ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Combining Bottleneck-BLSTM and Semi-Supervised Sparse NMF for Recognition of Conversational Speech in Highly Instationary Noise

Felix Weninger, Martin Wöllmer, Björn Schuller

Institute for Human-Machine Communication, Technische Universität München, Germany
(weninger|woellmer|schuller)@tum.de

10.21437/Interspeech.2012-108

Abstract

We address the speaker independent automatic recognition of spontaneous speech in highly variable noise by applying semi-supervised sparse non-negative matrix factorization (NMF) for speech enhancement coupled with our recently proposed front-end utilizing bottleneck (BN) features generated by a bidirectional Long Short-Term Memory (BLSTM) recurrent neural network. In our evaluation, we unite the noise corpus and evaluation protocol of the 2011 PASCAL CHiME challenge with the Buckeye database, and we demonstrate that the combination of NMF enhancement and BN-BLSTM front-end introduces significant and consistent gains in word accuracy in this highly challenging task at signal-to-noise ratios from -6 to 9 dB.

1. Introduction

Automatic speech recognition (ASR) in many realistic scenarios, including hands-free natural human-computer interaction and multimedia retrieval, has to deal with spontaneous speech on the one hand, and interfering audio sources on the other hand. As an additional challenge, in many situations, such as analysis of on-line videos, only one audio channel is available. To recognize spontaneous speech in challenging scenarios, extensions of the traditional ASR models have been proposed. Recently, it was shown that ASR front-ends employing so-called bidirectional Long Short-Term Memory (BLSTM) neural networks to generate context-sensitive probabilistic speech features lead to remarkable performance gains in conversational speech recognition [1]. Conversely, to increase robustness against background noise, efforts have been devoted to both robust recognizers and signal enhancement. Looking at previous studies in these fields, we observe that studies on speech enhancement and noise-robust ASR often look at simplistic recognition tasks. For example, the 2011 PASCAL CHiME Challenge data [2] features realistic noise from a domestic environment, yet utterances from the Grid corpus which are strictly adhering to a fixed grammar and a vocabulary of 51 phonetically similar keywords; furthermore, the Challenge task is speaker dependent, allowing to create precise models of how each speaker pronounces each word. A task that is more complex from the ASR point of view is provided by the Aurora 4 data; yet, it relies on read speech from the Wall Street Journal corpus and a very limited amount of noise data that is artificially scaled to produce different SNRs. As of today, systematic studies on recognition of conversational speech at varying levels of background noise are very sparse.

Hence, in this paper, we address the recognition of spontaneous conversational speech from the Buckeye database in realistic, variable and mostly non-stationary noise from a domestic environment as featured in the CHiME Challenge. To foster realism, we restrict ourselves to methods that are applicable to monaural signals. Furthermore, we enforce a stricly speaker independent evaluation for speech enhancement as well as ASR. To address the variability of spontaneous speech as well as non-stationary noise, we propose a combination of monaural speech enhancement by NMF (Section 2) and our Bottleneck Bidirectional Long Short-Term Memory (BN-BLSTM) tandem speech recognizer introduced in [1] (Section 3). We adapt the NMF methodology from our previous study on small vocabulary ASR in CHiME noise [3] to the large vocabulary ASR task at hand: We consider semi-supervised NMF relying on speaker independent phoneme models and unsupervised adaptation to background noise. Details of the experimental setup including the evaluation database are given in Section 4. Component-level evaluation of both speech enhancement and recognition is performed in Section 5 before concluding in Section 6.

2. NMF-based Speech Enhancement

NMF-based techniques for monaural speech enhancement, such as the ones used in this study, are based on the assumption that the wanted speech signal is corrupted by addition of interfering noise:

$$\mathbf{V} = \mathbf{V}^{(s)} + \mathbf{V}^{(n)},$$

where $\mathbf{V} \in \mathbb{R}_+^{M \times N}$ is an observed magnitude spectrogram of speech overlaid by interfering noise, $\mathbf{V}^{(s)}$ is the (true) spectrogram of the speech signal, and $\mathbf{V}^{(n)}$ is the (true) noise spectrogram. Furthermore, we assume that both $\mathbf{V}^{(s)}$ and $\mathbf{V}^{(n)}$ can be approximated as the product of speech and noise dictionaries $\mathbf{W}^{(s)} \in \mathbb{R}_+^{M \times R^{(s)}}$ and $\mathbf{W}^{(n)} \in \mathbb{R}_+^{M \times R^{(n)}}$ with non-negative coefficients (activations) $\mathbf{H}^{(s)} \in \mathbb{R}_+^{R^{(s)} \times N}$, $\mathbf{H}^{(n)} \in \mathbb{R}_+^{R^{(n)} \times N}$:

$$\mathbf{\Lambda} = \mathbf{\Lambda}^{(s)} + \mathbf{\Lambda}^{(n)} = \mathbf{W}^{(s)}\mathbf{H}^{(s)} + \mathbf{W}^{(n)}\mathbf{H}^{(n)}.$$

where $\mathbf{\Lambda}$, $\mathbf{\Lambda}^{(s)}$ and $\mathbf{\Lambda}^{(n)}$ denote approximations of $\mathbf{V}$, $\mathbf{V}^{(s)}$ and $\mathbf{V}^{(n)}$, respectively. In our semi-supervised NMF approach, we estimate a fixed speech dictionary $\mathbf{W}^{(s)}$ from training data as detailed in Section 4.2. In contrast, the noise dictionary $\mathbf{W}^{(n)}$ is estimated for each utterance along with $\mathbf{H}^{(s)}$ and $\mathbf{H}^{(n)}$ by iterative minimization of the following cost function:

$$c(\mathbf{W}^{(n)}, \mathbf{H}) = c_r(\mathbf{W}^{(n)}, \mathbf{H}) + \lambda\, c_s(\mathbf{H}) \qquad (1)$$

where c_r corresponds to the reconstruction error measured as the Kullback-Leibler divergence $d_1(\mathbf{V}|\mathbf{\Lambda})$ and c_s is an additive sparsity constraint; in our study, this simply corresponds to the L1 norm of $\mathbf{H}$, penalizing non-zero entries. For the minimization of (1) the standard multiplicative update NMF algorithm is applied, with a straightforward extension to include the sparsity constraint. Similar semi-supervised NMF approaches have been proven to be highly efficient for speech enhancement, e. g., in [4].

We optimize the value of λ on a held out development set (cf. Section 4.2). Informally, the purpose of sparsity is to force that only a few basis vectors can be active at a given time, which is a reasonable assumption if the basis vectors correspond to, e. g., phonemes, or spectra originating from different noise sources. The update rules are applied for a fixed number of iterations which is optimized on our held out development set (cf. Section 4.2). For speech enhancement, we obtain an estimate of the clean speech spectrogram, $\widehat{\mathbf{V}}^{(s)}$, by element-wise filtering of the observed spectrogram $\mathbf{V}$: $\widehat{\mathbf{V}}^{(s)} = (\mathbf{\Lambda}^{(s)}/\mathbf{\Lambda}) \otimes \mathbf{V}$. All experiments for this paper are based on the NMF implementations found in our open-source toolkit openBliSSART [5] to enforce reproducibility of our results.

3. Bottleneck-BLSTM based Speech Recognition

In tandem ASR systems, the output activations of neural networks trained on phoneme or phoneme state targets are used as probabilistic features, alternatively to (or in combination with) standard MFCC features. For enhanced probabilistic feature generation, standard multilayer perceptrons (MLP) can be replaced by bidirectional Long Short-Term Memory networks [6] which allow to access and model long-range temporal context information via so-called *memory blocks* substituting the conventional neurons in the network's hidden layers. Generally, bidirectional networks consist of two sets of hidden layers, one for forward and one for backward processing. This enables the incorporation of past and future context and captures for example co-articulation effects in human speech (for more details, see [6]). Combining BLSTM based feature generation with the 'bottleneck' idea was shown to lead to lower error rates in spontaneous speech recognition [1]. The bottleneck principle allows to generate tandem feature vectors of arbitrary size by using the activations of a narrow hidden (bottleneck) layer as features – rather than the logarithmized output activations corresponding to the estimated phoneme or phoneme state posteriors.

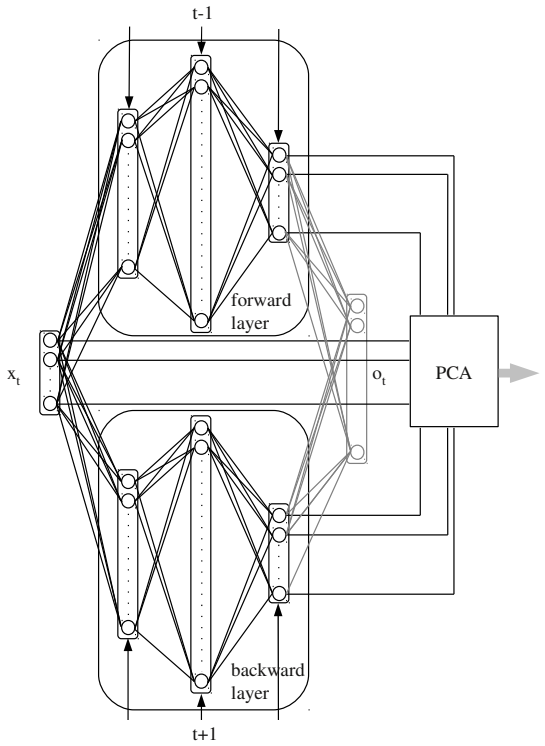
Figure 1 illustrates the detailed structure of the Bottleneck-BLSTM front-end applied for our experiments. Since we focus on *bidirectional* processing, we have two bottleneck layers: one within the network processing the speech sequence in forward direction and one within the network for backward processing. 39 cepstral mean and variance normalized MFCC features (including deltas and double deltas) are extracted from the speech signal every 10 ms using a window size of 25 ms. These features x_t serve as input for a BN-BLSTM network that is trained on framewise phoneme targets. During feature extraction, the activations of the output layer are ignored; only the activations of the forward and backward bottleneck layer are processed (i. e., the memory block outputs of the bottleneck layers). Together with the original MFCC features, the forward and backward bottleneck layer activations are concatenated to one large feature vector which is then decorrelated by Principal Component Analysis (PCA). In Figure 1, the connections between the bottleneck layers and the output layer are depicted in grey, indicating that the activations of the output layer (o_t) are only used during network training and not during BN-BLSTM feature extraction.

4. Experiments

4.1. Evaluation Database

Our choice of evaluation database was motivated by the lack of a noisy spontaneous speech database with a clean reference;

Figure 1: *Architecture of the Bottleneck-BLSTM front-end.*



while the COSINE corpus [7] provides realistic recordings of spontaneous speech in an outside environment, it is less suited to component level evaluation of source separation and ASR since even the close-talk speech in this corpus is corrupted by environmental noise. Thus, we used the Buckeye corpus [8] recorded in clean conditions and mixed with the CHiME noise corpus [2] to simulate spontaneous speech encountered in a noisy domestic environment.

The Buckeye corpus contains recordings of interviews with 40 speakers and was originally intended to study phonetic variation among speakers. The speech is highly spontaneous and contains a variety of non-linguistic vocalizations. Thus, we believe that this corpus is better suited to evaluation of speech separation in real-life conditions than, e. g., the popular TIMIT corpus of read speech, which is characterized by lower variation. Only the subjects' speech is used. The segmentation into utterances and the speaker-independent subdivision into training, development, and test set (stratified by speaker age and gender) exactly corresponds to the ASR experiments reported in [9].

The additive noise considered in this study is taken from the corpus of the 2011 PASCAL CHiME Challenge [2]. This corpus contains genuine recordings from a domestic environment obtained over a period of several weeks. Most of the noise is highly non-stationary due to abrupt changes such as appliances being turned on/off, impact noises such as banging doors, and interfering speakers; more details can be found in [2]. To create the noisy version of our evaluation database, we followed the protocol which was used to create the CHiME Challenge ASR task [2]: In the development and test set, we employ six signal-to-noise ratios (SNRs) ranging from 9 dB down to -6 dB in steps of 3 dB. After normalizing the speech signals to -6 dB maximum amplitude to avoid clipping after mixing with noise, we chose for each speech signal six noise segments from the CHiME develop-

ment/test noise matching the different SNRs. As proposed in [2], the noisy utterances are not constructed by artificial scaling of the speech or noise amplitudes, but by choosing noise segments as they were recorded in a real life situation. This means that noisy utterances at low SNRs occur in noise that naturally has high energy, such as broad band impact noise. The SNRs were measured on first order differences of speech and noise signals.

In addition, we created a multi-condition training set by mixing clean training speech with random segments of the six hours of training noise (disjoint from development and test noise) provided with the CHiME Challenge corpus. For this multi-condition training set, we added random segments of noise with the normalized speech utterances; this provides a good coverage of SNRs while not assuming any knowledge about the exact SNRs occuring in the test conditions. For the experiments reported in this paper, all signals were downmixed to mono by averaging channels.

4.2. Sparse Semi-Supervised NMF

To apply NMF on the development and test set, spectrograms of the signals were calculated by short-time Fourier Transform (STFT) using Hann windows of 25 ms length at 10 ms frame shift. A shorter window size and frame shift than in our previous study on the small vocabulary CHiME Challenge ASR task [3] have been chosen to cope with higher variability of spontaneous conversational speech.

To build a phoneme-dependent yet speaker-independent speech model for NMF, for each phoneme, the corresponding spectrograms were extracted from the Buckeye training set according to a forced alignment with the recognizer described in [1]. These concatenated phoneme spectrograms were reduced to a single dictionary atom by a 1-component NMF. The column-wise concatenation of these atoms builds the matrix $\mathbf{W}^{(s)}$. Thus, the number of speech components $R^{(s)}$ in semi-supervised NMF was equivalent to the number of phonemes (39). The advantage of such phoneme-dependent speech bases over unsupervisedly learnt ones has been shown in [10].

In addition, the number of noise components $R^{(n)}$ as well as the sparsity constant λ and the number of NMF iterations K were optimized in a limited three-dimensional grid search on a subset of the development set which consisted of 10 randomly selected utterances of each speaker at 6 SNRs (parameter ranges: $R^{(n)} \in \{4, 8, 12, 16\}$, $\lambda \in \{0, 0.01, 0.1, 1\}$, $K \in \{1, 2, 4, 8, 16, 32\}$). The separation performance was measured in terms of signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and signal-to-artifact ratio (SAR) [11]. SDR takes into account the trade-off between noise suppression and introduction of separation artifacts, i. e., loss of relevant speech information. The overall best SDR (8.8 dB on average from -6 to 9 dB SNR) was obtained for 4 noise components, 4 NMF iterations and $\lambda = 0.1$, which is a gain of more than 4 dB over the noisy data (average SDR = 4.5 dB). As can be seen from Figure 2, higher numbers of iterations tend to decrease SDR especially for a high number of noise components. More precisely, additional iterations increase noise suppression in terms of SIR at the expense of introducing artifacts (decreasing SAR); this can be explained by overfitting of the noise components to the speech, due to the mismatch of the speaker-independent speech model. Conversely, more noise components are only slightly beneficial for the SDR in the case of 1 or 2 iterations. This is somewhat expected, as the number of noise sources present in a single speech turn is limited, and thus overfitting occurs of too many parameters are estimated in the noise components.

Figure 2: *Signal-to-distortion ratio (SDR, top) and signal-to-interference ratio (SIR, bottom) on the noisy Buckeye development set (average across 6 SNRs from -6 to 9 dB), by number of noise components and number of iterations in semi-supervised sparse NMF ($\lambda = 0.1$).*

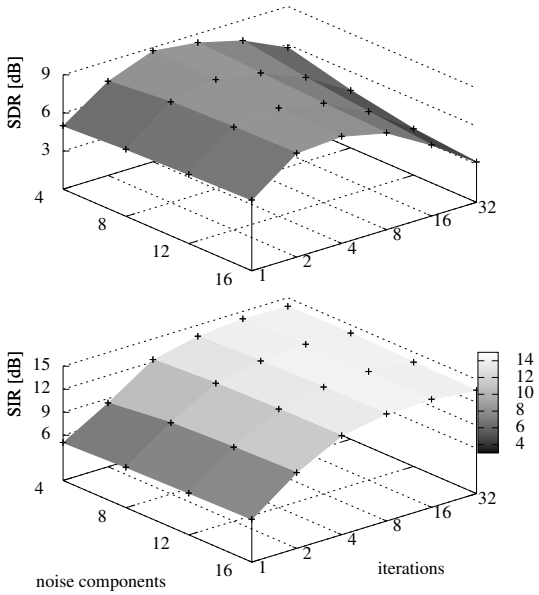
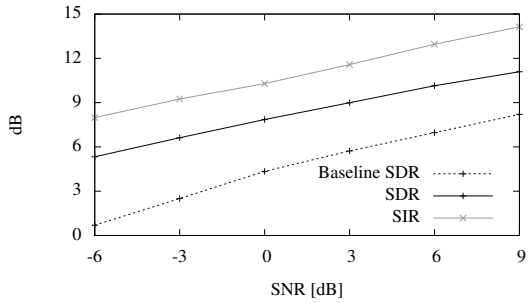


Figure 3: *Separation performance on noisy Buckeye test set: Baseline SDR and SDR / SIR after applying sparse semi-supervised NMF ($K = 4$, $R^{(n)} = 4$, $\lambda = 0.1$).*



4.3. BLSTM-HMM

For tandem feature generation, we trained a BN-BLSTM network consisting of three hidden layers (per input direction) on framewise phoneme targets obtained via HMM based forced alignment of the clean Buckeye training set. All network and training parameters, including the size of the hidden layers, learning rate, etc. were set exactly as in [1]. Only the first 39 principal components of the PCA-transformed BN-BLSTM feature vector were used as final features for tandem ASR. In the HMM system applied for processing the tandem and BN-BLSTM features each phoneme is represented by three emitting states (left-to-right HMMs) with 16 Gaussian mixtures. Tied-state cross-word triphone models with shared state transition probabilities were applied. Both, acoustic models and a back-off bigram language model were trained on the training set of the Buckeye corpus.

Table 1: *Word accuracies [%] on Buckeye test set at SNRs from -6 to 9 dB, on average across these SNRs, and for clean speech. MCT: multi-condition training.*

Front-end	**NMF**	**MCT**	**SNR**							**clean**
			-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	avg	
MFCC	✗	✗	21.21	23.11	25.40	27.85	30.85	34.48	27.15	50.97
MFCC	✗	✓	25.25	27.36	30.09	31.59	34.20	37.00	30.92	43.84
MFCC	✓	✗	23.06	25.32	27.17	29.65	32.56	36.48	29.04	50.54
MFCC	✓	✓	26.51	28.82	30.85	32.85	35.13	37.95	32.02	43.83
BN-BLSTM	✗	✗	22.73	25.08	28.13	30.51	35.16	39.04	30.11	58.21 [1]
BN-BLSTM	✗	✓	34.93	37.58	40.04	41.71	44.60	46.87	40.96	51.12
BN-BLSTM	✓	✗	24.47	26.79	29.75	32.18	36.53	40.74	31.74	57.94
BN-BLSTM	✓	✓	35.74	38.45	40.49	42.45	45.27	47.29	41.62	50.91

5. Results

We evaluate the separation by sparse semi-supervised NMF, with parameters optimized on the development set as described in Section 4.2, on the test set in Figure 3. We observe a constant and significant gain over the noisy SDR baseline; the SDR gain, however, decreases with increasing SNR, ranging from 4.6 dB (-6 dB SNR) down to 2.9 dB (9 dB SNR). Furthermore, NMF boosts the SIR by 7.3 dB at -6 dB SNR and by 5.9 dB at 9 dB SNR. Considering the word accuracies of ASR (Table 1), we observe drastic decreases in noisy conditions; however, by using NMF, we achieve a consistent gain of around 2 % absolute across all SNRs considered. The latter is in strong contrast to our earlier study [3] where a downgrade had to be accepted at high SNRs when using NMF. We attribute this to the optimization of the NMF parameters on SDR which effectively leads to using much less NMF iterations than in [3] (4 instead of 100). In contrast to NMF, using multi-condition training improves the performance of the MFCC front-end particularly in highly noisy conditions, but a severe downgrade of 7 % absolute is observed for clean speech; clean speech, in turn, seems to be largely unaffected by applying NMF. By combining NMF and MCT results on noisy speech can be further improved, but the downgrade for clean speech remains. This downgrade along with the low accuracies in noisy conditions indicate the difficulty of modeling highly variable speech and noise at the same time. The BN-BLSTM front-end delivers consistently higher word accuracies than the MFCC front-end, both with and without NMF; the gain by using the BN-BLSTM front-end instead of MFCC features increases with the SNR and the largest improvement (7 % absolute, up to 58.21 %) is found for clean speech. Finally, the overall best result across noisy speech (35.74 % to 45.27 %, mean = 41.62 %) and clean speech (50.91 %) is observed when combining BN-BLSTM with NMF and MCT. Overall, it seems that the BN-BLSTM can profit much more from training with noisy data than the HMM with MFCC features.

6. Conclusions

We have presented a large scale study on speaker independent recognition of spontaneous speech in various levels of interfering non-stationary noise. Significant gains could be achieved by a combination of NMF and BN-BLSTM, and optimization of NMF on SDR could avoid a downgrade in high SNRs and clean conditions. Still, the word accuracies indicate that this task remains highly demanding, especially due to the interfering speakers occuring in the 'noise'; the latter condition is especially challenging for monaural separation algorithms. Furthermore, we find that the drastic gains in SDR by NMF do not yield a corresponding boost of ASR accuracy, especially not in the case of the noise-robust BN-BLSTM frontend. This is in contrast to our results on 'command and control' speech in the original CHiME Challenge task [3] where we found BLSTM and NMF to be complementary; this observation could be attributed to insufficient power of simple linear, low dimensional NMF models in the case of spontaneous speech. Hence, it will be promising to compare exemplar-based enhancement methods based on the NMF framework. Besides, the mismatch of ASR accuracies and source separation metrics deserve further investigation, in order to enable optimization of source separation for ASR without costly task based evaluations in the future.

7. Acknowledgment

This study has received funding from the Federal Republic of Germany through grant nos. SCHU 2508/2-1 and /4-1.

8. References

[1] M. Wöllmer, B. Schuller, and G. Rigoll, "A novel Bottleneck-BLSTM front-end for feature-level context modeling in conversational speech recognition," in *Proc. of ASRU*, Waikoloa, Big Island, Hawaii, 2011, pp. 36–41.

[2] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 1918–1921.

[3] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments," in *Proc. of CHiME Workshop*, Florence, Italy, 2011, pp. 24–29.

[4] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. of WASPAA*, Mohonk, NY, United States, 2009, pp. 121–124.

[5] F. Weninger, A. Lehmann, and B. Schuller, "openBliSSART: Design and Evaluation of a Research Toolkit for Blind Source Separation in Audio Recognition Tasks," in *Proc. of ICASSP*, Prague, Czech Republic, 2011, pp. 1625–1628.

[6] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[7] A. Stupakov, E. Hanusa, D. Vijaywargi, D. Fox, and J. Bilmes, "The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments," *Computer Speech and Language*, vol. 26, no. 1, pp. 52–66, 2011.

[8] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, *Buckeye Corpus of Conversational Speech (2nd release)*. Columbus, OH, USA: Department of Psychology, Ohio State University (Distributor), 2007, [www.buckeyecorpus.osu.edu].

[9] F. Weninger, B. Schuller, M. Wöllmer, and G. Rigoll, "Localization of non-linguistic events in spontaneous speech by non-negative matrix factorization and long short-term memory," in *Proc. of ICASSP*, Prague, Czech Republic, 2011, pp. 5840–5843.

[10] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. of Interspeech*, Pittsburgh, PA, USA, 2006.

[11] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.