF and DNN without joint optimizing the masking function. The result of the case where both the noise and the speaker are unknown, the proposed model performs slightly worse compared to DNN without joint optimization with the masking function. Overall, it suggests

![Fig. 12. Speech denoising experimental results comparison between NMF, NN (without jointly optimizing masking function [15]), and DNN (with jointly optimizing masking function), when used on data that is not represented in training. We show the results of separation with (a) known speakers and noise, (b) with unseen speakers, (c) with unseen noise, and (d) with unseen speakers and noise.](image-url)
that the proposed approach is good at generalizing across speakers.

IV. Conclusion and Future Work

In this paper, we explore different deep learning architectures, including deep neural networks and deep recurrent neural networks for monaural source separation problems. We further enhance the results by jointly optimizing a soft mask layer with the networks and exploring the discriminative training criteria. We evaluate our proposed method for speech separation, singing voice separation, and speech denoising tasks. Overall, our proposed models achieve 2.30–4.98 dB SDR gain compared to the NMF baseline, while maintaining better SIRs and SARs in the TSP speech separation task. In the MIR-1K singing voice separation task, our proposed models achieve 2.30–2.48 dB GNSDR gain and 4.32–5.42 dB GSIR gain, compared to the previous proposed methods, while maintaining similar GSARs. Moreover, our proposed method also outperforms NMF and DNN baseline in various mismatch conditions in the TIMIT speech denoising task. To further improve the performance, one direction is to further explore using long short-term memory (LSTM) to model longer temporal information \( [34] \), which has shown great performance compared to conventional recurrent neural network as avoiding vanishing gradient properties. In addition, our proposed models can also be applied to many other applications such as robust ASR.

ACKNOWLEDGMENT

This research was supported by U.S. ARL and ARO under grant number W911NF-09-1-0383. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575.

REFERENCES

[1] P.-S. Huang, S. D. Chen, P. S. Smaragdis, and M. Hasegawa-Johnson, “Singing-voice separation from monaural recordings using robust principal component analysis,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 57–60.

[2] A. L. Maas, Q. V. Le, T. M. O’Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, “Recurrent neural networks for noise reduction in robust ASR,” in INTERSPEECH, 2012.

[3] P. Sprechmann, A. Bronstein, and G. Sapiro, “Real-time online singing voice separation from monaural recordings using robust low-rank modeling,” in Proceedings of the 15th International Society for Music Information Retrieval Conference, 2012.

[4] Y.-H. Yang, “Low-rank representation of both singing voice and music accompaniment via learned dictionaries,” in Proceedings of the 14th International Society for Music Information Retrieval Conference, November 4-8 2013.

[5] Y.-H. Yang, “On sparse and low-rank matrix decomposition for singing voice separation,” in ACM Multimedia, 2012.

[6] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 27, no. 2, pp. 113–120, Apr 1979.

[7] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[8] D. D. Lee and H. S. Seung, “Learning the parts of objects by nonnegative matrix factorization,” Nature, vol. 401, no. 6755, pp. 788–791, 1999.

[9] T. Hofmann, “Probabilistic latent semantic indexing,” in Proceedings of the international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999, pp. 50–57.

[10] P. Smaragdis, B. Raj, and M. Shashanka, “A probabilistic latent variable model for acoustic modeling,” Advances in models for acoustic processing, NIPS, vol. 148, 2006.

[11] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal Processing Magazine, vol. 29, pp. 82–97, Nov. 2012.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, 2012.

[13] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. PP, no. 99, pp. 1–1, 2014.

[14] F. Weninger, F. Eyben, and B. Schuller, “Single-channel speech separation with memory-enhanced recurrent neural networks,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2014, pp. 3709–3713.

[15] D. Liu, P. Smaragdis, and M. Kim, “Experiments on deep learning for speech denoising,” in Proceedings of the annual conference of the International Speech Communication Association (INTERSPEECH), 2014.

[16] S. Nie, H. Zhang, X. Zhang, and W. Liu, “Deep stacking networks with time series for speech separation,” in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, May 2014, pp. 6667–6671.

[17] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2013.

[18] Y. Wang and D. Wang, “Towards scaling up classification-based speech separation,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 7, pp. 1381–1390, 2013.

[19] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1849–1858, Dec 2014.

[20] Y. Tu, J. Du, Y. Xu, L.-R. Dai, and C.-H. Lee, “Deep neural network based speech separation for robust speech recognition,” in International Symposium on Chinese Spoken Language Processing, 2014.

[21] E. Grais, M. Sen, and H. Erdogan, “Deep neural networks for single channel source separation,” in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, May 2014, pp. 3734–3738.

[22] P.-S. Huang, M. Kim, P. Smaragdis, and M. Hasegawa-Johnson, “Deep learning for monaural speech separation,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 1562–1566.

[23] P.-S. Huang, M. Kim, M. Hasegawa-Johnson and P. Smaragdis, “Singing-voice separation from monaural recordings using deep recurrent neural networks,” in International Society for Music Information Retrieval (ISMIR), 2014.

[24] M. Hermans and B. Schrauwen, “Training and analysing deep recurrent neural networks,” in Advances in Neural Information Processing Systems, 2013, pp. 190–198.

[25] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, “How to construct deep recurrent neural networks,” in International Conference on Learning Representations, 2014.

[26] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in AISTATS (2013), 2011.

[27] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 6667–6671.

[28] E. Grais, M. Sen, and H. Erdogan, “Deep neural networks for single channel source separation,” in Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on, May 2014, pp. 3734–3738.

[29] P. Sprechmann, A. Bronstein, and G. Sapiro, “Real-time online singing voice separation from monaural recordings using robust low-rank modeling,” in Proceedings of the 15th International Society for Music Information Retrieval Conference, 2012.

[30] Y.-H. Yang, “Low-rank representation of both singing voice and music accompaniment via learned dictionaries,” in Proceedings of the 14th International Society for Music Information Retrieval Conference, November 4-8 2013.

[31] Y.-H. Yang, “On sparse and low-rank matrix decomposition for singing voice separation,” in ACM Multimedia, 2012.

[32] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 27, no. 2, pp. 113–120, Apr 1979.

[33] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[34] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, no. 6755, pp. 788–791, 1999.
[31] P. Kabal, “Tsp speech database.”
[32] J. Li, D. Yu, J.-T. Huang, and Y. Gong, “Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM,” in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 131–136.
[33] C.-L. Hsu and J.-S. Jang, “On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, Feb. 2010.
[34] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

Po-Sen Huang Biography text here.

Minje Kim Biography text here.

Mark Hasegawa-Johnson Biography text here.

Paris Smaragdis Biography text here.