The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/119427

Please be advised that this information was generated on 2019-12-11 and may be subject to change.
TARGET SPEAKER SEPARATION IN A MULTISOURCE ENVIRONMENT USING SPEAKER-DEPENDENT POSTFILTER AND NOISE ESTIMATION

Pejman Mowlaee\(^\dagger\) and Rahim Saeidi\(^\ddagger\)

\(^\dagger\)Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria
\(^\ddagger\)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 priori 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 combinatorial 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 combinatorial 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-
3.1. Spatial filtering (pre-processor)

Assume \( x_l(n) \) and \( x_r(n) \) with \( n = 0, \ldots, N - 1 \) denote the \( n \)th 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

\[
z_c(n) = x_c(n) * h_c(n) + d_c(n),
\]

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 \( k = 0, \ldots, K/2 - 1 \), we obtain

\[
Z_c,k = X_c,k H_{c,k} + D_{c,k}.
\]

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 \( \tau \) between the channels. The output of the spatial filter is given as the sum of the time aligned right and left signals \( Z_{c,k} = Z_{r,k} e^{j\tau} \) and \( \hat{Z}_{l,k} = \hat{Z}_{l,k} \) and we have: \( Y_{k}^{BF} = \hat{Z}_{l,k} \). 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]:

\[
W_{k}^{post} = \frac{2\phi_{1,k}}{\phi_{1,k} + \phi_{2,k}},
\]

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: \( Y_{k}^{BF} = W_{k}^{post} Y_{k}^{BF} \).

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 \( \alpha_t = 0.95 \), where \( \eta \) is the smoothing factor used to smooth the power spectrum of noisy speech, \( \gamma \) 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 \( \alpha_t \) is the coefficient used in updating the speech-presence probability. The gain function, \( G_k \), is calculated based on estimations of a priori and \textit{a posteriori} SNR values denoted by \( \xi_k \) and \( \gamma_k \) [15], and is given by:

\[
G_k = \frac{\xi_k}{1 + \xi_k} \exp \left( \frac{1}{2} \int_{0}^{\infty} \frac{1}{t} dt \right),
\]

with \( \nu_k = \frac{\xi_k}{\xi_k + \gamma_k} \). Applying \( G_k \) to the beamformer output, \( Y_{k}^{BF} \) we obtain

\[
|Y_{k}^{SA}| = G_k |Y_{k}^{BF}|,
\]

which together with the background noise estimate \( |D_{k}^{\text{est}}| \) 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, \( \hat{|D}_k^n| \) found by the noise tracker, we produce a binary mask \( \hat{G}_{k,0} \) as below

\[
\hat{G}_{k,0} = \begin{cases} 
1, & |\hat{D}_k^n| < |Y_k^{LSA}| \\
0, & \text{Otherwise}
\end{cases}
\]  

(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 \( \lambda \) be the probability density function for modeling the spectral amplitudes of the target speaker signal. Here, we assume that \( \lambda \sim \mathcal{N}(u_m, \Sigma_m) \) 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, \( |Y_k^{LSA}| \), 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:

\[
\mu_{m*} = \min_m \sum_k \left[ \frac{(Y_k^{LSA} - \mu_{k,m})^2}{2\sigma_{k,m}^2} - \ln \frac{u_m}{\sqrt{2\pi\sigma_{k,m}}} \right],
\]  

(8)

where \( \mu_{m*} \) 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 \( |\hat{X}_k^{ML}| = \mu_{k,m*} \).

3.3.2. Signal reconstruction using soft mask

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

\[
\hat{d}_k^n = \hat{y}_k^{BF} - \hat{X}_k^{ML} - |\hat{D}_k^n|.
\]  

(9)

Calculation of \( \hat{d}_k^n \) 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 spectral signal of the target speaker, we produce the following soft mask gain function

\[
\hat{G}_k = \begin{cases} 
\frac{|\hat{X}_k^{ML}|}{\hat{G}_{min}}, & |\hat{X}_k^{ML}| > |\hat{D}_k^n| \\
\frac{\sqrt{|X_k^{ML}|^2 + \max(|D_k^n|^2, |D_k^n|^2)}}{G_{min}}, & \text{Otherwise}
\end{cases}
\]  

(10)

where we define \( |\hat{D}_k^n| = (1 - \hat{G}_k^n)|Y_k^{BF}| \) and \( |\hat{D}_k^n| = \sqrt{|X_k^{ML}|^2 + \max(|D_k^n|^2, |D_k^n|^2)} \). with \( |\hat{D}_k^n| \) as the estimation for the non-stationary noise with \( \hat{G}_k^n \). Finally, using a K-point inverse DFT, the time domain enhanced speech \( \hat{y}_n \) is obtained as

\[
\hat{y}_n = \text{DFT}^{-1} \{ \hat{G}_k^n Y_k^{BF} e^{j\omega_k^{BF}} \}.
\]  

(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 SISPEC [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.
\[
p_m(\{Y^{LSA}\}) = \frac{1}{(2\pi)^{K/2+1}|\Sigma_m|^{1/2}} \exp\left[\frac{-(|Y^{LSA}| - \mu_m)^T \Sigma_m^{-1} (|Y^{LSA}| - \mu_m)}{2}\right]
\]  

(7)

<table>
<thead>
<tr>
<th>Method</th>
<th>-6</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td>-6.6</td>
<td>0.1</td>
<td>-4.2</td>
<td>0.0</td>
<td>-3.8</td>
<td>0.1</td>
<td>0.7</td>
<td>0.0</td>
<td>5.3</td>
</tr>
<tr>
<td>MMSE-LSA [9]</td>
<td>-5.4</td>
<td>0.2</td>
<td>-3.3</td>
<td>0.2</td>
<td>-3.0</td>
<td>0.3</td>
<td>2.0</td>
<td>0.5</td>
<td>3.4</td>
</tr>
<tr>
<td>MMSE-LSA [10]</td>
<td>-5.4</td>
<td>0.2</td>
<td>-3.3</td>
<td>0.2</td>
<td>-3.0</td>
<td>0.3</td>
<td>2.0</td>
<td>0.5</td>
<td>3.4</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.4</td>
<td>0.5</td>
<td>1.2</td>
<td>0.0</td>
<td>2.0</td>
<td>0.2</td>
<td>3.6</td>
<td>0.2</td>
<td>5.3</td>
</tr>
</tbody>
</table>

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.

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

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 proposed 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.

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.
7. REFERENCES


