TARGET SPEAKER SEPARATION IN A MULTISOURCE ENVIRONMENT USING SPEAKER-DEPENDENT POSTFILTER AND NOISE ESTIMATION

Pejman Mowlaee[†] and Rahim Saeidi[‡]

[†]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria
[‡] Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
pejman.mowlaee@tugraz.at rahim.saeidi@let.ru.nl

ABSTRACT

In this paper, we present a novel system for enhancing a target speech corrupted in a non-stationary real-life noise scenario. The proposed system consists of one spatial beamformer based on GCC-PHAT-estimated time-delay of arrival followed by three postfilters applied in a sequential way, namely: Wiener filter, minimum mean square error estimator (MMSE) of the log-amplitude, and a model-driven postfilter (MDP) that relies on particular speech signal statistics captured by target speaker Gaussian mixture model. The beamformer accounts for the directional interferences while the MMSE speech enhancement suppresses the stationary background noise, and MDP contributes to suppress the non-stationary sources from the binaural mixture. In our evaluation, multiple objective quality metrics are used to report the speech enhancement and separation performance, averaged on the CHiME development set. The proposed system performs better than standard state-of-the-art techniques and shows comparable performance with other systems submitted to the CHiME challenge. More precisely, it is successful in suppressing the non-stationary interfering sources at different SNR levels supported by the relatively high scores for signal-to-interference-ratio.

Index Terms: Multisource noise, speech enhancement, speech quality, non-stationary noise.

1. INTRODUCTION

Target speaker separation describes the problem of estimating an unknown clean speech signal recorded by one or several microphones in a noisy environment with possible presence of competing speaker(s). The problem finds applications in many different areas of speech communications, including mobile telephony, robust automatic speech recognition and hearing aids. The research in this area has been carried on for decades - with reporting some successful high quality speech enhancement systems. As a noise reduction device is expected to work in noisy environment without a prior knowledge of the noise type, recent research effort has been directed toward studying the robustness of these algorithms in nonstationary noise, including low signal-to-noise ratios (SNRs) [1].

As one step toward studying the problem of enhancing a target speech signal in a multisource environment with nonstationary background noise, recently, the PASCAL challenge, termed as computational hearing in multisource environments (CHiME) was organized [2]. The challenge addresses several critical aspects on the original problem of enhancing and recognizing of a target speech from its noisy version observed in a real-life listening environment mainly characterized by rather low SNR ratios whereas the noise sources are unpredictable, abrupt and highly non-stationary.

Motivated by the recent advances for handling non-stationary noise in speech enhancement [3–8], in this paper we propose a combinative approach to deal with multisource background noise (stationary as well as non-stationary noise sources) in a binaural setup. The proposed system utilizes several postfilters for handling the stationary part of interferences and novel GMM-based speaker models to estimate target speech and further to estimate the non-stationary part of the noise. The performance of the proposed algorithm is evaluated on the CHiME challenge corpus using several instrumental metrics. The performance of the proposed combinative signal-dependent approach is compared to two well-known state-of-the-art signal-independent algorithms in [9, 10] as well as the two top-performing systems [11, 12] that participated in CHiME challenge. Throughout our study we report how much improvement is achievable by incorporating speaker-dependent filters inside the speech enhancement algorithm to successfully handle the nonstationary noise.

2. PREVIOUS METHODS

Previous noise reduction techniques are classified as single and multi-channel. In a multichannel scenario, a beamformer algorithm leads to a promising cancellation of directional noise sources. Still, the usefulness of the beamforming techniques for enhancement purpose gets quite limited, especially when used individually under highly non-stationary or diffused noise scenarios [13]. For single-channel speech enhancement methods, a minimum mean square error (MMSE) estimator in the amplitude (MMSE-STSA) [10] and in the log-amplitude (MMSE-LSA) [9] domain are well-known for dealing well with the stationary additive noise scenario while other algorithms were suggested to handle non-stationary noise types [3, 14]. These techniques mainly rely on noise estimates typically provided by a noise estimation scheme (noise power spectral density (PSD) trackers [4, 14]) in a decision-directed manner, and further assume that the noise signal shows less changes in its second order statistics compared to that of the target speech signal. Such an assumption is not valid for real-life scenarios where the noise signal is highly time-varying and unpredictable or when the noise signal has a statistical characteristic close to the speech. Therefore, the achievable performance obtained by the methods in this group, gets limited when used in such adverse noise conditions [15].

To take advantage of both groups, several methods on combining a beamforming stage with a speech enhancement stage as a post-processor have been suggested [5, 16]. The post-processor at-

The work of Pejman mowlaee was partially funded by the European project DIRHA (FP7-ICT-2011-7-288121), by ASD (Acoustic Sensing & Design) and Speech Processing Solutions GmbH Vienna. The work of Rahim Saeidi was funded by the European Community's Seventh Framework Programme (FP7 2007-2013) under grant agreement no. 238803.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

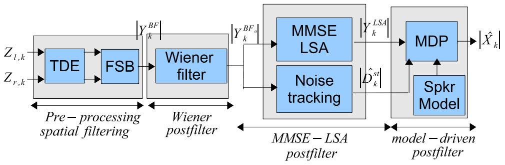

Fig. 1. Block diagram for the proposed system.

Algorithm 1 Steps taken in the proposed system in Fig. 1.

```
Spatial Filtering (Pre-processor)
Align the two channels based on the time-delay estimate $\hat{\tau}$ [19].
Filter-and-sum beamformer.
Wiener postfilter
Apply coherence-based Wiener postfilter [20].
MMSE-LSA postfilter
Using MMSE-LSA in (5) with noise tracking algorithm of [4].
Model-driven postfilter
ML speech estimation using target speaker GMM model in (8).
Recover target signal by applying a mask on the noisy signal (11).
```

tempts to reduce the stationary part of noise that was not canceled by the spatial filter. Also, to reduce the amount of the spectral outliers responsible for the musical noise produced at the output of the single-channel speech enhancement algorithms, some previous studies developed the idea of applying a post-filter based on a pre-trained model on the clean target speech spectra as a constraint for speech enhancement purposes [6–8]. Exploiting an adaptive postfilter was first suggested to enhance the perceptual quality of coded speech by emphasizing formants and pitch harmonics of speech [17]. More recently, the authors in [6] showed that a clean speech codebook is effective in introducing intraframe constrains. A two pass filtering technique composed of a logSTSA filter followed by a post-filter based on vector quantization (VQ) trained on the linear predictor coefficients of the clean speech was presented in [7]. They reported satisfactory improvement in perceptual quality of speech by removing the musical noise of MMSE-LSA for pink and white noise [7]. Finally, we recently suggested the idea of incorporating a VQ codebook of the target speaker as a postfilter and fused it to a noise tracker in a single-channel scenario [18]. The postfilter stage provides the maximum likelihood (ML) speech estimate based on the target speaker model, while the noise tracker provides an estimation of the background noise. The preliminary results on multisource noisy data provided in [1] showed improvement over state-of-the-art signal-independent single-channel speech enhancement techniques which solely rely on noise statistics [9, 10, 14].

3. PROPOSED SYSTEM

The block diagram of the proposed system is shown in Fig. 1. First, the time delay between channels is estimated using the phase transform generalized cross correlation (GCC-PHAT) method [19]. Based on the time-aligned signals, a coherence-based Wiener beamformer [20] is applied. The enhanced single-channel output by the spatial filtering stage is further sent to the MMSE-LSA algorithm [9] using the noise tracker in [4]. Finally, we apply a model-driven postfilter (MDP) which provides the ML speech estimate based on trained speaker models in the form of Gaussian mixture models (GMMs), taking advantage of a good interference cancellation property by model-driven separation systems [21] and perceptual quality enhancement in speech coding [17]. The steps taken are described in Alg. 1. In the following, we present each step in detail.

3.1. Spatial filtering (pre-processor)

Assume $x_l(n)$ and $x_r(n)$ with $n = 0, \cdots, N-1$ denote the nth sample of the left and right time-domain clean speech signals at each frame where N is the signal length in samples. The received signal at each channel experiences the reverberation effect introduced by the acoustic transfer function from the source to each microphone denoted by $h_l(n)$ and $h_r(n)$ with additive background noise denoted by $d_l(n)$ and $d_r(n)$, respectively. Then the binaural noisy observation at the left/right channels is given by

$$z_c(n) = x_c(n) * h_c(n) + d_c(n), \tag{1}$$

where $c = l$ and $c = r$ gives the signal for the left and right channels, respectively. Taking the K-point discrete Fourier transform (DFT) with $k \in 0, \ldots, K/2+1$, we obtain

$$Z_{c,k} = X_{c,k}H_{c,k} + D_{c,k}. \tag{2}$$

Assuming the left channel as the reference signal, the PHAT-weighted generalized cross-correlation (GCC-PHAT) algorithm in [19] is used to provide the time-delay estimate (TDE) of arrival $\hat{\tau}$ between the channels. The output of the spatial filter is given as the sum of the time aligned right and left signals $\tilde{Z}_{r,k} = Z_{r,k}e^{jk\hat{\tau}}$ and $\tilde{Z}_{l,k} = Z_{l,k}$ and we have: $Y_k^{\mathrm{BF}} = \frac{\tilde{Z}_{l,k}+\tilde{Z}_{r,k}}{2}$. Let $\phi_{ij,k}$ with $i = x_l, j = x_r$ be the cross-power spectral density between left and right microphones while for $i = j = \{x_l, x_r\}$, it denotes the auto-power spectral density of left and right microphones, respectively. The Wiener beamformer given by [20]:

$$W_k^{\mathrm{post}} = \frac{2\phi_{\tilde{z}_{l,k}\tilde{z}_{r,k}}}{\phi_{x_{l,k}x_{l,k}} + \phi_{x_{r,k}x_{r,k}}}, \tag{3}$$

is known as a good approximation when there is no correlation between the desired signal and noise as well as if the noise at each channel is uncorrelated. The power spectral densities are approximated using a time recursive averaging, with smoothing parameter of 0.9. The enhanced output is given by: $Y_k^{\mathrm{BFo}} = W_k^{\mathrm{post}}Y_k^{\mathrm{BF}}$.

3.2. Handling stationary noise

Given the beamformer output signal, we apply a single-channel speech enhancement gain function in order to reduce the stationary background noise in the noisy signal. For this we apply the MMSE-LSA noise suppression rule [9] and the noise tracker in [4]. The periodogram of the input signal is smoothed by a first order recursive equation. Based on pilot experiments, we set the key parameters in [4] as: $\eta = 0.7$, $\gamma = 0.998$ and $\alpha_d = 0.95$, where η is the smoothing factor used to smooth the power spectrum of noisy speech, γ is the parameter used to track the minimum of the periodogram of the noisy speech via continuously averaging spectral values of the noisy speech at previous frames, and α_d is the coefficient used in updating the speech-presence probability. The gain function, G_k, is calculated based on estimations of *a priori* and *a posteriori* SNR values denoted by ξ_k and γ_k [15], and is given by:

$$G_k = \frac{\xi_k}{1+\xi_k}\exp\left(\frac{1}{2}\int_{\nu_k}^{\infty}\frac{e^{-t}}{t}dt\right), \tag{4}$$

with $\nu_k = \frac{\xi_k}{\xi_k+1}\gamma_k$. Applying G_k to the beamformer output, $|Y_k^{\mathrm{BFo}}|$ we obtain

$$|Y_k^{\mathrm{LSA}}| = G_k|Y_k^{\mathrm{BFo}}|, \tag{5}$$

which together with the background noise estimate $|\hat{D}_k^{st}|$ is passed to the next step called a model-driven postfilter (MDP).

3.3. Handling non-stationary noise

So far, both spatial and spectral speech estimations function independently from the spectral constraint of the target source, and as a consequence, the gain function G_k leads to musical noise. To suppress the remaining musical noise, we propose to incorporate a postfilter by imposing the target speaker's spectral constraints captured by the Gaussian mixture models learned from the channel-distorted clean speech training data. The proposed model-driven postfilter (MDP) is implemented in two steps: 1) ML speech estimation, and 2) signal reconstruction using a soft mask gain function. In the following, we explain the two steps in details.

3.3.1. ML speech estimation

Based on the estimated background noise, $|\hat{D}_k^{st}|$ found by the noise tracker, we produce a binary mask $\hat{G}_{k,0}$ as below

$$\hat{G}_{k,0} = \left\{ \begin{array}{lll} 1 & , & |\hat{D}_k^{\text{st}}| < |Y_k^{\text{LSA}}| \\ 0 & , & \text{Otherwise} \end{array} \right. . \tag{6}$$

The mask acts like a target speaker activity detector and mostly rejects the speech pauses and noise only regions in the observed noisy signal. This is needed to avoid modeling these regions using the GMM inference. For the regions recognized as noise-only, we apply the spectral gain floor of $20 \log_{10} G_{min} = -25$dB, as suggested by [3].

Let λ be the probability density function for modeling the spectral amplitudes of the target speaker signal. Here, we assume that $\lambda \sim \{\mathcal{N}(w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)\}_{m=1}^M$ is modeled by a GMM where the model parameters are Gaussian weights, means and covariance respectively and M is the model order. The mixture weights are positive and further satisfy the constraint $\sum_{m=1}^M w_m = 1$. Hence, given the model of target speaker and the input enhanced spectrum, $|Y_k^{\text{LSA}}|$, the goal is to find the Gaussian of the model that provides the highest likelihood defined in (7). Assuming diagonal covariance matrices for each Gaussian, from the maximization of the log-likelihood function, the selected mean vector is found as the solution to the following minimization criterion:

$$\boldsymbol{\mu}_{m^*} = \min_m \sum_{k=0}^{K/2+1} \left[\frac{(|Y_k^{\text{LSA}}| - \mu_{k,m})^2}{2\sigma_{k,m}^2} - \ln(\frac{w_m}{\sqrt{2\pi}\sigma_{k,m}}) \right], \tag{8}$$

where $\boldsymbol{\mu}_{m^*}$ is the mean of the Gaussian in the speaker GMM that maximizes the a posteriori probability of the model given the input. We obtain the ML speech estimate as $|\hat{X}_k^{\text{ML}}| = \mu_{k,m^*}$.

3.3.2. Signal reconstruction using soft mask

The ML speech estimate $|\hat{X}_k^{\text{ML}}|$, as an estimate for reverberated clean speech, and $|\hat{D}_k^{\text{st}}|$, as our estimate for the stationary noise spectrum are used to find the non-stationary part of noise, $\hat{d}_n^{\text{nst}}$, as below

$$\hat{d}_n^{\text{nst}} = y_n^{\text{BFo}} - \hat{x}_n^{\text{ML}} - \hat{d}_n^{\text{st}}. \tag{9}$$

Calculation of $\hat{d}_n^{\text{nst}}$ in the time-domain is motivated by the fact that performing the calculation in the spectral-domain leads to negative spectrum amplitudes in some frequency bins, where flooring these amplitudes introduces musical noise. To recover the speech signal of the target speaker, we produce the following soft mask gain function

$$\hat{G}_k = \left\{ \begin{array}{ll} \frac{|\hat{X}_k^{\text{ML}}|}{\sqrt{|\hat{X}_k^{\text{ML}}|^2 + \max(|\hat{D}_k^{\text{st}}|^2, |\hat{D}_k^{\text{nst}}|^2)}} & , |\hat{X}_k^{\text{ML}}| > |\hat{D}_k^{\text{st}}| \\ G_{\text{min}} & , \text{Otherwise} \end{array} \right. , \tag{10}$$

where we define $|\hat{D}_k^{\text{nst}}| = (1 - \tilde{G}_k^2)|Y^{\text{BFo}}|$ and $|\hat{Z}_k^w| = \sqrt{|\hat{X}_k^{\text{ML}}|^2 + |\hat{D}_k^{\text{st}}|^2}$. with $|\hat{D}_k^{\text{nst}}|$ as the estimation for the non-stationary noise with $\tilde{G}_k = \frac{|\hat{Z}_k^w|}{|Y_k^{\text{BFo}}|}$. Finally, using a K-point inverse DFT, the time domain enhanced speech $\hat{x}_n$ is obtained as

$$\hat{x}_n = \text{DFT}^{-1}\{\hat{G}_k |Y^{\text{BFo}}| e^{j\angle Y^{\text{BFo}}}\}. \tag{11}$$

4. EXPERIMENTAL SETUP

4.1. System configuration and speech corpus

A window length of 32 ms and a frame shift of 8 ms were used at the sampling frequency of 16 kHz. GMMs were used to model for the spectral amplitude of the target speaker. The speaker models are trained using the binaural clean reverberated training data provided for each speaker [2]. In this way, the GMMs learn the average room impulse responses and the speaker characteristics. All 500 utterances from the training set are utilized to train a 512 component GMM for each speaker using 10 iterations of the EM algorithm [22].

For performance evaluation, we conducted our experiments on the PASCAL CHiME corpus produced by [2] via convolving the clean speech signals with the real room impulse response to simulate the reverberant environment as well as adding a wide range of noises coming from sources at different locations. The CHiME corpus consists of 34,000 utterances from 18 males and 16 females where the sentences follow a unique grammatical structure. The training set is used to train speaker models, while the development set is used to report the system performance in terms of target speaker separation quality. Averaged on the whole development set, we report segmental SNR (SSNR) to measure speech enhancement performance and BSS EVAL [23] metrics including signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and signal-to-artifact ratio (SAR) to report the separation performance. In all our evaluations, the objective metric is calculated at the left ear using the reverberant target speech as the reference signal.

5. EXPERIMENTAL RESULTS

5.1. Experiment 1: spectrogram analysis

Figure 2 illustrates an example to give indications about how the proposed system deals with background noise composed of stationary and non-stationary parts. The results are shown for two utterances selected from the SiSEC [24] development database corrupted at a signal-to-noise ratio of -3 dB. The reverberated version of the clean signals are used as reference signal to calculate the metrics. The proposed system is capable recovering the most parts of the target speaker spectrogram via effectively rejecting the interference signal. The SSNR improvement is shown in subplot 5 where for further highlight the capability of the proposed system in recovering the target speech signal; in the spectrograms, the regions where SSNR gets improved are marked by black dashed boxes.

5.2. Experiment 2: improvements in speech quality

We compare the performance of the model-driven speech enhancement system with the state-of-the-art speech enhancement methods: MMSE-STSA [10] and MMSE-LSA [9]. For a fair comparison, the beamformer output is used as the input signal to the speech enhancement methods studied here. The SDR and SIR results are shown in Table 1, and averaged on 600 sentences of the development set

$$p_m(|\mathbf{Y}^{\mathrm{LSA}}|) = \frac{1}{(2\pi)^{\frac{K/2+1}{2}}|\mathbf{\Sigma}_m|^{\frac{1}{2}}} \exp\left[-\frac{(|\mathbf{Y}^{\mathrm{LSA}}| - \boldsymbol{\mu}_m)^T \mathbf{\Sigma}_m^{-1} (|\mathbf{Y}^{\mathrm{LSA}}| - \boldsymbol{\mu}_m)}{2}\right] \tag{7}$$

Method	-6	-3	0	3	6	9
Noisy	-6.6±0.1	-4.2±0.0	-1.8±0.1	0.7±0.1	3.3±0.0	5.5±0.1
MMSE-LSA [9]	-5.5±0.2	-2.7±0.3	-0.1±0.3	2.6±0.3	5.4±0.3	7.8±0.3
MMSE-STSA [10]	-5.4±0.2	-2.6±0.3	-0.1±0.3	2.6±0.3	5.4±0.3	7.8±0.3
Proposed	**0.4±0.3**	**1.23±0.3**	**2.57±0.2**	**3.6±0.2**	4.5±0.2	5.1± 0.1

Method	-6	-3	0	3	6	9
Noisy	-6.6±0.1	-4.2±0.1	-1.8±0.1	0.7±0.1	3.3±0.1	5.5±0.1
MMSE-LSA [9]	-5.5±0.2	-2.7±0.3	-0.1±0.3	2.6±0.3	5.4±0.3	7.8±0.3
MMSE-STSA [10]	-5.4±0.2	-2.6±0.3	-0.1±0.3	2.7±0.3	5.4±0.3	7.8±0.3
Proposed	**6.77±0.3**	**7.86±0.3**	**10.2±0.2**	**12.4±0.2**	**14.4±0.2**	**17.0±0.1**

Table 1. Comparing SDR (left) and SIR (right) results for the proposed method versus two state-of-the-art speech enhancement algorithms

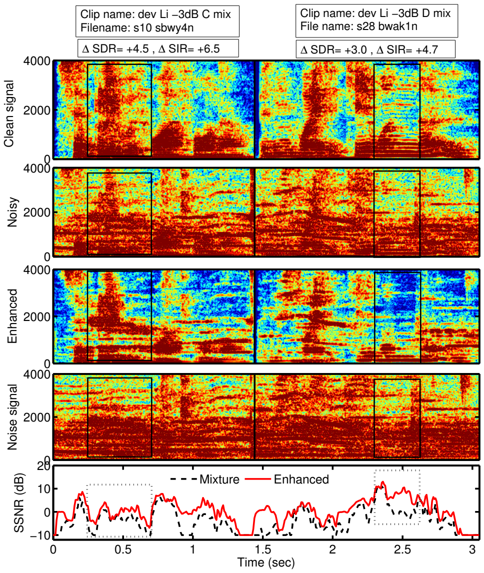

Fig. 2. Showing spectrogram of clean, noisy input, enhanced speech, and noise reference signals for input SNR of -3 (dB). Absolute improvement compared to noisy signal in terms of SDR and SIR are shown, per clip.

group to six input SNRs of -6 to 9 decibels. Considerable improvement versus the state-of-the-art speech enhancement techniques are attained in SDR for input SNR $\leq$ 3dB using the proposed method in this paper. In terms of SIR, we consistently outperform the standard approaches in [9, 10] with a wide margin.

We further compare the results of our method with those who participated in the CHiME challenge. We have received the full enhanced development set files from two participants whose systems are called: 1) data separation based on target signal cancellation and noise masking [12], and 2) non-negative matrix factorization bidirectional long short-term memory (NMF-BLSTM) [11]. Figure 3 shows the BSS EVAL results averaged over 600 sentences in the development set, grouped at different input SNRs. In terms of SDR, it is evident that the proposed system is in line with other top-performing systems submitted to the CHiME challenge and marginally outperforms the NMF-BLSTM in [11] in SNR $= -6$dB. However, the SIR results reveal that the proposed method achieves a consistent improvement at all SNR levels compared to the NMF-BLST approach in [11] but only better than [12] for SNR $\geq$ 3dB. The system in [11] appears to be the best performing system in terms of SAR.

In analysis of the scores of the instrumental quality metrics reported in Figure 3 and Table 1, one should remind the difference between the noise characteristic at low and high SNR scenarios. Low SNRs down to -6 dB are designed as background highly non-stationary energetic events while SNRs up to 9 dB are fairly stationary ambient noise. Therefore, the improvement in performance indicates that the model-driven postfilter stage is capable of handling non-stationary noises.

From the experimental results, it was observed that the proposed system offers a high interference cancellation property, especially at low SNR levels. At high SNR levels, the results indicate that the proposed model-based system will not exceed the metrics evaluated on the unprocessed signal, because of the saturation behavior offered by the model-driven enhancement method.

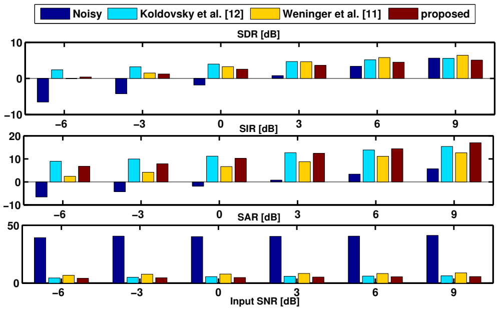

Fig. 3. Comparing the target separation performance of the proposed method versus systems participated in the CHiME challenge.

Acknowledgements

The authors would like to thank Dr. Felix Weninger, and Dr. Zbynek Koldovsky for their help in sharing wave files of their systems submitted to the CHiME challenge as benchmarks in our experiments.

6. CONCLUSION

We presented a multi-stage target speech separation system for processing binaural recordings in environments that may be corrupted by stationary or non-stationary noise. The proposed system combined a spatial beamformer and a GMM-based model-driven postfilter to handle spatial interference and non-stationary noise, respectively. The performance of the proposed system was compared with the state-of-the art speech enhancement methods as well as two benchmark systems submitted to CHiME challenge. The presented system provides consistent improvement over benchmarks in terms of SSNR and SIR. Compared to noisy observation, the proposed system, at -3 dB input SNR on average achieves 4.5 dB improvement in SDR and 9.8 dB in SIR.

7. REFERENCES

[1] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The Pascal CHiME speech separation and recognition challenge," *to appear in Computer Speech and Language*, 2012.

[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. INTERSPEECH*, 2010.

[3] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[4] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, no. 2, pp. 220–231, 2006.

[5] I. Cohen, S. Gannot, and B. Berdugo, "An integrated real-time beamforming and postfiltering system for nonstationary noise environments," *EURASIP Journal on Applied Signal Processing*, , no. 11, pp. 1064–1073, 2003.

[6] T.V. Sreenivas and P. Kirnapure, "Codebook constrained wiener filtering for speech enhancement," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 4, no. 5, pp. 383 –389, Sept. 1996.

[7] J. Wung, S. Miyabe, and B.-H. Juang, "Speech enhancement using minimum mean-square error estimation and a post-filter derived from vector quantization of clean speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2009, pp. 4657–4660.

[8] J. Wung, S. Miyabe, and B.-H. Juang, "Speech enhancement based on a log-spectral amplitude estimator and a postfilter derived from clean speech codebook," in *Proc. European Signal Processing Conf.*, 2010, pp. 999–1003.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443 – 445, Apr 1985.

[10] 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, 1984.

[11] 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 *Proceedings Machine Listening in Multisource Environments, CHiME 2011, satellite workshop of INTERSPEECH 2011*, 2011, pp. 24–29.

[12] Z. Koldovsky, J. Malek, M. Balik, and J. Nouza, "CHiME data separation based on target signal cancellation and noise masking," in *Proceedings Machine Listening in Multisource Environments, CHiME 2011, satellite workshop of INTERSPEECH 2011*, 2011, pp. 47–50.

[13] O.L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926 – 935, Aug. 1972.

[14] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Acoustics, Speech, and Signal Processing, International Conference*, 2010, pp. 4266–4269.

[15] P. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, 2007.

[16] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 6, pp. 561 – 571, nov. 2004.

[17] J. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 59–71, 1995.

[18] P. Mowlaee, R. Saeidi, and R. Martin, "Model-driven speech enhancement for multisource reverberant environment: signal separation evaluation campaign (SiSEC 2011)," in *Proc. the 10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA)*, 2012, pp. 454–461.

[19] M.S. Brandstein and H.F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Acoustics, Speech, and Signal Processing, International Conference*, Apr. 1997, vol. 1, pp. 375 –378.

[20] C. Marro, Y. Mahieux, and K.U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 6, no. 3, pp. 240 –259, May 1998.

[21] M. Cooke, J.R. Hershey, and S.J. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[22] S. Young, *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department, 2006.

[23] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462 –1469, July 2006.

[24] S. Araki, F. Nesta, E. Vincent, Z. Kodovsky, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign SiSEC 2011: Audio source separation," in *Proc. the 10th Int. Conf. on Latent Variable Analysis and Signal Separation*, 2012.