MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Phase-sensitive and Recognition-boosted Speech Separation Using Deep Recurrent Neural Networks

Erdogan, H.; Hershey, J.R.; Watanabe, S.; Le Roux, J.

TR2015-031 April 2015

Abstract

Separation of speech embedded in non-stationary interference is a challenging problem that has recently seen dramatic improvements using deep network-based methods. Previous work has shown that estimating a masking function to be applied to the noisy spectrum is a viable approach that can be improved by using a signal-approximation based objective function. Better modeling of dynamics through deep recurrent networks has also been shown to improve performance. Here we pursue both of these directions. We develop a phase-sensitive objective function based on the signal-to-noise ratio (SNR) of the reconstructed signal, and show that in experiments it yields uniformly better results in terms of signal-to-distortion ratio (SDR). We also investigate improvements to the modeling of dynamics, using bidirectional recurrent networks, as well as by incorporating speech recognition outputs in the form of alignment vectors concatenated with the spectral input features. Both methods yield further improvements, pointing to tighter integration of recognition with separation as a promising future direction.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

PHASE-SENSITIVE AND RECOGNITION-BOOSTED SPEECH SEPARATION USING DEEP RECURRENT NEURAL NETWORKS

Hakan Erdogan[⋆†] *John R. Hershey*[⋆] *Shinji Watanabe*[⋆] *Jonathan Le Roux*[⋆]

[⋆] Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA
[†]Sabanci University, Orhanli Tuzla, 34956, Istanbul, Turkey
haerdogan@sabanciuniv.edu, {hershey, watanabe, leroux}@merl.com

ABSTRACT

Separation of speech embedded in non-stationary interference is a challenging problem that has recently seen dramatic improvements using deep network-based methods. Previous work has shown that estimating a masking function to be applied to the noisy spectrum is a viable approach that can be improved by using a signal-approximation based objective function. Better modeling of dynamics through deep recurrent networks has also been shown to improve performance. Here we pursue both of these directions. We develop a phase-sensitive objective function based on the signal-to-noise ratio (SNR) of the reconstructed signal, and show that in experiments it yields uniformly better results in terms of signal-to-distortion ratio (SDR). We also investigate improvements to the modeling of dynamics, using bidirectional recurrent networks, as well as by incorporating speech recognition outputs in the form of alignment vectors concatenated with the spectral input features. Both methods yield further improvements, pointing to tighter integration of recognition with separation as a promising future direction.

Index Terms— speech enhancement, speech separation, deep networks, LSTM, ASR

1. INTRODUCTION

The goal of single-channel speech separation is to recover a target speaker from a mixture of background signals. Whereas speech enhancement focuses on stationary or nearly stationary backgrounds, speech separation refers to the case where the background is highly non-stationary and can contain difficult sources such as music or other speech signals. This problem has traditionally been addressed using model-based approaches, for example based on hidden Markov models (HMMs) [1], or non-negative matrix factorization (NMF) [2] and its extensions [3–5].

Lately, however, there has been increasing interest in purely data-driven discriminative approaches, such as deep neural networks and recurrent neural networks, which perform surprisingly well [6–11] . Supervised learning of time-frequency masks for the noisy spectrum has been investigated in [11–15], using stereo training data in which noisy speech is the input, and a target time-frequency mask based on the corresponding clean speech data forms the output. Subsequent work [6] focused on modeling dynamics well using long short-term memory (LSTM) recurrent neural networks which helped achieve state of the art performance on a difficult task with non-stationary interference. Here, we investigate whether the objective or in other words, the cost function and dynamics can be further improved.

We first explore improvements to the objective function. In previous work [6], the network estimates a filter or frequency-domain masking function that is applied to the noisy spectrum to produce an estimate of the clean spectrum. The objective function computes error in the amplitude spectrum domain between the speech estimate and the clean speech target. The reconstructed speech estimate retains the phase of the noisy input.

However, when noisy phase is used, the phase error interacts with the amplitude, and the best reconstruction in terms of signal-to-noise ratio (SNR) is obtained with amplitudes that differ from the clean speech amplitudes. Here we consider directly using a phase-sensitive objective function based on the error in the complex spectrum, which includes both amplitude and phase error. This allows the estimated amplitudes to compensate for the use of the noisy phases.

To improve the modeling of dynamics, we consider bidirectional recurrent networks, and the use of higher-level language information provided by integration with a speech recognizer. The LSTM recurrent network was used in prior work [6] to provide a causal system that could be used in on-line inference mode for low-latency enhancement. Here we consider whether bidirectional recurrent networks provide additional gains. This also provides a better baseline for investigating the use of higher level information from a language model, obtained by integration with an automatic speech recognizer, since the recognizer also uses bidirectional inference in the form of an utterance-level Viterbi algorithm.

Language models have previously been integrated into model-based speech separation systems [1, 16]. Feed-forward neural networks, in contrast to probabilistic models, support information flow only in one direction, from input to output. This leaves us with a chicken and egg problem. The speech separation network can benefit from the recognized state sequences, and the recognition system can benefit from the output of the speech separation system. In the absence of a fully integrated system, one might envision a system that alternates between separation and recognition in order to obtain benefits in both tasks.

Here we investigate using a noise-robust recognizer as the first pass. The recognized state sequences are combined with noisy features and used as input to a recurrent neural network trained to reconstruct the speech.

We present experiments demonstrating improvements from the phase-sensitive objective function, the use of bidirectional recurrent networks and the integration with the recognizer.

This work was performed while H. Erdogan was on sabbatical at MERL from Sabanci University.

2. SEPARATION WITH TIME-FREQUENCY MASKS

Time-frequency filtering methods estimate a filter or masking function to multiply by the frequency-domain feature representation of the noisy speech, in order to form an estimate of the clean speech.

We consider the complex short-time spectrum of the noisy speech, $y_{f,t}$, the noise, $n_{f,t}$, and the speech, $s_{f,t}$, obtained via discrete Fourier transform of windowed frames of the time-domain signals. Given an estimated masking function $\hat{a}_{f,t}$, the clean speech is estimated by $\hat{s}_{f,t} = \hat{a}_{f,t} y_{f,t}$. In parallel training, the clean and noisy speech signals are provided, and an estimator $\hat{\boldsymbol{a}}(\boldsymbol{y}|\boldsymbol{\theta})$ for the masking function is trained by minimizing an objective function, $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{f,t} D(\hat{a}_{f,t})$ where $\hat{\boldsymbol{a}}$, $\boldsymbol{y}$ denote variables for all time-frequency bins and $\boldsymbol{\theta}$ are parameters. In the rest of this section, we drop f, t and consider a single time-frequency bin.

Various objective functions have been used, of which there are two types: mask approximation (MA) and signal approximation (SA). The MA objective functions compute a target mask $a^*(s, y)$, and then measure the error between the estimated mask and the target mask: $D_{\mathrm{ma}}(\hat{a}) = D(a^*||\hat{a})$. SA objectives measure error between the filtered signal and the target clean speech: $D_{\mathrm{sa}}(\hat{a}) = D(s||\hat{a}y)$.

In mask approximation, a reasonable cost function is the squared error $D_{\mathrm{ma}}(\hat{a}) = |\hat{a} - a^*|^2$ although other error functions can be used. Various "ideal" masks have been used for a^*; the most common have been the so-called "ideal binary mask"(IBM) [17], and the "ideal ratio mask"(IRM) [18]. See Table 1 for the formulas.

Table 1. Various masking functions a for computing a speech estimate $\hat{s} = ay$, their formula in terms of a, and conditions for optimality. In the IBM, $\delta(x)$ is 1 if the expression x is true and 0 otherwise.

target mask/filter	formula	optimality principle
IBM:	$a^{\mathrm{ibm}} = \delta(\lvert s\rvert > \lvert n\rvert)$,	max SNR $a \in \{0, 1\}$
IRM:	$a^{\mathrm{irm}} = \dfrac{\lvert s\rvert}{\lvert s\rvert + \lvert n\rvert}$,	max SNR $\theta_{\mathrm{s}} = \theta_{\mathrm{n}}$,
"Wiener like":	$a^{\mathrm{wf}} = \dfrac{\lvert s\rvert^2}{\lvert s\rvert^2 + \lvert n\rvert^2}$,	max SNR, expected power
ideal amplitude:	$a^{\mathrm{iaf}} = \lvert s\rvert/\lvert y\rvert$,	exact $\lvert\hat{s}\rvert$, max SNR $\theta_{\mathrm{s}} = \theta_{\mathrm{y}}$
phase-sensitive filter:	$a^{\mathrm{psf}} = \dfrac{\lvert s\rvert}{\lvert y\rvert}\cos(\theta)$,	max SNR given $a \in \mathbb{R}$
ideal complex filter:	$a^{\mathrm{icf}} = s/y$,	max SNR given $a \in \mathbb{C}$

The IBM is better than any other binary mask in terms of signal to noise ratio (SNR). But binary masks are not the best one can do. The a^{irm} is a sub-optimal masking function in terms of SNR. It is optimal in the special case that the phase of s and n are the same, but most of the time they are not. The IRM also does not estimate the amplitude $|s|$ well, since $|a^{\mathrm{irm}} y| \neq |s|$. The Wiener filter, $E|s|^2/(E|s|^2 + E|n|^2)$ has optimal average SNR when the average power statistics are computed for a stationary signal. In the case of single frame processing the "Wiener-like" filter $a^{\mathrm{wf}} = |s|^2/(|s|^2 + |n|^2)$ is, like the IRM, sub-optimal. From the standpoint of estimating the amplitude, the optimal filter is $a^{\mathrm{iaf}} = |s|/|y|$ since then $|\hat{s}| = a^{\mathrm{iaf}}|y| = |s|$.

However, the ideal mask in terms of SNR is easy to derive: the filter that maximizes SNR $= \log |s|^2/|\hat{s} - s|^2$, or equivalently which minimizes $D(s||ay) = |ay - s|^2$, is the complex filter $a^{\mathrm{icf}} = s/y$. Estimating this involves phase estimation, which is potentially tricky, and here we restrict ourselves to using $a \in \mathbb{R}$, and keeping the noisy phases. The optimal phase-sensitive filter (PSF) under this constraint is

$$a^{\mathrm{psf}} = \mathrm{Re}\left(\frac{s}{y}\right) = \frac{|s|}{|y|}\mathrm{Re}\left(e^{i(\theta^{\mathrm{s}} - \theta^{\mathrm{y}})}\right) = \frac{|s|}{|y|}\cos(\theta), \qquad (1)$$

where $\theta = \theta^{\mathrm{s}} - \theta^{\mathrm{y}}$.

Rather than using an MA objective function, [6] showed that using a magnitude spectrum approximation (MSA) $D_{\mathrm{msa}}(\hat{a}) = (\hat{a}|y| - |s|)^2$ leads to a significant improvement. Here we propose using a phase-sensitive spectrum approximation (PSA) $D_{\mathrm{psa}}(\hat{a}) = |\hat{a}y - s|^2$, from which we derive the optimal real-valued filter. This distortion measure is equivalent up to an additive constant to using $D_{\mathrm{psa}}(\hat{a}) = (\hat{a}|y| - |s|\cos(\theta))^2$. Despite the use of the MSA/PSA objective function, it is still desirable to have the output of the network be a mask, since then the entire dynamic range of the data does not have to be covered by the output of the network. It is convenient in this context to truncate a to between 0 and 1, to fit the range of a sigmoid unit. In addition, we conjecture that mask prediction would also avoid global variance problems reported in [10].

To investigate the performance of various ideal masking functions and filters, we performed oracle experiments on the CHiME-2 development set [19]. We applied each filter to the noisy data and computed signal to distortion ratio (SDR) results at various input SNR levels. We present our results in Table 2. In these experiments, the oracle phase-sensitive filter significantly outperforms the other filters in terms of SDR. The truncation to the range of $[0, 1]$ causes a loss of 1.59 dB, but the truncated phase-sensitive filter is still significantly better than the other methods.

Table 2. SDR results (in dB) at various SNR levels on the CHiME-2 development (dt) data using various oracle masks.

dt	-6 dB	9 dB	Avg
IBM	14.56	20.89	17.59
IRM	14.13	20.69	17.29
"Wiener-like"	15.20	21.49	18.21
ideal amplitude	13.97	21.35	17.52
phase sensitive filter	17.74	24.09	20.76
truncated PSF	16.13	22.49	19.17

3. LSTM AND BLSTM NETWORKS FOR SEPARATION

For sequential data prediction, recurrent neural networks seem to be the right neural network model since they make use of the context information by connecting hidden nodes to their counterparts in the previous step of the sequence. Alternatively, to enable the use of context in deep feed-forward neural networks, one needs to explicitly concatenate multiple input vectors from neighboring sequence steps and enrich the input data.

However, because of the "vanishing or exploding gradients" problem that can occur during back-propagation through time, it is quite hard to train good performing RNNs for sequence prediction. An old and efficient trick is to introduce LSTM structures in RNNs which help alleviate the gradient problems [20]. LSTM introduces the concept of a "memory cell" with input, output and forget gates which are also basically recurrent units that have outputs in the range $[0, 1]$ and modify the scalars or vectors stored in the cells using the multiplication operation. One can think of LSTM as replacing the hidden nodes of RNNs with memory cells and introducing additional

gates to control the flow of information into and out from the cell. The recurrent connection from each memory cell to itself is just 1 so that the gradient never vanishes or blows up as back-propagation through time is performed [20]. It has been shown that using LSTM structures, one can learn sequential prediction networks which are able to make use of long-term contextual information [20]. In contrast, the performance of deep neural networks on speech separation is suboptimal to LSTM-RNNs as shown in [6].

In this work, we also investigate use of the bidirectional long short-term memory (BLSTM) networks [21]. In BLSTMs, there are recurrent connections in both forward and backward directions. A BLSTM can make use of contextual information from both sides in the sequence. In subsequent work, single directional LSTMs were used to enable real-time performance of the enhancement system [6].

Although the BLSTM neural network makes use of forward and backward context information, it still does not have access to a language model hence cannot explicitly use long-term word-level information. Therefore in the next section we turn to investigating whether additional information can be derived from a speech recognition system. Since speech recognition uses bidirectional inference in the form of the Viterbi algorithm, the BLSTM is an appropriate neural network baseline and has the same algorithmic latency requirements as the speech recognition system.

4. INCORPORATING SPEECH RECOGNITION INFORMATION

Speech enhancement and recognition can be considered as different but related problems. A good speech enhancement system can certainly be used as an input module to a speech recognition system. Conversely, speech recognition might be used to improve speech enhancement since it incorporates additional information such as the language model. However, it is not clear how to build a multi-task recurrent neural network system for both tasks.

In this paper, we simply perform a first step towards the integration of speech recognition and enhancement problems. For each frame of noisy input signal to be enhanced, we use the most likely state or phone that aligns to that frame as obtained from a speech recognizer. We have 2004 tied states and 42 phones[1] in our recognition system for the CHiME-2 dataset. The alignment information is provided as an extra feature added to the input of the LSTM network. We experimented with different kinds of features for the alignment information. First, we used a one-hot representation to indicate the frame-level state or phone. When done for the context-dependent states, this yields a large vector which could pose difficulties for learning. We also experimented with using continuous features derived by averaging the spectral features, aligned to each state or phoneme, calculated from the training data. This yields a shorter input representation and provides some kind of similarity-preserving coding of each state. In addition, the information is in the same domain as the other noisy spectral input which could make it easier for the network to utilize when predicting the mask.

5. EXPERIMENTS AND DISCUSSION

We performed experiments on the 2nd CHiME speech separation and recognition challenge (CHiME-2) medium vocabulary track database [19]. The noises in this database are highly non stationary and extremely challenging. They include TV in the background, noises from household appliances, children talking and making noises while playing and similar real-life living room sounds. The noisy signals are formed by mixing clean speech utterances from the Wall Street Journal (WSJ-0) corpus of read speech with recorded noises at SNR levels -6, -3, 0, 3, 6 and 9 dB. There is a 7138 utterance training set which includes various noisy mixtures, a 2460 utterance development set which is derived from 410 clean speech utterances, each mixed with a noise signal at six different noise levels. Similarly, there is an evaluation set that includes 1980 utterances derived from 330 clean speech signals.

We first measured in Table 1 the performance of the various ideal masks that we introduced in Section 2, providing upper bounds on performance. These results show that, using the phase-sensitive ideal filter, one can achieve about 2 dB higher SDR performance as compared to using the IRM, even with the truncated version. This result encourages us to use the phase-sensitive objective function D_{psa} in training neural networks instead of trying to approximate only the spectral magnitude.

In our experiments, all input vectors were mean-and-variance normalized using the training data statistics. For the alignment vectors, we normalized using clean alignments on the training data. The networks were trained with a momentum-based mini-batch stochastic gradient algorithm. The momentum coefficient was 0.9 and the learning rate was 10^{-6}. Validation cost on the development set was used as a stopping criteria. If the validation cost did not decrease for at least 10 epochs, the training is stopped. A zero-mean and $\sigma = 0.1$ Gaussian noise was added to the training samples to improve robustness. This setup is based on [6].

Our baseline network has two LSTM layers with 256 nodes each, trained using 100-bin log-mel-filterbank features as the input. Prior work [6] showed that careful multi-stage training of the LSTM network is essential to obtain good speech separation performance. Following this recipe, the network was first trained to predict a 100-bin mel-transformed mask output using the mask approximation cost function. In order to map to a full spectrum domain, a sigmoid layer was added, with weights initialized to the Mel-transform's regularized pseudo-inverse [2]. The network training objective function was modified to be the magnitude spectrum approximation cost and all layers of the network continued to be trained with the new cost function. This LSTM-MSA baseline system gave stellar results surpassing the state-of-the-art NMF-based speech enhancement results by at least 2.8 dB consistently in each SNR condition. Thus, our baseline is an extremely strong one which is quite hard to beat.

As ASR-based input features for each frame, we used one-best state-level or phone-level alignments from a noise-robust recognizer, trained on the CHiME-2 training set. We use a multi-stage training approach to incorporate the alignment information. We first train a spectral LSTM network which only uses log-mel-filterbank inputs. After obtaining a well performing network, we add the new alignment inputs to the input layer of the network and initialize the additional input weights to be zero. This ensures that the new network starts exactly from the earlier network and changes the weights according to the network's cost criterion to improve the results. Random initialization of the network did not work as well as this informed initialization.

Our initial experiment involved using clean speech signal's frame-level alignment with the reference transcript both in training and development set. This experiment acted as an oracle upper limit on the performance when the frame-level state information is used. Oracle alignment information aids speech enhancement by provid-

[1] Actual number of phones is higher but we merge similar phones and different varieties of the same phone.

[2] This step is slightly different than the one in [6].

ing the active state at each frame which the network learns to exploit to get a better mask $a_{f,t}$ at each time-frequency bin. The initial features were in the form of a one-hot vector, with one input for each state. However, we expected that this form of state information would not be easy to learn for the network, so we also experimented with adding the state information in the form of an average feature vector of frames that align to a particular state, learned from the training data.

In order to have realistic results, we needed to obtain alignments from noisy data. We performed DNN-based speech recognition on pre-processed speech data enhanced with the LSTM network trained with magnitude DFT features from [6] to obtain the one-best decoded transcriptions. These transcriptions were aligned with the enhanced speech data to obtain noisy alignment features. We further attempted to train the network using these noisy alignment features starting from the network trained with the clean alignment features. The accuracy of noisy alignments in development and evaluation sets is 50-55% for state-level alignments, whereas it is 65-70% for phone-level ones. In our experiments, further training with the noisy alignments did not always reduce the objective function on the validation data. If it was not reduced, we just used the network trained with the clean alignments.

The results using the oracle and noisy alignment information are given in Table 3. In the table, the "input" column describes the input used; 'mfb' stands for log-mel-filterbank features with 100 bins. For the alignment part of the input, we use three symbols to describe the input: 'oa' for oracle alignment versus 'na' for noisy alignment, followed by 'ph' for phone alignment versus 'st' for state alignment, and finally '1h' for one-hot, versus 'sm' for spectral mean and 'spm' for spectral power mean. In the "cost" column, 'MA' stands for mask approximation and 'MSA' stands for magnitude spectrum approximation. We only provide SDR values for -6 and 9 dB SNR cases and the average over all six SNR values for brevity. The different forms of alignment information (phoneme versus state) and features (one-hot versus spectral average) all yield similar results. This shows that the network is using the alignment information in whatever form we provide it. In other experiments, we arbitrarily use the "na,st,sm" variety of the alignment features.

Table 3. SDR results (in dB) on the CHiME-2 development data set using alignment information as additional inputs.

Network	Cost	Input	-6 dB	9 dB	Avg
LSTM 2x256	MA	mfb	8.77	16.71	12.76
LSTM 2x256	MSA	mfb	9.24	16.93	13.03
LSTM 2x256	MSA	mfb+oa,ph,1h	10.00	17.28	13.60
LSTM 2x256	MSA	mfb+oa,st,1h	9.98	17.32	13.58
LSTM 2x256	MSA	mfb+oa,st,sm	10.09	17.33	13.63
LSTM 2x256	MSA	mfb+oa,st,spm	10.17	17.39	13.69
LSTM 2x256	MSA	mfb+na,st,sm	9.64	17.12	13.36
LSTM 2x256	MSA	mfb+na,st,spm	9.59	17.15	13.33

We wanted to see if recognition information would still provide gains after other improvements. We sought to improve the LSTM baseline by considering a bidirectional LSTM network, as well as using a phase-sensitive spectrum approximation (PSA) cost function as discussed in Sections 2 and 3. In Table 4, we observe that BLSTM with 2 layers and 256 nodes at each layer provides 0.14 dB improvement over a similar LSTM network. However since half of the nodes are used for each direction (forward and backward) it makes sense to increase the number of nodes of the BLSTM network. When we used 384 nodes at each layer, we obtain 0.43 dB improvement over the LSTM-MA baseline. With MSA cost, LSTM-MSA baseline still improves by about 0.4 dB when using the BLSTM network. Using the PSA cost and starting the new network from the MSA-based one, the MSA enhancement result can be further improved by about 0.3 dB as can be seen in Table 4. Adding the noisy alignment inputs at this stage still provides about 0.2 dB improvement over the previous best result which constitutes an overall improvement of 0.9 dB over the LSTM-MSA baseline.

We provide the results on the CHiME-2 evaluation set in Table 5 in terms of SDR and speech-to-interference ratio (SIR). The improvements are similar to the development set improvements and we still see a combined improvement of about 0.90 dB in SDR and 2.5 dB in SIR.

Table 4. SDR results (in dB) on the CHiME-2 development data set using BLSTM and PSA.

Network	Cost	Input	-6 dB	9 dB	Avg
LSTM 2x256	MA	mfb	8.77	16.71	12.76
BLSTM 2x256	MA	mfb	8.92	16.83	12.90
BLSTM 2x384	MA	mfb	9.39	16.97	13.19
LSTM 2x256	MSA	mfb	9.24	16.93	13.03
BLSTM 2x384	MSA	mfb	9.76	17.28	13.45
LSTM 2x256	PSA	mfb	9.71	17.09	13.36
BLSTM 2x384	PSA	mfb	10.21	17.43	13.76
LSTM 2x256	MSA	mfb+na,st,sm	9.64	17.92	13.36
BLSTM 2x384	PSA	mfb+na,st,sm	10.50	17.56	13.97

Table 5. SDR results (in dB) on the CHiME-2 evaluation data set.

Network	Cost	Input	Avg-SDR	Avg-SIR
LSTM 2x256	MSA	mfb	13.83	17.53
BLSTM 2x384	MSA	mfb	14.22	18.24
LSTM 2x256	PSA	mfb	14.14	19.20
BLSTM 2x384	PSA	mfb	14.51	19.78
BLSTM 2x384	PSA	mfb+na,st,sm	**14.75**	**20.46**

6. CONCLUSIONS AND FUTURE WORK

We improved the speech separation performance on the CHiME-2 database by about 0.90 dB using (a) bidirectionality of the recurrent network, (b) phase-sensitive spectrum approximation, and (c) incorporating speech recognition alignment information within the LSTM-DRNN framework for speech enhancement. These improvements may be applicable in other speech processing problems such as bandwidth extension and voice conversion. It is interesting to see that the improvements are mostly additive.

In our view, promising future directions include prediction of the target phase rather than using the noisy phase, and tighter integration of language model information in speech separation.

7. ACKNOWLEDGEMENTS

We thank Felix Weninger for providing the computational setup for the experiments in [6] which is used to obtain the baseline results in this paper. The first author is partially funded by The Scientific and Technological Research Council of Turkey (TUBITAK) under the BIDEB-2219 program.

8. REFERENCES

[1] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer Speech and Language*, vol. 24, no. 1, Jan. 2010.

[2] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. of NIPS*, Vancouver, Canada, 2001.

[3] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Proc. of Interspeech*, Makuhari, Japan, 2010.

[4] C. Févotte, J. Le Roux, and J. R. Hershey, "Non-negative dynamical sytem with application to speech and audio," in *Proc. of ICASSP*, 2013.

[5] E. Grais and H. Erdogan, "Regularized nonnegative matrix factorization using Gaussian mixture priors for supervised single channel source separation," *Computer Speech and Language*, vol. 27, no. 3, 2013.

[6] F. J. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *GlobalSIP Machine Learning Applications in Speech Processing Symposium*, 2014.

[7] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. of ICASSP*, Florence, Italy, 2014.

[8] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Proc. of ICASSP*, Florence, Italy, 2014.

[9] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *Signal Processing Letters*, vol. 21, no. 1, 2014.

[10] ——, "Global variance equalization for improving deep neural network based speech enhancement," in *Proc. of ICASSP*, Florence, Italy, 2014.

[11] 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, 2014.

[12] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. of ICASSP*, Vancouver, Canada, 2013.

[13] J. Le Roux, S. Watanabe, and J. Hershey, "Ensemble learning for speech enhancement," in *Proc. of WASPAA*, Oct. 2013.

[14] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. of ICASSP*, Florence, Italy, 2014.

[15] S. Gonzalez and M. Brookes, "Mask-based enhancement for very low quality speech," in *Proc. of ICASSP*, Florence, Italy, 2014.

[16] T. T. Kristjansson, J. R. Hershey, P. A. Olsen, S. J. Rennie, and R. Gopinath, "Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system," in *Proc. Interspeech 2006*, Pittsburgh, 2006.

[17] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Communication*, vol. 51, no. 3, 2009.

[18] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. of ICASSP*. IEEE, 2013.

[19] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. of ICASSP*, Vancouver, Canada, 2013.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, 1997.

[21] M. Wöllmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise," in *Proc. of ICASSP*, Vancouver, Canada, 2013.