DUAL-CHANNEL NOISE REDUCTION BASED ON A MIXTURE OF CIRCULAR-SYMMETRIC COMPLEX GAUSSIANS ON UNIT HYPERSPHERE

Jalal Taghia†, Rainer Martin†, Jalil Taghia⋆ and Arne Leijon⋆

† Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany
† Email: {Jalal.Taghia,Rainer.Martin}@rub.de
⋆ Sound and Image Processing Lab, KTH Royal Institute of Technology, Stockholm, Sweden
⋆ Email: {Taghia,Leijon}@kth.se

ABSTRACT

In this paper a model-based dual-channel noise reduction approach is presented which is an alternative to conventional noise reduction algorithms essentially due to its independence of the noise power spectral density estimation and of any prior knowledge about the spatial noise field characteristics. We use a mixture of circular-symmetric complex-Gaussian distributions projected on the unit hypersphere for modeling the complex discrete Fourier transform coefficients of noisy speech signals in the frequency domain. According to the derived mixture model, clustering of the noise and the target speech components is performed depending on their direction of arrival. A soft masking strategy is proposed for speech enhancement based on responsibilities assigned to the target speech class in each time-frequency bin. Our experimental results show that the proposed approach is more robust than conventional dual-channel noise reduction systems based on the single- and dual-channel noise power spectral density estimators.

Index Terms— Speech enhancement, dual-channel noise reduction, soft masking, mixture of Gaussians

1. INTRODUCTION

Noise reduction algorithms are an integral part of speech communication systems and have been in great demand for an increasing number of speech applications. Most single- and dual-channel noise reduction algorithms rely considerably on the estimation of the noise power spectral density (PSD) [1]. Fulfilling this requirement is quite challenging in a reverberant environment with non-stationary noise sources. Existing approaches for noise PSD estimation can be categorized into single- and multi-channel classes based on the number of microphones which are used in the process of noise reduction.

Single-channel noise power estimators do not exploit information on the coherence and the direction-of-arrival (DOA) of signals, and they are relatively robust against reverberation and multiple sources. However, not using these properties limits the performance of single-channel noise PSD estimators such that they can achieve considerable noise reduction only at the expense of introducing speech distortion in a speech enhancement framework. In [2] several single-channel noise PSD estimators have been compared in different adverse acoustic environments for a wide range of input signal-to-noise ratios (SNRs). The performance of algorithms was evaluated in terms of both the mean estimation error and the estimation error variance between the reference and the estimated noise PSDs. According to the experiments and performance measures in [2], it was concluded that the most robust noise estimator is the minimum mean-squared error (MMSE) based approach [3].

Recently, in [4] the performance of single- and dual-channel noise power estimators was evaluated in the context of mobile phones. In this study, dual-channel noise power estimators were considered for a dual-microphone mobile phone and two different scenarios of microphone alignments (such as "bottom-top" and "bottom-bottom" alignments) were taken into account. In both alignments the coherence of speech signals was regarded to be roughly close to one and the noise field was assumed to be diffuse. In their results it turned out that the single-channel MMSE based algorithm is superior to most of recently proposed dual-channel noise PSD estimators (e.g. [5], [6], [7]) in terms of the mean estimation error between the reference and the estimated noise PSDs. Generally speaking, except for the assumption on the independence between desired signals and noise signals, most dual-channel noise PSD estimators rely on some additional constraints so that violating them degrades the performance. 1) They require prior knowledge about the noise field coherence. Some approaches assume that noise signals received by different microphones are uncorrelated (e.g. [5]), and some others assume that the noise field is diffuse and the noise coherence function can be estimated by a predefined model (e.g. [6], [7]). 2) Target speech signals are assumed to be highly coherent and their power spectra are equal at two microphones by following equal attenuation paths. 3) Noise power spectral densities at two channels are approximately equal where this assumption is only realistic for diffuse noise fields. Any violation from the aforementioned assumptions can be realized in particular scenarios where for instance directional (point) noise sources are present in the environment or the target speech is not in front. Directional noise sources are highly coherent not only at low frequencies but also over entire frequency range, and the PSDs of the noise signals at different microphones are not necessarily the same depending on their direction of arrival. The noise reduction task is challenging for highly correlated noise signals like directional noise sources since the coherence function cannot be simply estimated with a predefined model as one could employ for the diffuse noise field. In [8] a multi-channel approach was proposed for blind source separation and speech enhancement. Their approach exploits a sparseness model for the observations and an expectation maximization (EM) algorithm is employed for a mixture of complex Watson distribution to perform the dominant source detection and source-to-microphone transfer function ratio (TFR) estimation.

In this paper we propose a model-based dual-channel noise reduction approach which is an alternative to conventional noise reduction algorithms essentially due to its independence of the noise PSD estimation and of any prior knowledge about the noise field characteristics. The proposed approach can be considered as an extension

to the source separation algorithms introduced in [9], [10], and [11]. Our approach is motivated by the recent work in [11] which uses the variational Bayes framework instead of EM method for estimating the model parameters of a mixture of circular-symmetric complex-Gaussian distributions projected on the unit hypersphere.

The noise reduction in our proposed algorithm is performed by bin-wise clustering of the noisy speech DFT coefficients into the noise and the target speech classes depending on their direction of arrival. Finally, a soft masking strategy is introduced for speech enhancement based on responsibilities assigned to the target speech class in each time-frequency bin. The important advantage of the proposed noise reduction approach is to deal with scenarios in which the common aforementioned assumptions for the conventional dual-channel noise reduction systems are violated for example by having directional noise sources in the environment. By providing different adverse noise environments we show in our experimental results that the proposed algorithm results in a more robust speech enhancement performance in comparison to conventional dual-channel noise reduction systems based on the single- and dual-channel noise PSD estimators.

2. PROPOSED APPROACH

The observed noisy signals at two microphones at time instant t can be modeled as

$$\begin{aligned} y_n(t) &= g_n(t) * s(t) + v_n(t) \\ &= x_n(t) + v_n(t), \;\; n = 1, 2 \end{aligned} \tag{1}$$

where $s(t)$ is a speech signal impinging on the two microphones, $g_n(t)$ is the channel impulse response between the source of $s(t)$ and the n^{th} microphone, and $y_n(t)$, $x_n(t)$ and $v_n(t)$ are respectively the noisy speech, the clean speech (i.e. the noise free component), and the noise signal at the n^{th} microphone. It is assumed that the noise signal $v_n(t)$ is statistically independent of the clean speech component $s(t)$. Moreover, the clean speech component $s(t)$ and all the noise components are supposed to be zero-mean random processes. In the DFT domain the signal model (1) is written as

$$Y_n(k,l) = X_n(k,l) + V_n(k,l), \;\; n = 1, 2 \tag{2}$$

where $k = 0, 1, \cdots, K-1$ is the frequency-bin index, K the DFT length, and l is the time-frame index. By considering the vector notation $\mathbf{y}(k,l) \triangleq [Y_1(k,l),\, Y_2(k,l)]^T$, $\mathbf{x}(k,l) \triangleq [X_1(k,l),\, X_2(k,l)]^T$, $\mathbf{v}(k,l) \triangleq [V_1(k,l),\, V_2(k,l)]^T$ of the signal model (1) we have $\mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{v}(k,l)$, where T denotes the transpose operator.

In our procedure, first of all the observation vectors $\mathbf{y}(k,l)$ are normalized such that they have a unit-norm i.e. $\bar{\mathbf{y}}(k,l) = \frac{\mathbf{y}(k,l)}{\|\mathbf{y}(k,l)\|}$. Then, a pre-whitening is performed by multiplying $\bar{\mathbf{y}}(k,l)$ by the whitening matrix $\mathbf{Q}$, as: $\bar{\mathbf{y}}(k,l) \leftarrow \mathbf{Q}\,\bar{\mathbf{y}}(k,l)$, where $\mathbf{Q} = \sqrt{\mathbf{\Lambda}}\,\mathbf{U}^{\mathbf{H}}$ and H is the Hermitian operator. $\mathbf{U}$ and $\mathbf{\Lambda}$ are calculated from the eigenvalue decomposition of the covariance matrix $\mathrm{E}[\bar{\mathbf{y}}\bar{\mathbf{y}}^H] = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$. The normalization procedure is performed one more time after whitening. In the rest of paper, for simplicity in notation, the frequency-bin index is omitted and for instance the observation vector $\bar{\mathbf{y}}(k,l)$ is denoted by $\bar{\mathbf{y}}_l$. In our mixture model, for each observation $\bar{\mathbf{y}}_l$, there is a corresponding latent variable $\mathbf{z}_l$ which forms a 1-of-M binary vector with elements z_{lm}, where M indicates the number of components in the mixture model. Let $\bar{\mathbf{Y}} = \{\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_L\}$ denote the observation matrix in a particular frequency bin, and $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_L\}$ denote the matrix of latent variables, where L indicates the whole number of time frames. The conditional distribution of $\mathbf{Z}$ given the mixing coefficients $\boldsymbol{\gamma} = \{\gamma_m\}$ can be expressed by

$$p(\mathbf{Z}|\boldsymbol{\gamma}) = \prod_{l=1}^{L}\prod_{m=1}^{M} \gamma_m^{z_{lm}}. \tag{3}$$

Hence, the mixture model is defined by expressing the conditional distribution of $\bar{\mathbf{Y}}$ and latent variables $\mathbf{Z}$ given the component parameters $\boldsymbol{\mu} = \{\boldsymbol{\mu}_m\}$ and $\boldsymbol{\lambda} = \{\lambda_m\}$ as

$$p(\bar{\mathbf{Y}},\, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\lambda}) = \prod_{l=1}^{L}\prod_{m=1}^{M} (\gamma_m\, \widetilde{\mathbb{N}}_c(\bar{\mathbf{y}}_l|\boldsymbol{\mu}_m, \lambda_m^{-1}))^{z_{lm}}, \tag{4}$$

where $\widetilde{\mathbb{N}}_c(\bar{\mathbf{y}}_l|\boldsymbol{\mu}_m, \lambda_m^{-1})$ denotes a circular-symmetric complex-Gaussian density function projected on the unit hypersphere,

$$\widetilde{\mathbb{N}}_c(\bar{\mathbf{y}}_l|\boldsymbol{\mu}_m, \lambda_m^{-1}) = \frac{1}{(\pi\lambda_m^{-1})^{D-1}} e^{-\lambda_m \|\bar{\mathbf{y}}_l - (\boldsymbol{\mu}_m^H \bar{\mathbf{y}}_l)\boldsymbol{\mu}_m\|^2}, \tag{5}$$

where D is the dimension of the observation vector $\bar{\mathbf{y}}_l$ (in our case due to the dual-channel system $D = 2$). $\boldsymbol{\mu}_m$ is the centroid with unit norm, $\boldsymbol{\mu}_m^H\boldsymbol{\mu}_m = 1$, and λ_m is the precision which is scalar and the same for all m. $(\boldsymbol{\mu}_m^H\bar{\mathbf{y}}_l)\boldsymbol{\mu}_m$ is the orthogonal projection of $\bar{\mathbf{y}}_1$ onto the subspace spanned by $\boldsymbol{\mu}_m$, hence, the distance $\|\bar{\mathbf{y}}_l - (\boldsymbol{\mu}_m^H\bar{\mathbf{y}}_l)\boldsymbol{\mu}_m\|^2$ determines the dependency of $\bar{\mathbf{y}}_l$ to the m^{th} class. According to Bayes' theorem, we can define the corresponding posterior probabilities once we have observed $\bar{\mathbf{y}}_l$ which are known as responsibilities ξ_{lm}

$$\xi_{lm} = \frac{\gamma_m\, \widetilde{\mathbb{N}}_c(\bar{\mathbf{y}}_l|\boldsymbol{\mu}_m, \lambda_m^{-1})}{\sum_{m=1}^{M} \gamma_m\, \widetilde{\mathbb{N}}_c(\bar{\mathbf{y}}_l|\boldsymbol{\mu}_m, \lambda_m^{-1})} \tag{6}$$

where $\sum_{m=1}^{M} \xi_{lm} = 1,\ \forall\, l \in \{1,\, 2,\, \ldots,\, L\}$.

The variational Bayes approach [11] can be employed to estimate the model parameters $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\gamma}$ of the mixture model (4) by imposing certain prior distributions over the model parameters. Typically, conjugate priors are used such that the prior and posterior will have the same functional form and, hence, optimization procedures can be carried out in an iterative manner. The learning task consists of the optimization of the variational distribution of the latent variables $\mathbf{Z}$ and component parameters $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$. Optimization of the posterior distribution of latent variables leads to a set of responsibilities ξ_{lm} which show how responsible the m^{th} component is for modeling of $\bar{\mathbf{y}}_l$. The reader is referred to [11] for more details on the variational Bayes approach. In this paper, we learn the mixture model (4) and derive the responsibilities ξ_{lm} for the complex-DFT points of the observed noisy speech $\bar{\mathbf{y}}_l$ per frequency-bin. Since the number of sources in our noise reduction scenario is two we set the number of mixture components equal to two (i.e. $M = 2$) for all frequency bins. To transform the noisy speech signals to the DFT domain, the Hann window length as well as the DFT length is set to 2048 samples (128 ms) and the amount of overlap between the frames is set to 75 percent of the frame length (i.e. 96 ms). The parameters for the spectral analysis of noisy speech signals are determined based on our initial experiments and take a trade-off between the performance and the computational complexity of the variational Bayes approach into account. We follow a three-stage approach for the noise reduction. In the first stage the main task is a bin-wise clustering. The bin-wise clustering is performed based on the responsibilities ξ_{lm} ($m = 1, 2$) where the responsibility for a particular component determines the presence probability of a certain source (either target speech or noise) in each time-frequency bin.

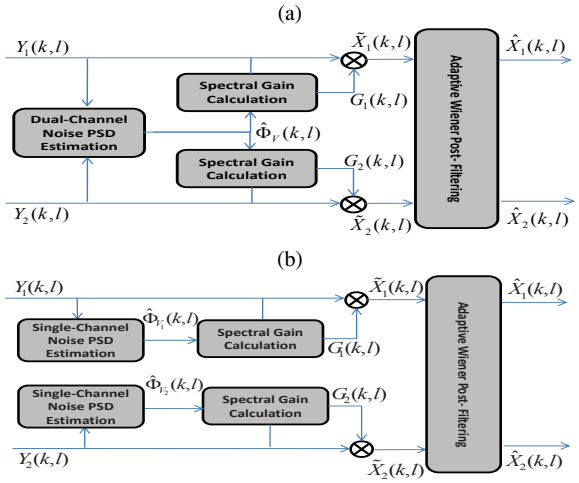

Fig. 1. Block diagrams of noise reduction systems used for: (a) conventional dual-channel noise PSD estimators, and (b) the single-channel noise PSD estimator.

Hence, per frequency-bin we obtain a pattern of responsibilities for each source over time-frame indices. Since the component assigned to a particular source in a frequency-bin can be different to the component assigned to the same source in another frequency-bin, we need to perform a permutation alignment procedure in the second stage to resolve the disorder along frequency-bins (known as permutation ambiguity). Here, we employ the permutation alignment procedure proposed in [10] which is based on the correlation between responsibility patterns in adjacent frequency-bins. After the permutation alignment, we can assume that a particular component is associated with the same source along all frequency bins. The task in the third stage of our approach is to segregate the signals based on a soft masking procedure where at the n^{th} channel ($n \in \{1, 2\}$), the enhanced speech $\hat{X}_n(k,l)$ is derived by

$$\hat{X}_n(k,l) = \xi_{lm^*}(k)\, Y_n(k,l) \tag{7}$$

where m^* ($m^* \in \{1, 2\}$) is the assigned component to the target speech after the permutation alignment which is fixed for all frequency-bins. One could also employ a binary masking process (by assigning zero and one, respectively, to the lower and the higher responsibilities in each time-frequency bin) as used in [11], but in our experiments we observed that the soft masking approach is slightly more effective in the speech enhancement. Hence, our experimental results are presented based on the proposed soft-masking procedure.

3. EVALUATION FRAMEWORK

In this section we describe our framework to evaluate the proposed approach and to compare its performance with the performance of conventional dual-channel noise reduction systems. Fig. 1, (a)-(b), shows two different noise reduction systems where one is based on conventional dual-channel noise PSD estimators and the other one is based on the single-channel noise PSD estimator (MMSE algorithm [3]). Conventional dual-channel noise PSD estimators which we consider in our evaluation framework are: the enhanced coherence-based (ECoh) algorithm [6] and the binaural noise estimation (BNE) algorithm [7]. The required tuning parameters for the implementation of the aforementioned algorithms are chosen as proposed by authors of these works. The noise reduction system presented in Fig. 1 (a) is similar to the dual-channel noise reduction system proposed in [5]. The spectral gain calculation is performed in the frequency domain for each channel based on the decision-directed approach [12] (with smoothing factor equal to 0.9) and the single-channel minimum mean-square error log-spectral amplitude estimator (MMSE-LSA) $G_n(k,l)$ [13]. Since the MMSE-LSA estimator in the structure of the used noise reduction system tends to provide a strong attenuation of speech components, we adopt the procedure introduced in [5] for smoothing the gain values as $G_n(k,l) \leftarrow G_n^{0.7}(k,l)$ to preserve speech components appropriately. As shown in Fig. 1, the spectral gain calculation is followed by an adaptive Wiener post-filter as proposed in [5]. The adaptive Wiener post-filtering is performed by applying the post-filter coefficients $W(k,l)$ to the complex DFT coefficients $\tilde{X}_1(k,l)$ and $\tilde{X}_2(k,l)$ of the noisy speech signals processed by spectral gains $G_1(k,l)$ and $G_2(k,l)$. The Wiener post-filter is estimated as follows [5]

$$W(k,l) = \frac{4\,.\left|\tilde{\Phi}_{\tilde{X}_1\tilde{X}_2}(k,l)\right|^2}{\left(\tilde{\Phi}_{\tilde{X}_1\tilde{X}_1}(k,l) + \tilde{\Phi}_{\tilde{X}_2\tilde{X}_2}(k,l)\right)^2}, \tag{8}$$

where the DFT power spectra $\tilde{\Phi}_{\tilde{X}_1\tilde{X}_1}(k,l)$, $\tilde{\Phi}_{\tilde{X}_2\tilde{X}_2}(k,l)$, and $\tilde{\Phi}_{\tilde{X}_1\tilde{X}_2}(k,l)$ are estimated using the first order IIR filters, e.g.

$$\tilde{\Phi}_{\tilde{X}_1\tilde{X}_2}(k,l) = \beta\,\tilde{\Phi}_{\tilde{X}_1\tilde{X}_2}(k,l-1) + (1-\beta)\,\tilde{X}_1(k,l)\tilde{X}_2^*(k,l) \tag{9}$$

where β is a constant close to one (in the experimental results $\beta = 0.9$), and $*$ is the complex-conjugate operator. In the following, we briefly introduce two different cases describing conditions for the target speech (clean speech signal) and the background noise field in our experiments. The size of the room is approximately $5 \times 4 \times 3$ meters and the reverberation time T_{60} is approximately 0.2s. In the following cases, the corresponding head related transfer functions (HRTFs) are used to produce noisy speech signals.

Case (A): In this case the target speaker is assumed to be about 1 meter away from the dummy head and exactly in front of it at the azimuth angle of 0 degrees. The background noise field is diffuse. To model the diffuse noise field we use the binaural recordings of two different diffuse noise types such as party and children playing. The binaural recordings are taken from ICRA Natural Sounds database [14] Head and Torso Simulator (HATS) recordings which are derived in real life scenarios. According to ICRA Natural Sounds database the selected noise types are described as follows [14]: 1) party noise has been recorded in a large hall and represents the sound of a large party with about 60 people talking all around the listener; and, 2) children playing noise has been made indoor at a big playing field. The children produce a lot of noise including strong impulse noises from various toys. Case (B): In this case (similar to Case (A)) the target speaker is assumed to be about 1 meter far from the dummy head and exactly in front of it at the azimuth angle of 0 degrees. The noise signal reaches to each microphone from a point noise source which is 1 meter far from the dummy head and located at an azimuth angle of 45 degrees anticlockwise from the front. In this case, point noise sources (i.e. single-channel recorded noise signals) like babble (produced by a large crowd and taken from NOISEX-92 [15]) and sinusoidal white Gaussian noise (WGN) are used. Sinusoidal WGN is obtained through modulating WGN by the function, $h(t) = 1 + \sin\left(\frac{2\pi t}{f_s}.\, f_{\mathrm{mod}}\right)$, where t is the sample index, f_s the sampling frequency, and f_{mod} indicates the varying modulation frequency which linearly increases in 30 s from 0 Hz to 0.5

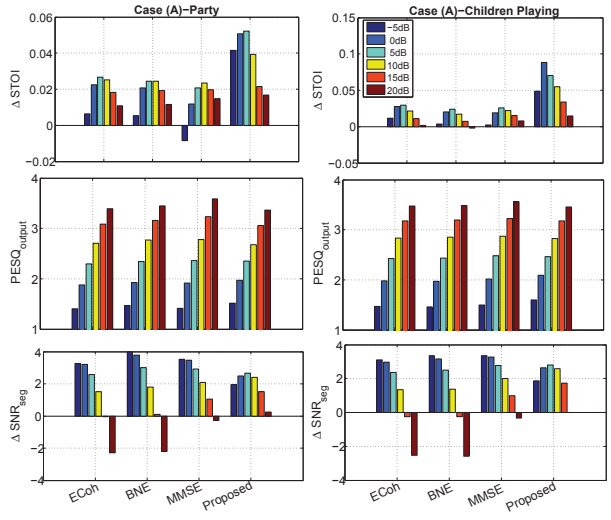

Fig. 2. Performance of the proposed approach and the conventional dual-channel noise reduction systems in terms of $\Delta STOI$, $PESQ_{output}$, and $\Delta SSNR$ for the diffuse noise field case (i.e. Case (A)) and two different noise signals.

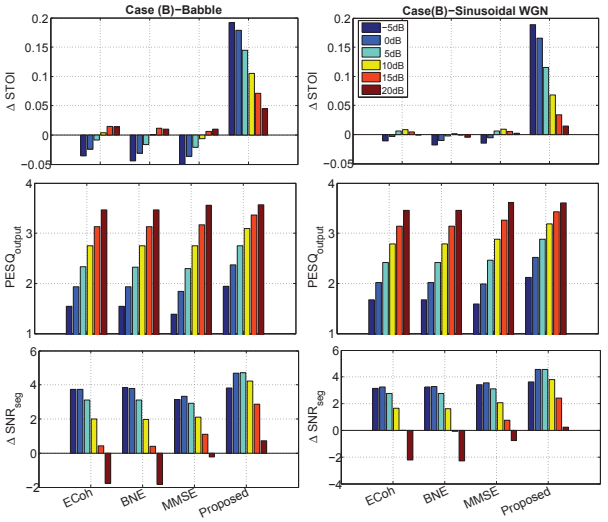

Fig. 3. Performance of the proposed approach and the conventional dual-channel noise reduction systems in terms of $\Delta STOI$, $PESQ_{output}$, and $\Delta SSNR$ for the directional noise case (i.e. Case (B)) and two different noise signals.

Hz. In this paper, due to the space limit, we only present the results for the above cases in which the target speech is assumed to be in front. It should be noted that in our experiments this assumption basically would be in favor of the dual-channel noise PSD estimators selected in our paper (see Section 1) and subsequently the corresponding noise reduction systems.

4. EXPERIMENTAL RESULTS

In this section we present experimental results for the performance of the proposed approach and the conventional dual-channel noise reduction systems (mentioned in Section 3 and shown in Fig. 1) which are based on the single-channel noise PSD estimator (i.e. MMSE algorithm) and the selected dual-channel noise PSD estimators such as ECoh, and BNE algorithms. Except for the proposed approach (where its settings are described in Section 3), all other algorithms were implemented in a DFT-based spectral analysis-synthesis system using overlapping square-root periodic Hann windows. The window length as well as the DFT length are 512 samples (32 ms) and the amount of overlap between the frames is 256 samples (16 ms). We measure the performance with respect to the speech quality enhancement by using the improvement in the segmental SNR [1] (indicated by $\Delta SSNR$) and the Perceptual Evaluation of Speech Quality measure [16] (as implemented in [1]) for the enhanced speech (indicated by $PESQ_{output}$). Furthermore we evaluate the performance of algorithms in terms of the improvement in the speech intelligibility by means of the Short Time Objective Intelligibility measure [17] (indicated by $\Delta STOI$). The sampling frequency of all signals used in this work is 16 kHz. The clean speech signal has a total duration of 40 s taken from the TIMIT database [18] including two male and two female speech signals (of four different speakers) where each one has a duration of 10 s. In the experiments for each type of noises, we change the input overall SNR from -5 dB to 20 dB in 5 dB steps. In Fig. 2 we show the results for the diffuse noise field case (i.e. Case (A)) in terms of the aforementioned performance measures. The experimental results relevant to the case in which noise signals are directional (i.e. Case (B)) are presented in Fig. 3. Furthermore, we considered a case in the experiments in which completely uncorrelated WGN is present at the two microphones. In this case less noise reduction is expected. Here, we only report our observation about the results that for this special case still a segmental SNR improvement of 2.5 dB (or equivalently an overall SNR improvement of 4 dB) can be achieved by our proposed approach.

5. CONCLUSION

In this paper we extended a model-based dual-channel speech enhancement method and evaluated it in a number of different scenarios with diffuse and directional noise. The important difference of the proposed method to conventional noise reduction algorithms is its independence of the noise power spectral density estimation and of any prior knowledge about the noise field characteristics. Similar to a source separation method (e.g., [10], [11]) we exploit a mixture of circular-symmetric complex-Gaussian distributions projected on the unit hypersphere to model the complex DFT coefficients of noisy speech signals in the frequency domain. The noise reduction was performed by bin-wise clustering of the noisy speech DFT coefficients into the noise and the target speech classes depending on their DOA. We proposed a soft mask based on the responsibilities assigned to the target speech class in each time-frequency bin. From the experimental results one can observe that the proposed approach is relatively robust to the type of noise field (i.e. either diffuse or directional noise). The proposed approach proved to be quite effective in speech enhancement for directional noise sources (according to experiments in Case (B)). Furthermore, the proposed approach achieved a relatively good performance for diffuse noise field comparing to the conventional noise reduction systems (in particular for the children playing noise). It is noteworthy that the performance of our proposed method for uncorrelated WGN present at two microphones is approximately close to that of a delay-and-sum beamformer. The potential of the proposed approach in increasing the intelligibility for various conditions in our experiments is indicated by the STOI measure while for the conventional noise reduction systems in some conditions a degraded intelligibility is predicted.

6. REFERENCES

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

[2] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4640–4643, Prague, May 2011.

[3] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4266–4269, March 2010.

[4] C. Gerglotz, M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, "Evaluation of single- and dual-channel noise power spectral density estimation algorithms for mobile phones," *22nd Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Aachen, Germany*, pp. 1–10, September 2011.

[5] M. Dörbecker and S. Ernst, "Combination of two channel spectral substraction and adaptive wiener post-filtering for noise reduction and dereverberation," *In Proceedings European Signal Processing Conference (EUSIPCO), Trieste, Italy*, 1996.

[6] M. Jeub, C. Nelke, H. Krüger, C. Beaugeant, and P. Vary, "Robust dual-channel noise power spectral density estimation," *In Proceedings European Signal Processing Conference (EUSIPCO), Barcelona, Spain*, 2011.

[7] A. H. Kamkar-Parsi and M. Bouchard, "Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment," *IEEE Transactions on Audio, Speech, and Language Processing*, 2009.

[8] D. H. T. Vu and R. Haeb-Umbach, "An EM approach to integrated multichannel speech separation and noise suppression," *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, August 2010.

[9] P. D. O'Grady and B. A. Pearlmutter, "The LOST algorithm: finding lines and separating speech mixtures.," *EURASIP Journal on Advances in Signal Processing, ISSN 1687-6172*, pp. 1–17, 2008.

[10] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516 –527, 2011.

[11] J. Taghia, N. Mohammadiha, and A. Leijon, "A variational Bayes approach to the undetermined blind source separation with automatic determination of the number of sources," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 253–256, Kyoto 2012.

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

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

[14] A. P. Bjerg and J. N. Larsen, "Recording of natural sounds for hearing aid measurements and fitting," *Master's Dissertation. Oersted, Denmark: Danish Technical University (DTU), Acoustic Technology*, May 2006.

[15] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[16] "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation P.862*, Geneva, February 2001.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, September 2011.

[18] "TIMIT, acoustic-phonetic continuous speech corpus," DARPA, NIST Speech Disc 1-1.1, Oct. 1990.