Far Field Speech Enhancement at Low SNR in Presence of Nonstationary Noise Based on Spectral Masking and MVDR Beamforming

Sergei Astapov[1(✉)], Aleksandr Lavrentyev[2], and Evgeniy Shuranov[1,2]

[1] Department of Speech Information Systems, ITMO University, Kronverksky prospekt 49, St. Petersburg 197101, Russia
{astapov,lavrentyev,shuranov}@speechpro.com
[2] Speech Technology Center, Krasutskogo Street 4, St. Petersburg 196084, Russia

Abstract. Low Signal to Noise Ratio (SNR) conditions are highly likely during remote speech acquisition. This paper handles a method of remote speech multi-channel signal processing for speech enhancement in presence of strong nonstationary noise. The presented approach builds upon the Minimum Variance Distortionless response (MVDR) method, additionally filtering the multi-channel signal prior to MVDR beamforming coefficient estimation with a spectral mask. This mask is obtained by applying mixture observation vector clustering based on a spatial correlation model, which is estimated by a Complex Gaussian Mixture Model (CGMM). The posterior probabilities obtained during the CGMM Expectation-Maximization (EM) algorithm are used to estimate the cumulative noise mask, which is applied to the mixture. The masked mixture is then used to calculate the MVDR covariance matrix and beamforming coefficients. The method is tested on four mixtures acquired using a 66 microphone array at various low SNR. The results are compared to conventional MVDR and several other methods and validated using the Signal to Distortion Ratio (SDR) improvement metric. The results show that the presented method gives SDR improvement no less than 1–1.5 dB in the majority of cases, compared to MVDR, and performs best specifically at low SNR of −15 – −20 dB.

Keywords: Speech enhancement · Low SNR · Microphone array
Nonstationary noise · MVDR
Complex Gaussian Mixture Model (CGMM)

1 Introduction

Advances in close proximity speech enhancement and recognition have paved the way for various voice control related services. However, speech enhancement in far field scenarios at low Signal to Noise Ratios (SNR) still poses a problem for tasks situated with remote speech signal acquisition and processing [2,10].

A. Karpov et al. (Eds.): SPECOM 2018, LNAI 11096, pp. 21–31, 2018.
https://doi.org/10.1007/978-3-319-99579-3_3

Due to acoustic wave diffusion and acoustic signal energy attenuation, reverberation [6] in enclosed spaces and physical limitations of the microphone transducer aperture, the speech signal may distort even at relatively low noise levels. Furthermore, assuming that noise sources may appear in a closer proximity to the microphone than the speaker, a low SNR scenario is highly likely.

It has become a common practice to use microphone arrays (MA) for remote speech acquisition and apply multi-channel signal processing methods for speech enhancement [3,15]. Single-channel methods are most effective in cases of narrow-band or stationary noise, where noise statistics can be estimated by, e.g., the Wiener filter, or the signal of interest can be unmixed using, e.g., ICA [8]. Though dual-channel adaptive noise cancellation [2] may be applied in presence of nonstationary wide-band noise, it has its spatial limitations. Multi-channel speech processing, however, allows reducing both diffuse and spatially coherent noise by applying various beamforming techniques [15] and adaptive cancelers [3]. Spatially coherent noise incoming from point noise sources can be canceled by steering a null beamformer in their direction. Such an approach is sensitive to steering vector inaccuracy, does not consider multi-path signal propagation (inc. reverberation) and is prone to partial target signal cancellation. Methods like Minimum Variance Distortionless Response (MVDR), which calculate the beamforming weight coefficients by estimating signal-noise mixture covariance matrices, are generally more robust, but can suffer from estimation errors [3]. The robustness of adaptive beamformers is increased by applying single [5] or multi-channel [4] masks generated through dereverberation [6], array frequency and phase response estimation [7], source separation [9] and other algorithms, which often employ deep learning for purposes of speech recognition [12].

In this paper we attempt to increase the robustness of MVDR for speech enhancement under heavy wide-band nonstationary noise by applying spectral masking to the multi-channel signal mixture prior to calculating the MVDR beamforming weight coefficients. The spectral mask is obtained by applying observation vector clustering based on a spatial correlation model, which is estimated by a Complex Gaussian Mixture Model (CGMM). As a basis for the CGMM Expectation-Maximization (EM) algorithm we adopt a method proposed by Araki et al. [1]. The method is originally used for MVDR steering vector estimation, however, we apply the posteriors obtained after EM directly to the multi-channel signal. The approach is tested on several signal mixtures acquired *in situ* using a 66 microphone MA. The results are validated using the Signal to Distortion Ratio (SDR) metric and compared to the results of several other methods based on the classical Delay-Sum Beamformer (DSB).

2 Preliminary Information

This section regards the problem formulation and provides essential information about the methods applied in our approach to speech enhancement, namely MVDR beamforming, observation vector clustering via the CGMM EM algorithm and the DSB variations used for comparison with our approach.

2.1 Problem Formulation

The entire speech enhancement process is performed in the frequency domain. Let $s(t, f)$ be the Short-Time Fourier Transform (STFT) coefficient of a clean speech signal at time instance t and frequency bin f, and $\mathbf{h}_s(f) = [h_1, \ldots, h_M]^T$ its steering vector, where M is the number of MA channels. The observation vector $\mathbf{y}(t, f) = [y_1(t.f), \ldots, y_M(t, f)]^T$ then has the form

$$\mathbf{y}(t, f) = s(t, f)\mathbf{h}_s(f) + \sum_{k=1}^{N} n_k(t, f)\mathbf{h}_k(f) + \mathbf{n}_d(t, f), \tag{1}$$

where $n_k(t, f)$ is spatially coherent noise produced by a point source k, $\mathbf{h}_k(f)$ is its steering vector and $\mathbf{n}_d(t, f)$ is the diffuse noise component. (Note that all other acoustic sources, including other speakers not-of-interest, are considered spatially coherent noise.) In this paper we assume that the direction to the speaker and, therefore, the acoustic wave propagation vector in the far field, are known (i.e., measured or estimated with zero error). On the other hand, the power spectral densities (PSD) of speech signal and noise components are unknown. We also assume that $M \geq N$. The problem is then to estimate the speech signal $\hat{s}(t, f)$ from the observation vector $\mathbf{y}(t, f)$.

2.2 MVDR Beamforming

The MVDR beamformer output at time instance t and frequency f is given as

$$\hat{s}(t, f) = \mathbf{w}^H(f)\mathbf{y}(t, f), \tag{2}$$

where $\mathbf{w}(f)$ is a $M \times 1$ vector of the beamforming weight coefficients and $(\cdot)^H$ denotes the conjugate transpose of a vector [3]. The optimum weights are selected to minimize the MA output power while maintaining unity gain in the direction of the steering vector of the desired signal $\mathbf{h}_s(f)$:

$$\mathbf{w}(f) = \frac{\mathbf{R}^{-1}(f)\mathbf{h}_s(f)}{\mathbf{h}_s^H(f)\mathbf{R}^{-1}(f)\mathbf{h}_s(f)}, \tag{3}$$

where $\mathbf{R}(f)$ is a $M \times M$ covariance matrix of the signal-noise mixture, which is conventionally calculated as $\mathbf{R}(f) = \sum_t \mathbf{y}(t, f)\mathbf{y}^H(t, f)$, and $\mathbf{h}_s(f)$, in our case of known direction to the speaker, is calculated based on the Time Difference of Arrival (TDOA) τ_i between the first and the i-th microphone as $\mathbf{h}_s(f) = \left[1, e^{-j2\pi f\tau_2}, \ldots, e^{-j2\pi f\tau_i}, \ldots, e^{-j2\pi f\tau_M}\right]^T$.

In our experiments we calculate TDOA directly using known direction to speaker and also validate the measurements by applying two methods for mutual reassurance. The first method consists of an exhaustive search for Angles of Arrival (AOA) in spherical coordinates (θ, φ), which give cumulative spectral energy maxima in the range of $\theta, \varphi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. We calculate them by applying DSB beamforming to the AOA spherical plane in this given range with a discrete

step $\Delta_{\theta,\varphi}$ and calculate the total spectral energy along all frequency bins. For the second method we calculate the MUltiple SIgnal Classification (MUSIC) pseudospectrum in the same AOA range. (AOA estimation and speaker tracking can alternatively be performed using audio-visual methods developed in the Speech Technology Center [11].)

2.3 Observation Vector Clustering with CGMM

Araki et al. [1] attempt to solve the speaker separation problem in a meeting scenario by clustering the signal mixture observation vectors. We revise their EM algorithm according to our task and the definition of the signal mixture (1).

Assuming that the speech signal $s(t,f)$ and spatially coherent noise $n_k(t,f)$ follow a Gaussian distribution of zero mean and variance $|n_k(t,f)|^2 = \phi_{tfk}$:

$$p\left(n_k(t,f);\phi_{tfk}\right) = \mathcal{N}\left(0,\phi_{tfk}\right), \tag{4}$$

the observation vectors follow a complex Gaussian mixture model:

$$p\left(\mathbf{y}(t,f);\lambda\right) = \sum_{k=1}^{N+1} \alpha_{fk} p\left(\mathbf{y}(t,f)|C(t,f)=k;\lambda\right); \tag{5}$$

$$p\left(\mathbf{y}(t,f)|C(t,f)=k;\lambda\right) = \mathcal{N}_c\left(0,\phi_{tfk}\mathbf{B}_{fk}\right), \tag{6}$$

where α_{fk} is a mixture weight $\left(\sum_k^{N+1}\alpha_{fk}=1\right)$, and $\mathbf{B}_{fk} = \hat{\mathbf{h}}_k(f)\hat{\mathbf{h}}_k^H(f)$ is a $M \times M$ spatial correlation matrix of noise source k. The value $C(t,f)=k$, $k=1,\ldots,N$, corresponds to the coherent noise classes and $C(t,f)=N+1$ corresponds to the speech signal class.

The log likelihood function is defined as

$$\mathcal{L}(\lambda) = \sum_t \sum_f \log p\left(\mathbf{y}(t,f);\lambda\right) = \sum_t \sum_f \log \sum_k \alpha_{fk}\mathcal{N}_c\left(0,\phi_{tfk}\mathbf{B}_{fk}\right), \tag{7}$$

where $\lambda = \{\lambda_k\} = \{\{\alpha_{fk},\phi_{tfk},\mathbf{B}_{fk}\}\}$ is the parameter set. The log likelihood function is maximized by using the EM algorithm. The posterior probability is hereafter denoted as $M_k(t,f) = p\left(C(t,f)=k|\mathbf{y}(t,f),\lambda\right)$; the posterior for the speech signal is denoted as $M_{N+1}(t,f)$.

E-step: Calculate the posterior:

$$M_k(t,f) = p\left(C(t,f)=k|\mathbf{y}(t,f),\lambda\right) = \frac{\alpha_{fk}p\left(\mathbf{y}(t,f)|\lambda_k\right)}{\sum_k \alpha_{fk}p\left(\mathbf{y}(t,f)|\lambda_k\right)}. \tag{8}$$

M-step: Calculate the parameters λ as:

$$\phi_{ftk} = \frac{1}{M}\mathbf{y}^H(t,f)\mathbf{B}_{fk}^{-1}\mathbf{y}(t,f), \tag{9}$$

$$\mathbf{B}_{fk} = \frac{\sum_t^T \frac{M_k(t,f)}{\phi_{tfk}}\mathbf{y}(t,f)\mathbf{y}^H(t,f)}{\sum_t^T M_k(t,f)}, \tag{10}$$

$$\alpha_{fk} = \frac{1}{T}\sum_t^T M_k(t,f). \tag{11}$$

2.4 Comparative Methods

We compare the proposed algorithm to four other methods, namely, DSB with adaptive frequency compensation, DSB with adaptive spectral subtraction, DSB with the Stolbov filter [13], and conventional MVDR (referred to as MVDR-MIX).

DSB with adaptive frequency compensation first performs beamforming in the estimated directions and then executes exponential smoothing [2] over the estimated power of the speech signal $\hat{s}(t,f)$ and the residual noise $\hat{n}(t,f)$:

$$\tilde{s}(t,f) = \hat{s}(t,f) - \frac{(1-\alpha)\hat{s}(t-1,f)\hat{n}^H(t-1,f) + \alpha\hat{s}(t,f)\hat{n}^H(t,f)}{(1-\alpha)\left|\hat{s}(t-1,f)\right|^2 + \alpha\left|\hat{s}(t,f)\right|^2}\hat{n}(t,f), \quad (12)$$

where $\tilde{s}(t,f)$ is the enhanced speech signal. In Sect. 4 this method is addressed as DSB-compensate.

The second method first performs DSB and afterwards applies adaptive spectral subtraction [14] to the estimated power of the speech signal $\hat{s}(t,f)$ and the residual noise $\hat{n}(t,f)$:

$$\left|\tilde{s}(t,f)\right|^2 = \left\{\max\left(0, 1 - \frac{\left|\hat{n}(t,f)\right|^2}{\left|\hat{s}(t,f)\right|^2}\right)\right\}\left|\hat{s}(t,f)\right|^2. \quad (13)$$

This method is hereafter denoted as DSB-spect subt.

The Stolbov filter is integrated into the DS beamformer, where each channel is independently processed using an adaptive noise suppressor prior to channel summing [13]. It is shown to provide good noise suppression in presence of wide-band noise. In Sect. 4 it is referred to as DSB-Stolbov.

3 Proposed Approach to Spectral Masking

Our proposed approach is aimed at canceling spatially coherent noise and also filtering diffuse noise by applying a spectral mask to the MA channels prior to calculating the MVDR beamforming coefficients. It is based on the procedure discussed in Sect. 2.3, however, due to the problems situated with inverting the CGMM spatial correlation matrices, this procedure is also adjusted.

3.1 Spectral Mask Application

The spectral mask is obtained by running the CGMM EM algorithm described by (8)–(11). The posteriors $M_k(t,f)$ are used as spectral-temporal coefficients to emphasize the signal prior to MVDR application.

As the sum of posteriors is $\sum_k^{N+1} M_k(t,f) = 1$, we estimate the cumulative noise mask as

$$M_{\mathbf{n}}(t,f) = \sum_{k=1}^{N} M_k(t,f) = 1 - M_{N+1}(t,f), \quad (14)$$

and apply it to the observation vector:

$$\tilde{\mathbf{y}}(t,f) = M_{\mathbf{n}}(t,f)\mathbf{y}(t,f). \tag{15}$$

The masked observation vector is then used to calculate the covariance matrix for (3) as $\tilde{\mathbf{R}}(f) = \sum_t \tilde{\mathbf{y}}(t,f)\tilde{\mathbf{y}}^H(t,f)$, after which the speech signal is estimated using (2) as $\hat{s}(t,f) = \tilde{\mathbf{w}}^H(f)\mathbf{y}(t,f)$. This approach is denoted in Sect. 4 as MVDR-CGMM.

Alternatively we also test a similar approach under the assumption of known speech signal PSD. Our aim is to test, how well the diffuse noise is filtered from the mixture by the MVDR-CGMM approach. Here we assume that the CGMM spatial correlation matrix corresponding to speech $\mathbf{B}_{f,N+1}$ and signal variance $\phi_{t,f,N+1}$ are *a priori* known:

$$\phi_{t,f,N+1}\mathbf{B}_{f,N+1} := |s(t,f)|^2\, \mathbf{h}_s(f)\mathbf{h}_s^H(f). \tag{16}$$

In this case the matrix (16) is fixed, i.e., Eqs. (9) and (10) are not applied on the M-step for $k = N+1$. For $k = 1, \ldots, N$ the EM algorithm is executed in a normal fashion. This approach is denoted as MVDR-CGMM-S in Sect. 4.

3.2 Avoiding Singularity in CGMM EM

Spatial correlation matrix singularity is highly probable under the assumption of unknown noise parameters and random noise source location. In such a case it is not guaranteed that the matrix will be full rank or Hermitian positive definite. We perform several adjustments over the CGMM EM algorithm to minimize the risk of converging to singular spatial correlation matrices.

First, we scale the multivariate Gaussian probability density function to the natural logarithm. The density function of a complex n-variate Gaussian $\mathbf{Z} \sim \mathcal{N}_c(\mu, \Gamma)$ is defined as

$$f(\mathbf{z};\mu,\Gamma) = \frac{e^{-(\mathbf{z}-\mu)^H\Gamma^{-1}(\mathbf{z}-\mu)}}{|\pi\Gamma|}, \quad \mathbf{z} \in \mathbb{C}^n, \tag{17}$$

where $\mathbf{z}$ is a complex vector, μ is the vector of mean values, and Γ is the complex variance. Substituting these arguments with ours and taking the natural logarithm yields:

$$\begin{aligned}\ln f\left(\mathbf{y};0,\phi\mathbf{B}\right) &= \ln\left(\frac{e^{-\mathbf{y}^H\phi^{-1}\mathbf{B}^{-1}\mathbf{y}}}{|\pi\phi\mathbf{B}|}\right) \\ &= -\frac{1}{\phi}\mathbf{y}^H\mathbf{B}^{-1}\mathbf{y} - M\left(\ln\pi + \ln\phi\right) - \ln|\mathbf{B}|\,.\end{aligned} \tag{18}$$

Second, we perform spatial matrix regularization during the M-step if its inverse condition number $\kappa^{-1}(\mathbf{B}) = \max\left(\sum_{j=1}^{M}|b_{ij}|\right)^{-1}$ is below some set value. If this is the case, a small increment is iteratively added to the main diagonal: $\mathbf{B} \leftarrow \mathbf{B} + \epsilon I$, until the condition number check is satisfied.

4 Experimental Results

The test signals for our experiments were acquired in a meeting room with a large table in the middle. Room parameters are as follows: dimensions $\mathrm{L} \times \mathrm{W} \times \mathrm{H} = 6 \times 6 \times 3.5\,\mathrm{m}$, reverberation time $\mathrm{T}_{60} = 0.6\,\mathrm{s}$. For signal acquisition we apply a rectangular MA, consisting of 6 rows of microphones, 11 in each row. The horizontal distance between successive microphones is equal to 35 mm, and the vertical distance – 50 mm. The microphone array is placed on the table at approximately 1.5 m from the wall facing the middle of the room. The speaker is standing facing the array 4 m away from it at AOA of $(\theta, \varphi)_s = (7.154°, 7.395°)$; the loudspeaker reproducing different types of noise is placed facing the array 4 m away from it at AOA of $(\theta, \varphi)_n = (-16.072°, -0.163°)$.

Table 1. Signal mixtures under test and their parameters.

Mixture name	Speaker	Noise	F_s, No. bits
speech+music1	male1	Solarstone - Solarcaster	16 kS/s, 16 bits
speech+music2	male1	Rammstein - Du Hast	16 kS/s, 16 bits
speech+babble	male2	Noisy crowded cafeteria, speech	16 kS/s, 16 bits
speech+white n	female	White noise in band [150, 6000] Hz	16 kS/s, 16 bits

To guarantee accurate SNR readings on the mixture, the speech and noise segments are acquired separately. For each type of noise we sum these speech and noise signals in the frequency domain, while also tuning their gains accordingly to produce the mixtures at specific SNR. To obtain the baseline SDR we then proceed with the following:

1. Apply DSB to the separate speech signal in the direction of the speaker. Obtain the clean speech signal (S).
2. Apply DSB to the mixture in the direction of the speaker. Obtain the enhanced speech signal on the mixture (MIX).
3. Calculate the baseline SDR as $\mathrm{SDR}(\mathrm{MIX}, \mathrm{S})$.

Afterwards all the presented speech enhancement methods are validated using the SDR improvement metric. Each method is applied to the signal mixture and a speech signal estimate $\mathrm{S}_{\mathrm{est}}$ is obtained. SDR improvement is then calculated as $\mathrm{SDR}_{\mathrm{imp}} = \mathrm{SDR}(\mathrm{S}_{\mathrm{est}}, \mathrm{S}) - \mathrm{SDR}(\mathrm{MIX}, \mathrm{S})$.

The signals under test and their components are presented in Table 1. These four combinations are mixed at different low SNR and put through the speech enhancement algorithms discussed in Sects. 2 and 3. STFT parameters for all tests remain the same and are as follows: window length 512 samples, Hann windowing function, overlap 256 samples. The results of SDR improvement are presented in Table 2. All three DSB variations fail to produce noteworthy SDR improvements over conventional DSB; DSB-compensate performs surprisingly

Table 2. Results of SDR improvement in dB for all mixtures under test.

Speech + ...mix at specific SNR	DSB-compensate	DSB-spect subt	DSB-Stolbov	MVDR-MIX	MVDR-CGMM	MVDR-CGMM-S
music1, –5 dB	–0.332	**0.280**	–1.738	–5.985	–1.150	0.113
music1, –10 dB	–0.387	0.149	–1.459	–1.511	**0.642**	0.140
music1, –15 dB	–0.420	–0.091	–1.480	2.513	3.009	**3.654**
music1, –20 dB	–0.473	–0.366	–1.551	5.696	5.772	**6.649**
music2, –5 dB	**2.057**	–0.304	–0.427	–4.505	–1.493	–1.270
music2, –10 dB	2.069	–0.596	–0.240	–0.206	0.230	**2.250**
music2, –15 dB	2.039	–1.050	–0.259	3.798	2.472	**5.591**
music2, –20 dB	1.968	–1.613	–0.197	6.941	5.287	**8.493**
babble, 5 dB	–5.506	–0.254	–3.322	–9.298	–4.635	**–0.149**
babble, 0 dB	–3.126	**–0.046**	–0.266	–5.697	–1.898	–0.085
babble, –5 dB	–2.008	–0.078	**1.264**	–2.581	0.058	–0.065
babble, –10 dB	–1.615	–0.201	**1.496**	–0.041	1.485	–0.037
white n, 5 dB	–8.953	–1.010	–11.330	18.366	30.578	**34.127**
white n, 0 dB	–5.813	0.797	–6.599	24.363	31.725	**33.582**
white n, –5 dB	–4.147	1.278	–2.683	29.244	33.487	**34.792**
white n, –10 dB	–3.514	1.165	–0.130	33.136	33.884	**34.398**
white n, –15 dB	–3.402	0.982	0.820	35.297	**35.477**	34.167

well in the music2 case, and DSB-Stolbov performs best in the babble noise case, which conforms with the results presented in [13]. Conventional MVDR-MIX performs significantly better, especially in lower SNR, which is expected, as given the speech source steering vector, very little room is left for estimation error. However, our MVDR-CGMM outperforms conventional MVDR, giving an improvement of no less than 1–1.5 dB in the majority of cases. This indicates the applicability of the proposed approach in low SNR conditions. MVDR-CGMM-S performs better than MVDR-CGMM at lower SNR mixtures. This may indicate insufficient filtering of diffuse noise and has to be investigated further. Handling stationary white noise does not pose a problem for any of the MVDR variations, however, our method performs better at higher SNR than the conventional MVDR.

An example of speech enhancement by MVDR-CGMM is presented in Fig. 1. Here the speech signal enters the mixture at the 10th second. It can be seen that the music noise dominates almost the entire band of frequencies, however, this noise is efficiently attenuated above 2 kHz. The rhythmic music pattern remains evident only below 2 kHz, and the utterances become distinguishable even at such low SNR. Application of the spectral mask to the speech+music1 mixture is portrayed in Fig. 2. The mask $M_{\mathbf{n}}(t, f) \in (0, 1)$ is presented in a blue-to-yellow color scheme. It clearly indicates the frequency bins belonging to speech, additionally emphasizing the speech signal in noise.

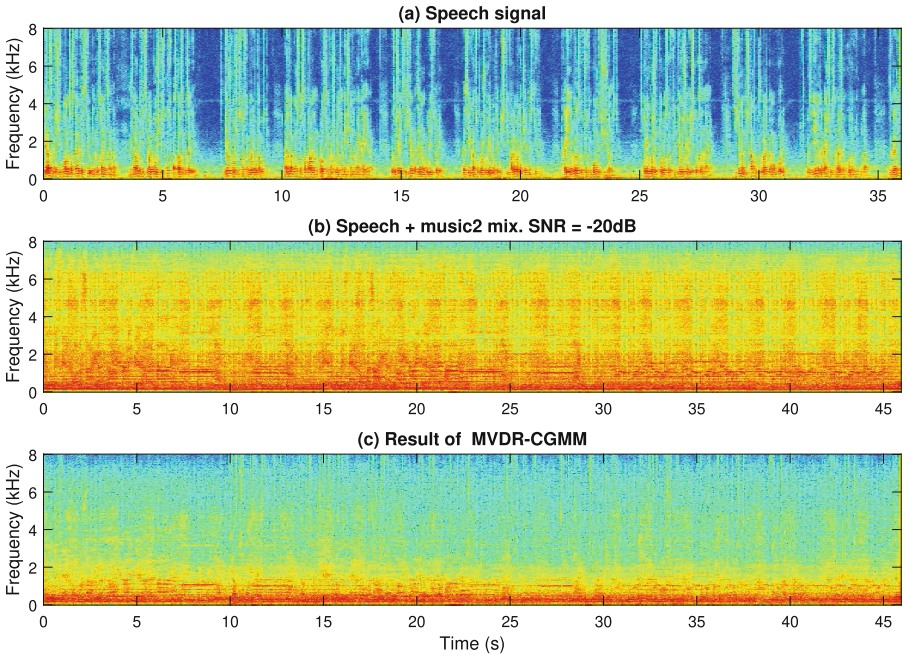

Fig. 1. Example of speech enhancement by MVDR-CGMM at SNR of –20 dB.

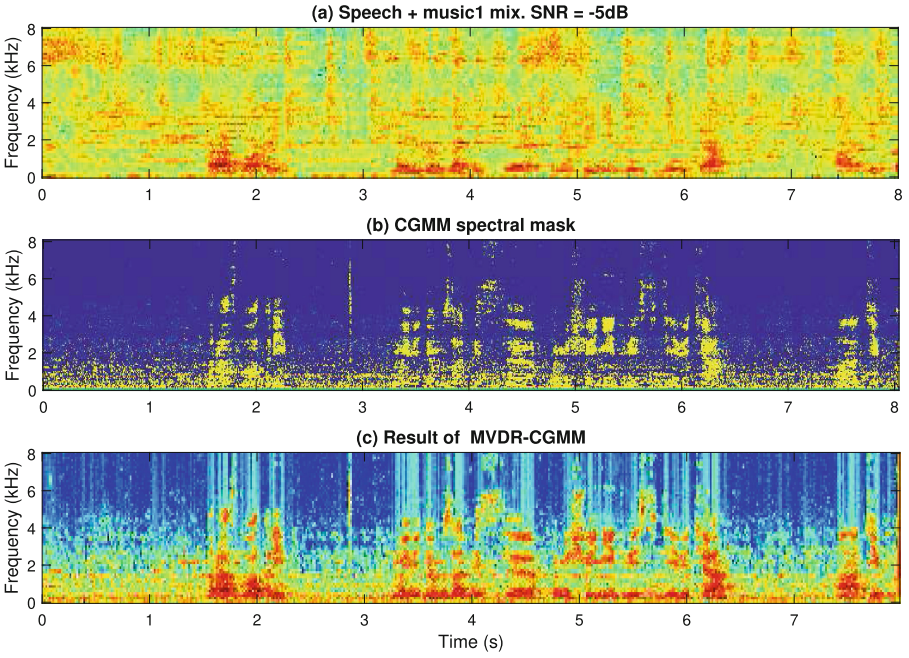

Fig. 2. Example of spectral mask application.

5 Conclusion

In this paper we have discussed the possibility of applying spectral masks for improved speech enhancement in low SNR conditions during remote speech acquisition. The established approach of spectral mask estimation has been proven to be applicable to speech enhancement, improving on the SDR results of the conventional MVDR and several other methods based on DSB. It has provided noticeable SDR improvement specifically at lower SNR conditions in presence of nonstationary noise. Research will be continued in the direction of meeting criteria for automatic speech recognition.

Acknowledgments. This research was financially supported by the Ministry of Education and Science of the Russian Federation, Contract 14.575.21.0132 (IDRFMEFI57517X0132).

References

1. Araki, S., Okada, M., Higuchi, T., Ogawa, A., Nakatani, T.: Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, pp. 385–389, March 2016
2. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Noise Reduction in Speech Processing. Springer Topics in Signal Processing, vol. 2. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-00296-0
3. Brandstein, M., Ward, D.: Microphone Arrays: Signal Processing Techniques and Applications. Digital Signal Processing, Heidelberg (2010). https://doi.org/10.1007/978-3-662-04619-7
4. Cauchi, B., et al.: Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech. EURASIP J. Adv. Signal Process. 61 (2015)
5. Erdogan, H., Hershey, J.R., Watanabe, S., Mandel, M.I., Le Roux, J.: Improved MVDR beamforming using single-channel mask prediction networks. In: Proceedings of Interspeech Conference, INTERSPEECH, pp. 1981–1985 (2016)
6. Habets, E.A.P., Benesty, J.: A two-stage beamforming approach for noise reduction and dereverberation. IEEE Trans. Audio Speech Lang. Process. **21**(5), 945–958 (2013)
7. Higuchi, T., Ito, N., Araki, S., Yoshioka, T., Delcroix, M., Nakatani, T.: Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(4), 780–793 (2017)
8. Hong, L., Rosca, J., Balan, R.: Independent component analysis based single channel speech enhancement. In: Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology, Darmstadt, pp. 522–525, December 2003
9. Jaureguiberry, X., Vincent, E., Richard, G.: Fusion methods for speech enhancement and audio source separation. IEEE Trans. Audio Speech Lang. Process. **24**(7), 1266–1279 (2016)

10. Korenevsky, M.L., Matveev, Y.N., Yakovlev, A.V.: Investigation and development of methods for improving robustness of automatic speech recognition algorithms in complex acoustic environments. In: Anisimov, K.V., et al. (eds.) Proceedings of the Scientific-Practical Conference "Research and Development - 2016", pp. 11–20. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-62870-7_2
11. Oleinik, A.: A lightweight face tracking system for video surveillance. In: Campilho, A., Karray, F. (eds.) ICIAR 2016. LNCS, vol. 9730. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41501-7_46
12. Prudnikov, A., Korenevsky, M., Aleinik, S.: Adaptive beamforming and adaptive training of DNN acoustic models for enhanced multichannel noisy speech recognition. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, pp. 401–408, December 2015
13. Stolbov, M., Aleinik, S.: Speech enhancement with microphone array using frequency-domain alignment technique. In: Proceedings of the Audio Engineering Society 54th International Conference, Audio Forensics, London, pp. 1–6, June 2014
14. Upadhyay, N., Karmakar, A.: Speech enhancement using spectral subtraction-type algorithms: a comparison and simulation study. Procedia Comput. Sci. **54**, 574–584 (2015)
15. Zhao, Y., Jensen, J.R., Christensen, M.G., Doclo, S., Chen, J.: Experimental study of robust beamforming techniques for acoustic applications. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 86–90, October 2017