Rank-1 Constrained Multichannel Wiener Filter for Speech Recognition in Noisy Environments

Ziteng Wang, Emmanuel Vincent, Romain Serizel, Yonghong Yan

To cite this version:

Ziteng Wang, Emmanuel Vincent, Romain Serizel, Yonghong Yan. Rank-1 Constrained Multichannel Wiener Filter for Speech Recognition in Noisy Environments. Computer Speech and Language, Elsevier, 2018, 49, pp.37-51. 10.1016/j.csl.2017.11.003 . hal-01634449

HAL Id: hal-01634449

https://hal.inria.fr/hal-01634449

Submitted on 14 Nov 2017

Rank-1 Constrained Multichannel Wiener Filter for Speech Recognition in Noisy Environments

Ziteng Wang[a,*], Emmanuel Vincent[b], Romain Serizel[c], Yonghong Yan[a]

[a] University of Chinese Academy of Sciences, Beijing, 100190, China
[b] Inria, F-54600, Villers-lès-Nancy, France
[c] Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France

Abstract

Multichannel linear filters, such as the Multichannel Wiener Filter (MWF) and the Generalized Eigenvalue (GEV) beamformer are popular signal processing techniques which can improve speech recognition performance. In this paper, we present an experimental study on these linear filters in a specific speech recognition task, namely the CHiME-4 challenge, which features real recordings in multiple noisy environments. Specifically, the rank-1 MWF is employed for noise reduction and a new constant residual noise power constraint is derived which enhances the recognition performance. To fulfill the underlying rank-1 assumption, the speech covariance matrix is reconstructed based on eigenvectors or generalized eigenvectors. Then the rank-1 constrained MWF is evaluated with alternative multichannel linear filters under the same framework, which involves a Bidirectional Long Short-Term Memory (BLSTM) network for mask estimation. The proposed filter outperforms alternative ones, leading to a 40% relative Word Error Rate (WER) reduction compared with the baseline Weighted Delay and Sum (WDAS) beamformer on the real test set, and a 15% relative WER reduction compared with the GEV-BAN method. The results also suggest that the speech recognition accuracy correlates more with the Mel-frequency cepstral coefficients (MFCC) feature variance than with the noise reduction or the speech distortion level.

Keywords: rank-1 multichannel Wiener filter, speech recognition, residual noise power, deep neural network.

1. Introduction

Robust machine speech recognition in real environments is a common interest for the signal processing and speech recognition communities [1]. It has been

[*] Corresponding author
Email address: wangziteng@hccl.ioa.ac.cn (Ziteng Wang)

a challenging task for decades. One main reason is that the target speech is corrupted by various background noises. Signal processing methods are able to extract the desired source from corrupted measurements and to improve the recognition accuracy. For this purpose, multichannel techniques improve over single-channel techniques by exploiting information not only in the time-frequency domain but also in the spatial domain.

Multichannel linear filters, also known as beamformers, have been amply investigated in the literature [2, 3]. Nevertheless, only a few approaches have found widespread use in the speech recognition community until recently; these include the Weighted Delay and Sum (WDAS) beamformer in BeamformIt [4] and the Minimum Variance Distortionless Response (MVDR) beamformer in BTK[1]. Recent works have explored more extensive beamforming implementations in the scope of speech recognition [5, 6, 7], and the outcomes of these works indeed benefit both signal processing and speech recognition communities. On the one hand, multichannel algorithms designed to suppress noise [8], reverberation [9] or competing speech, can be used as preprocessing steps for speech recognition. Though they are in general intended for improving the speech perceptual quality [10], some improvements are typically also achieved in terms of speech recognition performance. On the other hand, the speech recognition application inspires many new beamforming architectures [11, 12]. The recognition accuracy metric can also highlight an algorithm from a different perspective [13].

Remarkably, Deep Neural Network (DNN) based linear filtering has gained popularity with its success in recent speech recognition challenges [14, 15, 16, 17]. A regression DNN can be used to predict the speech spectra and combined with the classical multichannel Gaussian model to derive a Multichannel Wiener Filter (MWF) [14, 15]. Alternatively, a Bi-directional Long Short-Term Memory (BLSTM) network can be applied as a classification model to predict a spectral mask and combined with the MVDR beamformer or the Generalized Eigenvalue (GEV) beamformer [16, 17]. The mask is used in the calculation of the source covariance matrix, from which the linear filter coefficients are obtained. Deep neural networks have proved to be more capable of estimating the speech second-order statistics or the speech presence probability than traditional methods.

Among the above linear filters, the MVDR beamformer is theoretically designed to be distortionless [18], while the GEV beamformer is targeted to achieve maximum output Signal-to-Noise Ratio (SNR) [19]. MWF [20] is a Minimum Mean Square Error (MMSE) solution which allows for given noise reduction at the expense of some speech distortion. There exist other linear filter variants, such as the Speech Distortion Weighted MWF (SDW-MWF) [21, 22, 23] and the Variable Span (VS) linear filter [24]. The SDW-MWF involves a trade-off parameter which tunes the speech distortion versus the noise reduction. In the case of a single target source, it can be expressed in the form of a spatial-prediction MWF [25] or a rank-1 MWF [26]. Note that these linear filters are all

[1] http://distantspeechrecognition.sourceforge.net

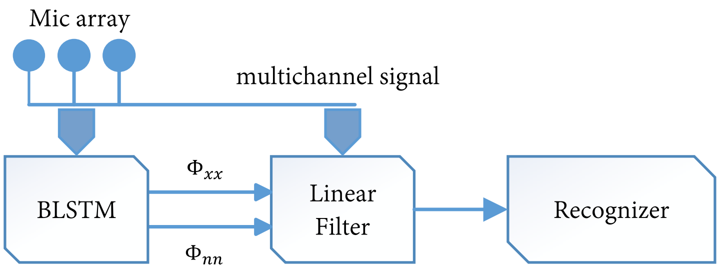

Figure 1: System illustration with BLSTM supported linear filters. $\mathbf{\Phi}_{xx}$ is the speech covariance matrix and $\mathbf{\Phi}_{nn}$ is the noise covariance matrix.

equivalent up to a scaling factor if formulated in a unified framework [24, 27, 28]. While the speech quality performance of these filters has been well studied, the comparison in terms of speech recognition performance is lacking. An interesting question is whether the already known speech quality performance can be related to the speech recognition accuracy.

In this paper, we provide an extensive experimental study of the relative performance of these multichannel linear filters, considering the real world speech recognition task in multiple noisy environments of the CHiME-4 challenge [29]. In particular, we focus on a family of rank-1 MWF variants. We propose a new constraint of constant residual noise power along both time and frequency, which links the rank-1 MWF and the GEV beamformer. This constraint is shown to enhance the speech recognition performance. To fulfill the underlying rank-1 assumption, we introduce a speech covariance matrix reconstruction process. The reconstruction is based on eigenvectors or generalized eigenvectors. In the experiments, all linear filters are supported by the same BLSTM network, which is used for mask estimation. An overview of the system is given in Fig. 1. We also introduce a novel feature variance metric that correlates well with the Word Error Rate (WER) and helps understanding the benefit of the proposed constant residual noise power constraint.

The rest of this paper is organized as follows. The multichannel signal processing problem is formulated in Section 2. In Section 3, the rank-1 MWF solution is first introduced. Three filter variants, including the novel constant residual noise power filter, are then derived separately. To fulfill the rank-1 assumption in practice, the eigenvector based speech covariance matrix reconstruction is discussed in Section 4. The speech recognition experiments, the BLSTM network for mask estimation, the results and the analysis are presented in Section 5. Conclusions are drawn in Section 6.

2. Problem formulation

The multichannel signal processing problem is formulated as follows. A target speech source s propagates in the acoustic space and impinges on an array of M microphones. The observations at time t are given by

$$y_m(t) = g_m * s(t) + n_m(t), \quad m = 1, 2, ..., M \tag{1}$$

where $*$ denotes convolution, g_m is the time-invariant acoustic impulse response from the source to the mth microphone and n_m is the undesired noise at microphone m. Under the narrowband assumption [28], the above model can be written in the frequency domain as

$$\begin{aligned} Y_m(l,k) &= G_m(k)S(l,k) + N_m(l,k) \\ &= X_m(l,k) + N_m(l,k), \quad m = 1, 2, ..., M \end{aligned} \tag{2}$$

where l and k are respectively the frame index and the frequency index. $Y_m(l,k)$, $S(l,k)$ and $N_m(l,k)$ denote the Short-Time Fourier Transform (STFT) coefficients of $y_m(t)$, $s(t)$ and $n_m(t)$, respectively, and $G_m(k)$ is the Fourier transform of g_m. $X_m(l,k) = G_m(k)S(l,k)$ is the narrowband approximation of the reverberated source.

Linear filtering techniques aim to design an optimal filter $\mathbf{h}(l,k) = [H_1(l,k), ..., H_M(l,k)]^T$ which extracts the desired source and suppresses the other components, where subscript ^{T} denotes transposition. This filter is applied to the observation vector $\mathbf{y}(l,k) = [Y_1(l,k),\ ...,\ Y_M(l,k)]^T$, and the filter output is

$$\begin{aligned} O(l,k) &= \mathbf{h}^H(l,k)\mathbf{y}(l,k) \\ &= \mathbf{h}^H(l,k)\mathbf{x}(l,k) + \mathbf{h}^H(l,k)\mathbf{n}(l,k) \end{aligned} \tag{3}$$

where ^{H} denotes Hermitian transpose, $\mathbf{x}(l,k) = [X_1(l,k),\ ...,\ X_M(l,k)]^T$ and $\mathbf{n}(l,k) = [N_1(l,k),\ ...,\ N_M(l,k)]^T$.

The filter coefficients can be derived by setting certain constraints on the filtered output, for instance, to achieve MMSE with respect to an arbitrary channel of the reverberated source, say $X_1(l,k)$. This is expressed as the optimization problem:

$$\min_{\mathbf{h}} E\{|\mathbf{h}^H(l,k)\mathbf{y}(l,k) - X_1(l,k)|^2\} \tag{4}$$

where $E\{\cdot\}$ means expectation. Assuming speech and noise are uncorrelated, we can rewrite (4) as

$$\min_{\mathbf{h}} E\{|\mathbf{h}^H(l,k)\mathbf{x}(l,k) - X_1(l,k)|^2\} + E\{|\mathbf{h}^H(l,k)\mathbf{n}(l,k)|^2\} \tag{5}$$

where the first term is the speech distortion and the second term is the residual noise power. A weight μ can be introduced to control the contribution of the second term:

$$\min_{\mathbf{h}} E\{|\mathbf{h}^H(l,k)\mathbf{x}(l,k) - X_1(l,k)|^2\} + \mu E\{|\mathbf{h}^H(l,k)\mathbf{n}(l,k)|^2\}. \tag{6}$$

The solution of this weighted optimization problem is known as the SDW-MWF [21]

$$\mathbf{h}_{\text{SDW-MWF}}(l,k) = (\mathbf{\Phi}_{xx}(l,k) + \mu\mathbf{\Phi}_{nn}(l,k))^{-1}\mathbf{\Phi}_{xx}(l,k)\mathbf{u}_1 \tag{7}$$

where $\mathbf{\Phi}_{xx}(l,k) = E\{\mathbf{x}(l,k)\mathbf{x}^H(l,k)\}$ is the speech covariance matrix, $\mathbf{\Phi}_{nn}(l,k) = E\{\mathbf{n}(l,k)\mathbf{n}^H(l,k)\}$ is the noise covariance matrix and $\mathbf{u}_1 = [1,\ 0,\ ...,\ 0]^T$ is an M-dimensional vector that projects on the first channel. The hyperparameter μ in the SDW-MWF controls the trade-off between speech distortion and noise reduction. A larger value of μ leads to more noise reduction at the expense of more speech distortion. Specially, the plain MWF is obtained with $\mu = 1$.

3. Rank-1 MWF variants

In the following, we first review the rank-1 MWF solution [26]. Then three filter variants, namely the minimum distortion filter, the plain rank-1 MWF and the new constant residual noise power filter, are derived separately by finding the proper trade-off parameter values. While the first two variants were discussed in [26], the last one is obtained here by analysing the GEV beamformer [19], a filter that maximizes the output SNR. We show that the GEV beamformer also features a constant residual noise power property over both time and frequency. The new rank-1 MWF variant is then derived following this constraint.

3.1. Rank-1 MWF

Under the narrowband approximation (2), the speech covariance matrix can be decomposed as

$$\mathbf{\Phi}_{xx}(l,k) = \phi_{ss}(l,k)\mathbf{g}(k)\mathbf{g}^H(k) \tag{8}$$

where ϕ_{ss} denotes the speech power spectral density and $\mathbf{g}(k) = [G_1(k), ...,\ G_M(k)]^T$ is the vector of acoustic transfer functions. This matrix is of rank-1. Thus $\mathbf{\Phi}_{nn}^{-1}(l,k)\mathbf{\Phi}_{xx}(l,k)$ is also of rank-1. Its unique non-zero eigenvalue is given by

$$\lambda(l,k) = \mathrm{tr}\{\mathbf{\Phi}_{nn}^{-1}(l,k)\mathbf{\Phi}_{xx}(l,k)\} \tag{9}$$

where $\mathrm{tr}\{\cdot\}$ is the trace operation. With Woodbury's identity and the fact that

$$\mathbf{g}^H(k)\mathbf{\Phi}_{nn}^{-1}(l,k)\mathbf{g}(k) = \mathrm{tr}\{\mathbf{\Phi}_{nn}^{-1}(l,k)\mathbf{g}(k)\mathbf{g}^H(k)\} \tag{10}$$

the SDW-MWF solution ends up in the rank-1 MWF

$$\mathbf{h}_{\mathrm{r1MWF}-\mu}(l,k) = \frac{\mathbf{\Phi}_{nn}^{-1}(l,k)\mathbf{\Phi}_{xx}(l,k)}{\mu + \lambda(l,k)}\mathbf{u}_1. \tag{11}$$

Similarly, the trade-off parameter μ controls the speech distortion and noise reduction performance. With different parameter values, the corresponding filter variants exhibit different properties.

3.2. Minimum distortion filter and plain rank-1 MWF

These two filter variants match the cases of $\mu = 0$ and $\mu = 1$, respectively:

$$\mathbf{h}_{\mathrm{r1MWF}-0}(l,k) = \frac{\mathbf{\Phi}_{nn}^{-1}(l,k)\mathbf{\Phi}_{xx}(l,k)}{\lambda(l,k)}\mathbf{u}_1, \tag{12}$$

$$\mathbf{h}_{\mathrm{r1MWF}-1}(l,k) = \frac{\mathbf{\Phi}_{nn}^{-1}(l,k)\mathbf{\Phi}_{xx}(l,k)}{1 + \lambda(l,k)}\mathbf{u}_1. \tag{13}$$

$\mathbf{h}_{\mathrm{r1MWF}-0}$ is indeed distortionless in theory.

3.3. *Constant residual noise power filter*

To derive the new filter variant, we first investigate the maximum SNR filter that is defined as

$$\mathbf{h}(l,k) = \underset{\mathbf{h}}{\operatorname{argmax}} \frac{\mathbf{h}^H(l,k)\boldsymbol{\Phi}_{xx}(l,k)\mathbf{h}(l,k)}{\mathbf{h}^H(l,k)\boldsymbol{\Phi}_{nn}(l,k)\mathbf{h}(l,k)}. \tag{14}$$

This is a generalized Rayleigh quotient and the GEV solution is

$$\mathbf{h}_{\mathrm{GEV}}(l,k) = \mathcal{P}\{\boldsymbol{\Phi}_{nn}^{-1}(l,k)\boldsymbol{\Phi}_{xx}(l,k)\} \tag{15}$$

where $\mathcal{P}\{\cdot\}$ takes the eigenvector corresponding to the largest eigenvalue, which is defined up to an arbitrary scale. An additional Blind Analytical Normalization (BAN) post-filter can be applied to control the speech distortion [19]. The output SNR of the GEV beamformer is equal to the largest eigenvalue of $\boldsymbol{\Phi}_{nn}^{-1}(l,k)\boldsymbol{\Phi}_{xx}(l,k)$, which is exactly λ in the rank-1 case.

Meanwhile, the two Hermitian matrices $\boldsymbol{\Phi}_{xx}(l,k)$ and $\boldsymbol{\Phi}_{nn}(l,k)$ can be jointly diagonalized as

$$\begin{cases} \mathbf{B}^H\boldsymbol{\Phi}_{xx}(l,k)\mathbf{B} = \boldsymbol{\Lambda} \\ \mathbf{B}^H\boldsymbol{\Phi}_{nn}(l,k)\mathbf{B} = \mathbf{I} \end{cases} \tag{16}$$

where $\mathbf{B}$ and $\boldsymbol{\Lambda}$ are respectively the eigenvector[2] and eigenvalue matrices of $\boldsymbol{\Phi}_{nn}^{-1}(l,k)\boldsymbol{\Phi}_{xx}(l,k)$, and $\mathbf{I}$ is the identity matrix [24]. If the diagonal elements of $\boldsymbol{\Lambda}$ are in descending order, then the GEV beamformer (15) can be chosen as the first column vector of $\mathbf{B}$. This is the usual choice made in the literature [19] and the one we also make in the following. We denote it by $\mathbf{h}^*_{\mathrm{GEV}}(l,k) = \mathcal{P}^*\{\boldsymbol{\Phi}_{nn}^{-1}(l,k)\boldsymbol{\Phi}_{xx}(l,k)\}$. By defining the residual noise as $\xi_n = \mathbf{h}^H\mathbf{n}$, we see that the residual noise power of the GEV is given by

$$E\{|\xi_n(l,k)|^2\} = \mathbf{h^*}^H_{\mathrm{GEV}}(l,k)\boldsymbol{\Phi}_{nn}(l,k)\mathbf{h}^*_{\mathrm{GEV}}(l,k) = 1, \tag{17}$$

which indicates constant residual noise power over both frequency and time.

Going back to the rank-1 MWF, it can be proved that the rank-1 MWF solution also satisfies (14) with an arbitrary trade-off parameter. The general expectation of the residual noise power is

$$\begin{aligned} E\{|\xi_n(l,k)|^2\} &= \mathbf{h}^H_{\mathrm{r1MWF}}(l,k)\boldsymbol{\Phi}_{nn}(l,k)\mathbf{h}_{\mathrm{r1MWF}}(l,k) \\ &= \frac{\mathbf{u}_1^T\boldsymbol{\Phi}_{xx}(l,k)\boldsymbol{\Phi}_{nn}^{-1}(l,k)\boldsymbol{\Phi}_{xx}(l,k)\mathbf{u}_1}{(\mu+\lambda(l,k))^2} \\ &= \frac{\phi_{x_1x_1}(l,k)\phi_{ss}(l,k)\mathbf{g}^H(k)\boldsymbol{\Phi}_{nn}^{-1}(l,k)\mathbf{g}(k)}{(\mu+\lambda(l,k))^2} \\ &= \frac{\phi_{x_1x_1}(l,k)\lambda(l,k)}{(\mu+\lambda(l,k))^2} \end{aligned} \tag{18}$$

[2] Note that the eigenvectors are not of unit norm here: they are scaled such that $\mathbf{B}^H\boldsymbol{\Phi}_{nn}(l,k)\mathbf{B} = \mathbf{I}$ holds.

in which the final step makes use of equation (10). Setting the residual noise power to a constant value $E\{|\xi_n(l,k)|^2\} = 1$ as in (17), and taking it into equation (18), we obtain

$$\mu_{\mathrm{G}}(l,k) = \sqrt{\phi_{x_1x_1}(l,k)\lambda(l,k)} - \lambda(l,k) \tag{19}$$

which has become frame and frequency dependent. Thus a rank-1 MWF filter which is similar to the GEV in terms of maximizing the output SNR and leading to constant residual noise power, but different in terms of projection direction, is given by

$$\mathbf{h}_{\mathrm{r1MWF}-\mu_{\mathrm{G}}}(l,k) = \frac{\mathbf{\Phi}_{nn}^{-1}(l,k)\mathbf{\Phi}_{xx}(l,k)}{\mu_{\mathrm{G}}(l,k) + \lambda(l,k)}\mathbf{u}_1. \tag{20}$$

This choice of $\mu = \mu_{\mathrm{G}}$ is new in the context of rank-1 MWF. Although it has been known that linear filters are equivalent up to a scaling factor [24, 27, 28], the factor that specifically relates the rank-1 MWF and GEV is given here by $\frac{1}{\mu_{\mathrm{G}}+\lambda}$ for the first time.

In [30], the residual noise power was chosen as constant over time. Here we restrict it to be constant along frequency too. Note that, under this constraint, the signal can be amplified in some noise-dominated frequency bins and weakened in some speech-dominated frequency bins, which induces speech distortion as does the GEV beamformer. Nevertheless, the derived three rank-1 MWF variants differ only by the spectral shape of the filtered signal. They all project in the spatial direction of $\mathbf{\Phi}_{nn}^{-1}(l,k)\mathbf{\Phi}_{xx}(l,k)\mathbf{u}_1$, but with different spectral gains.

4. Rank-1 constraint on the speech covariance matrix

The above linear filters are specified as functions of the covariance matrices: $\mathbf{\Phi}_{xx}(l,k)$ and $\mathbf{\Phi}_{nn}(l,k)$. In practice, the covariance matrices need to be estimated either by recursive smoothing

$$\tilde{\mathbf{\Phi}}_{xx}(l,k) = \alpha\tilde{\mathbf{\Phi}}_{xx}(l-1,k) + (1-\alpha)\mathcal{M}_x(l,k)\mathbf{y}(l,k)\mathbf{y}^H(l,k) \tag{21}$$

$$\tilde{\mathbf{\Phi}}_{nn}(l,k) = \alpha\tilde{\mathbf{\Phi}}_{nn}(l-1,k) + (1-\alpha)\mathcal{M}_n(l,k)\mathbf{y}(l,k)\mathbf{y}^H(l,k) \tag{22}$$

or by the arithmetic mean

$$\tilde{\mathbf{\Phi}}_{xx}(l,k) = \frac{1}{L}\sum_{l=-L/2}^{L/2-1}\mathcal{M}_x(l,k)\mathbf{y}(l,k)\mathbf{y}^H(l,k) \tag{23}$$

$$\tilde{\mathbf{\Phi}}_{nn}(l,k) = \frac{1}{L}\sum_{l=-L/2}^{L/2-1}\mathcal{M}_n(l,k)\mathbf{y}(l,k)\mathbf{y}^H(l,k) \tag{24}$$

where α is a forgetting factor, and $\mathcal{M}_x$, $\mathcal{M}_n$ represent the speech and noise masks or the speech and noise presence probabilities, respectively. Due to estimation errors or to the fact that the narrowband assumption (8) doesn't hold perfectly, the estimated speech covariance $\mathbf{\Phi}_{xx}(l,k)$ is not rank-1. In [23], using a low-rank approximation of the speech covariance matrix in the SDW-MWF effectively delivered better noise reduction performance. This motivates us to constrain the estimated speech covariance matrix to be rank-1 as follows.

The matrix can be decomposed into a rank-1 part and a remainder part:

$$\begin{aligned}\tilde{\mathbf{\Phi}}_{xx}(l,k) &= \mathbf{\Phi}_{r1}(l,k) + \mathbf{\Phi}_{z}(l,k) \\ &= \sigma_x(l,k)\mathbf{a}(l,k)\mathbf{a}^H(l,k) + \mathbf{\Phi}_{z}(l,k) \end{aligned} \tag{25}$$

where $\sigma_x = \mathrm{tr}\{\tilde{\mathbf{\Phi}}_{xx}(l,k)\}/\mathrm{tr}\{\mathbf{a}(l,k)\mathbf{a}^H(l,k)\}$, and $\mathbf{a}(l,k)$ is defined as the reconstruction vector. The remainder matrix $\mathbf{\Phi}_z(l,k)$ can be either treated as noise or simply ignored, leading to different interpretations of the filter [23]. We choose to ignore the remainder part here. $\mathbf{a}(l,k)$ is chosen from the eigenvector and the generalized eigenvector, that are defined as:

$$\mathbf{a}_{\mathrm{EVD}}(l,k) = \mathcal{P}\{\tilde{\mathbf{\Phi}}_{xx}(l,k)\} \tag{26}$$

$$\mathbf{a}_{\mathrm{GEVD}}(l,k) = \tilde{\mathbf{\Phi}}_{nn}\mathcal{P}\{\tilde{\mathbf{\Phi}}_{nn}^{-1}(l,k)\tilde{\mathbf{\Phi}}_{xx}(l,k)\}. \tag{27}$$

Note that $\mathbf{a}_{\mathrm{GEVD}}$ is interpreted as the desired source relative transfer function in [31]. These two expressions result in new EVD and GEVD based filters, respectively, that fulfill the rank-1 assumption used for deriving the rank-1 MWF. The new filters are given by

$$\tilde{\mathbf{h}}_{\mathrm{r1MWF}-\mu-\mathrm{evd/gevd}}(l,k) = \frac{\tilde{\mathbf{\Phi}}_{nn}^{-1}(l,k)\mathbf{\Phi}_{r1}(l,k)}{\mu + \lambda(l,k)}\mathbf{u}_1. \tag{28}$$

5. Experiments and analysis

5.1. The recognition task

The experiments are conducted on the CHiME-4 challenge data [29]. This dataset features real recordings in four daily noise environments: bus, cafeteria, street junction and pedestrian area. Sentences from the Wall Street Journal (WSJ0) 5k corpus are read from a tablet device. Then the audio signals are captured by a 6-channel microphone array embedded in the tablet frame. For subsequent processing, the signals are downsampled to 16kHz. Besides the real recordings, there are also artificially generated sentences. Clean WSJ0 samples are mixed with the environment noises at similar SNRs as the real data. The whole dataset is divided into disjoint training, development and evaluation sets. In the training set, there are 1600 real and 7138 simulated sentences, about 20 hours in total. In the development set and the test set, there are 1640 and 1320 sentences for each kind of data.

The recognition system is the official challenge baseline built with the Kaldi toolkit[3]. The inputs to the DNN acoustic model are Mel-frequency Cepstral Coefficient (MFCC) features processed by feature space Maximum Likelihood Linear Regression (fMLLR) transformation. The outputs are 1979 Hidden Markov

[3] https://github.com/kaldi-asr/kaldi/tree/master/egs/chime4

Model (HMM) probability states. The acoustic model has 7 layers that are trained under the state level Minimum Bayes Risk (sMBR) criterion. In the decoding phase, a 3-gram Language Model (LM) is used. Recurrent neural network (RNN) LM rescoring is not applied in our experiments: this is the only difference with respect to the official baseline. The results obtained here are not meant to be compared to the best CHiME-4 results, where advanced acoustic and RNN language models are applied.

5.2. Evaluation setup

Table 1: Linear filters involved in the evaluation. They are organized in terms of the projection direction and spectral gain in order to highlight their differences or similarities. The filter $\mathbf{h}$ is given by the product of the projection direction and the spectral gain. Note that $\mathbf{\Phi}_{xx}$, $\mathbf{\Phi}_{nn}$, $\mathbf{a}$, σ_x, λ and μ_{G} depend on time and frequency.

linear filter	reference	projection direction	spectral gain
MVDR	[18]	$\mathbf{\Phi}_{nn}^{-1}\mathbf{a}$, $\mathbf{a} = \mathcal{P}\{\mathbf{\Phi}_{xx}\}$	$\frac{\sqrt{\mathbf{a}^H\mathbf{a}}}{\mathbf{a}^H\mathbf{\Phi}_{nn}^{-1}\mathbf{a}}$
r1MWF-μ-evd	(25)(28)		$\frac{\sigma_x\mathbf{a}^H\mathbf{u}_1}{\mu+\lambda}$, $\mu = 1, \mu_{\mathrm{G}}$
r1MWF-μ	(11)	$\mathbf{\Phi}_{nn}^{-1}\mathbf{\Phi}_{xx}\mathbf{u}_1$	$\frac{1}{\mu+\lambda}$, $\mu = 0, 1, 5, 10, \mu_{\mathrm{G}}$
r1MWF-μ-gevd	(25)(28)	$\mathcal{P}^*\{\mathbf{\Phi}_{nn}^{-1}\mathbf{\Phi}_{xx}\}$, $\mathbf{a} = \mathbf{\Phi}_{nn}\mathcal{P}^*\{\mathbf{\Phi}_{nn}^{-1}\mathbf{\Phi}_{xx}\}$	$\frac{\sigma_x\mathbf{a}^H\mathbf{u}_1}{\mu+\lambda}$, $\mu = 1, \mu_{\mathrm{G}}$
GEV-BAN	(15)		BAN [21]
GEV			1
MWF	(7)	$(\mathbf{\Phi}_{xx} + \mathbf{\Phi}_{nn})^{-1}\mathbf{\Phi}_{xx}\mathbf{u}_1$	1
VS	[24]	$\mathbf{a}\mathbf{a}^H\mathbf{\Phi}_{xx}\mathbf{u}_1$, $\mathbf{a} = \mathcal{P}^*\{\mathbf{\Phi}_{nn}^{-1}\mathbf{\Phi}_{xx}\}$	$\frac{1}{\mu+\lambda}$, $\mu = 1$

The WDAS beamformer [4] is provided as the official baseline for CHiME-4. The linear filters involved in the evaluation are listed in Table 1. They are organized in terms of the projection direction and the spectral gain. GEV-BAN was the method used in the best CHiME-4 submissions [29].

The linear filters are based on the same BLSTM network which simultaneously predicts the speech mask $\mathcal{M}_x$ and the noise mask $\mathcal{M}_n$. In [13], the network was combined with MVDR and GEV. We extend the process here to other linear filters. The STFT is performed in 1024 points with 256 points shift. The magnitude spectrum vector of one frame is used as input. The network consists of one recurrent BLSTM layer with 256 nodes and two feed-forward hidden layers with 512 nodes each. The outputs are 1026 nodes for the speech mask and the noise mask. The target ideal masks are defined as

$$\mathcal{M}_x = \begin{cases} 1 & \mathrm{SNR} > LC_x, \\ 0 & \text{otherwise}, \end{cases} \tag{29}$$

$$\mathcal{M}_n = \begin{cases} 0 & \mathrm{SNR} > LC_n, \\ 1 & \text{otherwise}, \end{cases} \tag{30}$$

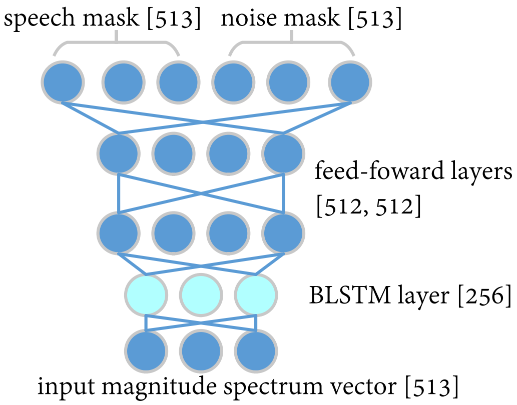

Figure 2: Illustration of the BLSTM network for mask prediction. The numbers in brackets indicate the number of nodes per layer.

where the thresholds for speech and noise detection LC_x and LC_n are set to be 0 dB and -10 dB, respectively. The thresholds are chosen to favor a speech/noise decision with low false acceptance rate. This results in more reliable covariance matrix estimation at the cost of discarding some time-frequency bins [16]. The ReLU activation function is used for all the hidden layers while the sigmoid function is chosen for the output layer. The network is totally single-channel based, i.e., it operates on each microphone signal independently. An illustration of the network architecture is shown in Fig. 2.

In the training stage, the network is trained with all the simulated training utterances from the 6 channels. The simulated data from the development set is used for cross validation and early stopping. The weights of the BLSTM layer are initialized from a uniform distribution ranging from -0.05 to 0.05. The other layers are initialized with samples from a normal distribution with zero mean and a variance of $\sqrt{1/u_{\text{in}}}$ with u_{in} denoting the number of input units. The Adam method [32] is employed to tune the network and the learning rate is adjusted adaptively. Cross-entropy loss is used as the optimization criterion. For better generalization performance, dropout is applied to all the hidden layers. The dropout rate is fixed to 0.5. Batch normalization [33] is applied to speed up the training process and help the network converge to a better local optimum.

In the test phase, the magnitude spectrum vector of the test signal is fed to the trained model and the output masks are in the [0, 1] range. The masks are obtained separately for each channel, and the median value is taken across channels. The median operation is robust to outliers in the case of microphone failure in the real recordings [13]. This value is then used to obtain $\tilde{\mathbf{\Phi}}_{xx}, \tilde{\mathbf{\Phi}}_{nn}$ using (23) and (24). The statistics are averaged on the whole sentence, which leads to time-invariant filters per utterance, that have shown to be more advantageous than time-varying ones for this speech recognition task [29]. For the rank-1 MWF, the reference channel is decided by cross-channel correlations. The channel which has the highest average correlation score with the other channels is selected as the reference.

The experimental setup follows the CHiME-4 challenge instructions: no extra information, such as the environment label, is exploited. The source code is available at `https://github.com/ZitengWang/nn_mask`.

5.3. Recognition results - Acoustic model trained on noisy data

Table 2: WERs (%) achieved by the DNN-sMBR system trained on noisy data. The best result for each dataset is in bold.

Acoustic model	data from channel 5				data from all 6 channels			
Dataset	dev		test		dev		test	
	simu	real	simu	real	simu	real	simu	real
Noisy channel 5	11.43	12.53	14.15	23.52	9.92	11.00	11.44	18.86
WDAS	9.07	8.14	14.20	15.04	8.09	7.30	11.97	12.86
MVDR	6.97	6.86	8.70	10.31	6.21	6.07	7.47	8.89
GEV-BAN	7.27	6.85	9.17	10.48	6.24	6.57	8.25	9.11
GEV	7.54	7.05	10.01	10.53	6.85	6.72	9.21	9.14
MWF	11.24	9.38	12.54	16.16	9.48	7.82	10.17	13.63
VS	5.41	6.53	**6.37**	10.22	**4.85**	5.58	5.30	8.56
r1MWF-0	5.83	6.68	7.03	11.40	5.18	5.83	5.79	9.54
r1MWF-1	5.86	6.70	7.07	11.44	5.22	5.84	5.85	9.74
r1MWF-5	6.01	6.83	7.12	11.71	5.31	6.04	6.00	10.15
r1MWF-10	6.20	6.97	7.41	12.00	5.44	6.15	6.21	10.49
r1MWF-μ_G	6.42	6.43	8.00	10.33	5.76	5.73	6.61	8.89
r1MWF-1-evd	5.82	6.83	7.05	11.17	5.09	5.66	5.99	9.56
r1MWF-1-gevd	**5.37**	6.59	6.40	10.26	4.86	**5.52**	**5.16**	8.47
r1MWF-μ_G-evd	6.05	6.12	7.63	9.25	5.41	5.54	6.14	8.09
r1MWF-μ_G-gevd	6.01	**6.03**	6.84	**8.74**	5.29	5.53	5.83	**7.71**

In the first experiment, two acoustic models are trained with the noisy data: one with utterances from the official channel 5 (∼20h) and the other with utterances from all 6 channels (∼120h). The involved linear filters are only applied to the development data and the test data. The WER results are given in Table 2.

From an overall perspective, the results of the linear filters follow the same trends for both acoustic models. The filers consistently enhance the recognition performance and lower WERs are achieved as expected with more training data. The performance difference between simulated data and real data is small on the development set. The overall higher error rates on the test set are due to the fact that the speakers of the test set speak in a less intelligible way [34]. The following discussions concentrate on the results achieved on the test set with the acoustic model trained on utterances from all 6 channels.

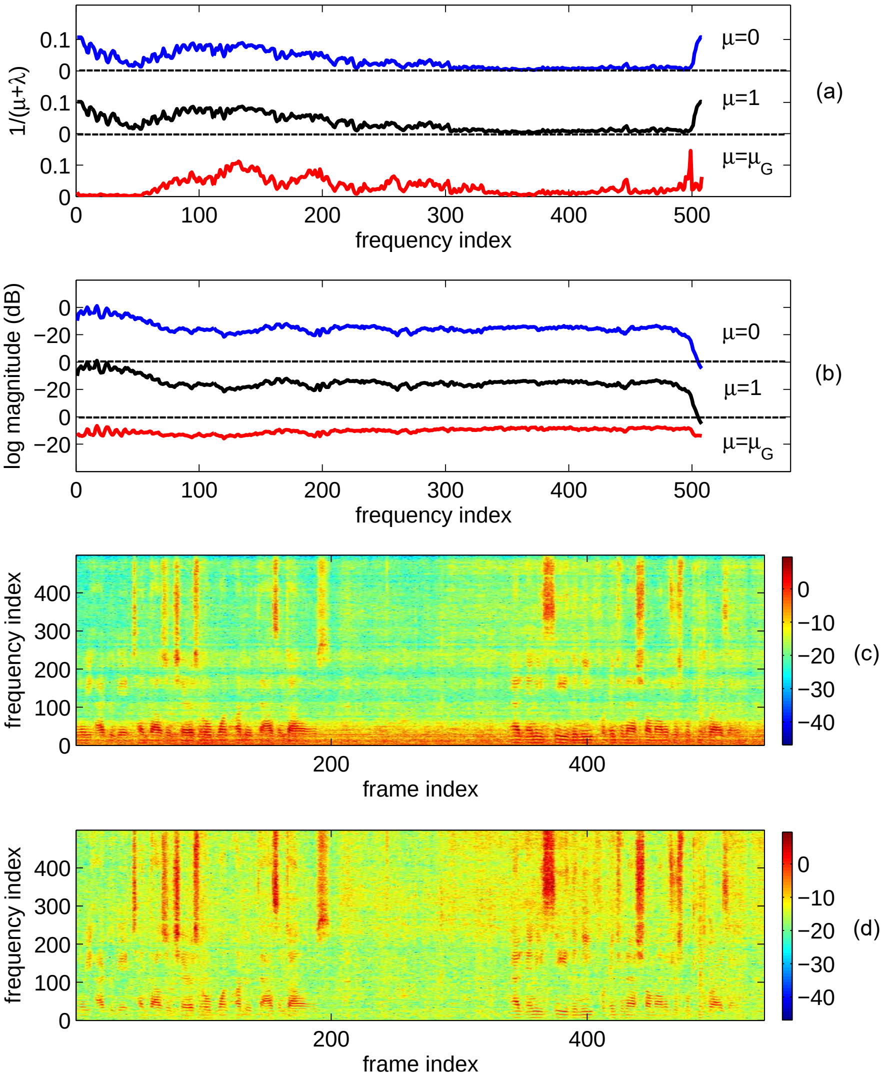

Figure 3: Illustration of the r1MWF-μ filter for an example sentence (M06_440C0201_BUS) from the real test set. (a) Spectral gain along frequency for different values of μ. (b) The corresponding log-magnitude of one frame of the filtered signals. (c) and (d) Log-magnitude spectrograms of the filtered signals with $\mu = 1$ and $\mu = \mu_G$, respectively.

Compared with the noisy baseline, it is obvious that all the multichannel methods improve the speech recognition performance. The WDAS beamformer
285 is a simple but effective technique, which delivers 35% relative WER reduction on the real data. The MWF achieves less reduction here partly due to its sensitivity to mask estimation errors [35]. The MVDR filter is theoretically

speech distortionless and further improvement is achieved from the WDAS filter. For instance, the WER is reduced from 12.86% to 8.89% on the real data. The GEV and GEV-BAN surprisingly lead to comparable results, despite the fact that BAN is believed to be crucial to the speech perceptual quality [19]. There is around 1% absolute difference on the simulated data though. The VS filter gets the lowest WER among the above ones. It is especially effective on the simulated data with an average 25% relative improvement from the MVDR filter. The recognition performance is clearly influenced by the projection direction of the beamformers as shown by the GEV, MWF filters and the VS, r1MWF-1 filters.

Regarding the rank-1 MWF variants without speech covariance matrix reconstruction, the distortionless r1MWF-0 works best on the simulated data while the residual noise power constrained r1MWF-μ_{G} works best on the real data. By changing the trade-off parameter μ from 0 to $\{1,5,10\}$, more noise reduction is achieved in the processed signal at the expense of more speech distortion. This results in worse recognition performance in this specific task: WERs increase as μ increases. Note that for the r1MWF-μ_{G}, this trade-off parameter is frequency dependent. In Fig. 3, the spectral gain along frequency and the filtered signals are shown for different parameter values. The r1MWF-μ_{G} has small gain in the low frequencies and puts more weight in the high frequencies, leading to relatively stable level of log-magnitudes as shown in Fig 3 (b) and Fig 3 (d). The differences in the (time-varying) spectral gain result in different recognition accuracies.

Additional improvement is observed with the speech covariance matrix reconstruction process. On the real test data, the WER is reduced from 8.89% to 8.09% for the r1MWF-μ_{G}-evd and 7.71% for the r1MWF-μ_{G}-gevd. Overall, the r1MWF-μ_{G}-gevd gives the best result. It achieves a 40% relative WER reduction compared with the baseline WDAS beamformer on the real test set and a 15% relative WER reduction compared with the GEV-BAN method.

An interesting experiment is to check the performance of these filters using the *correct* masks instead of the predicted ones. This presumably would help to partially discriminate the error rate caused by covariance estimation errors and limitations of the multichannel linear filters themselves. The correct masks for the simulated data are well defined, however, the ground truth underlying the real data is not readily available. The method in [29] is adopted for the ground truth estimation for real data and then the masks are calculated using (29) and (30). The recognition results are summarized in Table 3, with the percentages in brackets denoting the relative WER changes from the results in the left half of Table 2, that are obtained with the BLSTM predicted masks.

The relative performance between the linear filters is generally consistent with the previous results, though a reduction of WERs on the simulated data is observed and an overall increase of WERs is observed on the real data. For instance, the WERs of GEV-BAN on the test set decrease by 29% relative on simulated data and increase by 14% relative on real data. This indicates that GEV-BAN would benefit from better estimated masks on simulated data. This also indicates that the ground truth estimation process is not perfect and GEV-BAN is prone to covariance estimation errors. In comparison, the VS filter

Table 3: WERs (%) achieved by the DNN-sMBR system trained on noisy data from channel 5. These filters are computed from the *correct* masks. The percentages in brackets denote the relative WER changes from the results obtained with the BLSTM predicted masks. The best result for each dataset is in bold.

Dataset	dev		test	
	simu	real	simu	real
MVDR	6.25 (-10%)	7.01 (+ 2%)	7.18 (-17%)	10.90 (+ 6%)
GEV-BAN	5.87 (-19%)	7.95 (+16%)	6.51 (-29%)	11.94 (+14%)
GEV	6.70 (-11%)	8.35 (+18%)	7.42 (-26%)	12.84 (+22%)
MWF	7.67 (-31%)	8.65 (- 7%)	8.36 (-33%)	15.91 (- 2%)
VS	**4.92** (- 9%)	**6.29** (- 4%)	5.83 (- 8%)	10.43 (+ 2%)
r1MWF-0	5.00 (-14%)	6.44 (- 4%)	**5.74** (-18%)	11.16 (- 2%)
r1MWF-1	5.00 (-15%)	6.49 (- 3%)	5.75 (-19%)	11.23 (- 2%)
r1MWF-μ_G	5.81 (-10%)	6.71 (+ 4%)	6.68 (-17%)	10.94 (+ 6%)
r1MWF-μ_G-evd	5.65 (- 7%)	6.32 (+ 3%)	6.38 (-16%)	10.03 (+ 8%)
r1MWF-μ_G-gevd	5.56 (- 7%)	6.36 (+ 5%)	6.24 (- 9%)	**10.00** (+14%)

is more robust to mask misestimation and achieves the lowest WERs on the development set. Comparing r1MWF-μ_G-evd and r1MWF-μ_G-gevd to r1MWF-μ_G, the rank-1 constraint on the speech covariance matrix still leads to lower error rates. On the real test data, r1MWF-μ_G-gevd achieves a 16% relative WER reduction compared with the GEV-BAN method in this case.

5.4. Recognition results - Acoustic model trained on enhanced data

In the second experiment, the acoustic model is retrained with the filtered training data. The WERs are shown in Table 4. They are comparable to the left half of Table 2 in the sense that the amount of training data is the same.

On the real data, all linear filters generally achieve higher error rates than in the first experiment, except for the GEV filter. On the simulated data, the WERs are generally lower. The proposed r1MWF-μ_G-gevd is still the best on real data. Note that retraining the acoustic model every time is rather time-consuming and not efficient in practice. The results here provide a strong argument for noisy training, that extends the argument made specifically for the GEV-BAN in [13].

5.5. Analysis

The above results suggest that neither speech distortion nor noise reduction is straightforwardly correlated with the speech recognition performance. Indeed, the GEV introduces more speech distortion than the theoretically distortionless MVDR but it performs better in the second experiment. The r1MWF-5/10

Table 4: WERs (%) achieved by the DNN-sMBR system trained on enhanced data. The best result for each dataset is in bold.

Dataset	dev		test	
	simu	real	simu	real
Noisy	11.43	12.53	14.15	23.52
MVDR	6.80	6.97	8.61	11.58
GEV-BAN	6.59	7.14	7.43	10.62
GEV	6.83	7.01	7.70	9.91
VS	**5.59**	6.55	6.30	11.12
r1MWF-0	5.95	6.83	6.87	12.55
r1MWF-1	6.72	7.45	7.70	13.74
r1MWF-μ_G	6.89	7.35	7.82	12.07
r1MWF-1-evd	5.92	6.66	6.96	12.32
r1MWF-1-gevd	5.65	6.48	**6.13**	11.19
r1MWF-μ_G-evd	5.76	6.13	7.26	10.33
r1MWF-μ_G-gevd	5.79	**6.04**	6.48	**9.52**

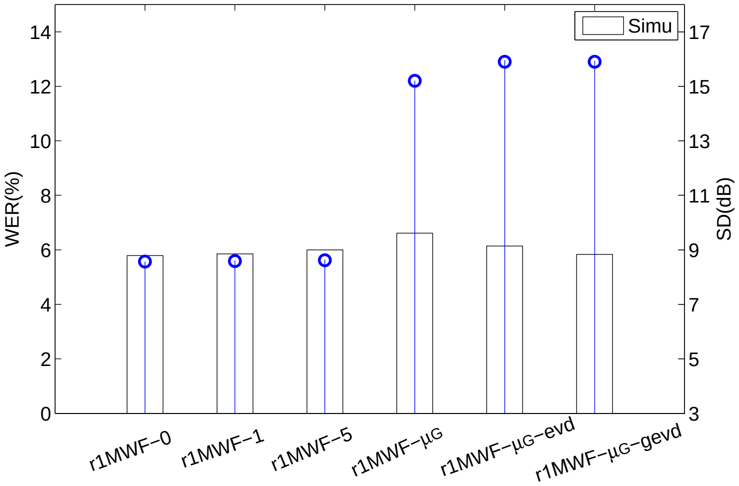

Figure 4: WERs achieved on the acoustic model trained on utterances from all 6 channels and SD scores of the r1MWF variants. WERs are represented by white bars and SD scores are marked by circles.

355 are supposed to deliver more noise reduction than the r1MWF-0 but they give higher WERs.

In the following, we investigate the rank-1 MWF variants and their WERs achieved on the noisy acoustic model trained on utterances from all 6 channels. In Fig. 4, the relation between the WERs and the speech distortion scores is

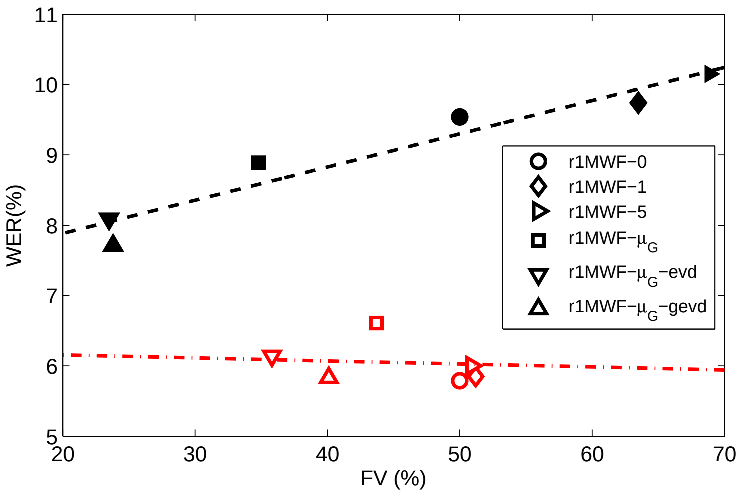

Figure 5: Relation between FV and WER for the real (solid black markers) and simulated (hollow red markers) data. The dashed lines show the linear regression results separately. A line with positive slope means positive correlation.

shown. The frequency-weighted log-spectral Signal Distortion (SD) metric [36] is defined as

$$\mathrm{SD} = \frac{1}{L}\sum_{l=1}^{l=L}\sqrt{\sum_{k=1}^{k=K}\mathrm{ERB}(k)\left(10\log_{10}\frac{\phi_o}{\phi_i}\right)^2 \mathrm{d}k} \tag{31}$$

where L is the number of frames, ϕ_o and ϕ_i are respectively the processed speech power spectrum and the clean speech power spectrum, and $\mathrm{ERB}(k)$ is the frequency-weighting factor giving equal weight to each auditory critical band. The SD scores are computed and averaged on the simulated test data. We observe that the r1MWF-μ_{G} introduces much larger distortion than the r1MWF-0/1, from about 9 dB to 16 dB. But the WER only increases slightly. Clearly, there is no strong correlation between the two.

In order to explain the recognition performance, we investigate the variance of the input features corresponding to each HMM state in Fig. 5. The intuition is that smaller Feature Variance (FV) implies an easier classification task for the neural network acoustic model. We expect the constant residual noise power property of the r1MWF-μ_{G} to translate into a smaller FV for the processed speech. The HMM state corresponding to each feature vector is first obtained by forced alignment on enhanced data separately. Note that the alignments of the simulated data can be obtained using the clean speech, nevertheless, similar results are observed here. The FV is calculated over all the feature vectors belonging to each HMM state for each method

$$V(j) = \frac{1}{I} \sum_{i=1}^{I} \mathrm{Var}(i, j) \tag{32}$$

where $\mathrm{Var}(i, j)$ means the variance of the ith feature in the jth state. We pick the FV of the r1MWF-0 as a baseline and define the metric

$$\mathrm{FV} = \frac{100}{\sum_j c_j} \sum_{j=1}^{J} c_j \cdot \mathbb{I}(V_{\mathrm{test}}(j) > V_{\mathrm{baseline}}(j)) \tag{33}$$

that is the weighted percentage of states for which the FV is larger than the baseline. c_j denotes the number of occurrences of the jth state. $\mathbb{I}(\cdot)$ is an indicator function the value of which is 1 for true arguments and 0 for false. For a comparable method, the value is expected to be around 50%. On the real data, the r1MWF-1 and r1MWF-5 have higher percentages (62.7% and 68.4%) and corresponding higher WERs. For the r1MWF-μ_{G}-evd and r1MWF-μ_{G}-gevd, lower percentages (23.1% and 23.6%) correlate with lower WERs. However, the correlation is not always valid on the simulated data as shown by the r1MWF-μ_{G}: it has 43.7% states with smaller FV and yet a higher WER than the baseline.

The FV metric provides another view from the feature side to explain the performance of the constant residual noise filter. Note that a global scale factor only results in a shift in the 0th MFCC value and will not affect the feature variance. The computation of FV also avoids the time-consuming decoding procedure that is required for WER.

6. Conclusion

Multichannel linear filters are generally designed to improve the speech perceptual quality but not specifically to improve the speech recognition accuracy. As a matter of fact, the choice of the optimal filter may be different for different tasks. In the scenario of a single target source, the popular SDW-MWF can be formulated as the rank-1 MWF. We derived a family of rank-1 MWF variants and evaluated their performance for speech recognition in multiple noisy environments. We defined a constant residual noise power constraint to find the trade-off parameter which links the rank-1 MWF filter and the GEV beamformer. We showed that this constraint brings more speech distortion, however, it benefits the speech recognition performance on the real data. To fulfill the underlying rank-1 assumption, speech covariance matrix reconstruction is proposed. The reconstruction based on eigenvectors or generalized eigenvectors subsequently improves the recognition accuracy. With experiments conducted on the CHiME-4 dataset, the final r1MWF-μ_{G}-gevd filter achieved a 40% relative WER reduction compared with the baseline WDAS beamformer on the real test set and a 15% relative WER reduction compared with the GEV-BAN method. For future research, we would like to see how the performance is impacted for corpora with higher reverberation time where the narrowband approximation becomes more erroneous.

In the speech recognition task, it is observed that multi-condition noisy training works well and sometimes outperforms retraining with enhanced data. So when new signal processing methods are applied, a reasonable practice is to process only the test data. Another finding is that the speech perceptual quality is not straightforwardly related to the speech recognition performance. An investigation from the perspective of feature variance is provided. The work puts forward the need for novel signal or feature metrics that correlate better with the WER.

7. Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by the China Scholarship Council (No. 201604910623). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

References

[1] J. Li, L. Deng, Y. Gong, R. Haeb-Umbach, An overview of noise-robust automatic speech recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 22 (4) (2014) 745–777.

[2] J. Benesty, J. Chen, Y. Huang, B. Rafaely, Microphone array signal processing, Journal of the Acoustical Society of America 125 (6) (2009) 4097–4098.

[3] M. Brandstein, D. Ward, Microphone Arrays: Signal Processing Techniques and Applications, Springer Science & Business Media, 2013.

[4] X. Anguera, C. Wooters, J. Hernando, Acoustic beamforming for speaker diarization of meetings, IEEE Transactions on Audio, Speech, and Language Processing 15 (7) (2007) 2011–2022.

[5] K. Kumatani, J. McDonough, B. Raj, Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors, IEEE Signal Processing Magazine 29 (6) (2012) 127–140.

[6] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, R. Maas, The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2013, pp. 1–4.

[7] J. Barker, R. Marxer, E. Vincent, S. Watanabe, The third CHiME speech separation and recognition challenge: Dataset, task and baselines, in: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015, pp. 504–511.

[8] S. Gannot, D. Burshtein, E. Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech, IEEE Transactions on Signal Processing 49 (8) (2001) 1614–1626.

[9] T. Yoshioka, T. Nakatani, Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening, IEEE Transactions on Audio, Speech, and Language Processing 20 (10) (2012) 2707–2720.

[10] T. Van den Bogaert, S. Doclo, J. Wouters, M. Moonen, Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids, The Journal of the Acoustical Society of America 125 (1) (2009) 360–371.

[11] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, D. Yu, Deep beamforming networks for multi-channel speech recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5745–5749.

[12] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, M. Bacchiani, Neural network adaptive beamforming for robust multichannel speech recognition, in: Proc. Interspeech, 2016, pp. 1976–1980.

[13] J. Heymann, L. Drude, R. Haeb-Umbach, Neural network based spectral mask estimation for acoustic beamforming, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 196–200.

[14] S. Sivasankaran, A. A. Nugraha, E. Vincent, J. A. Morales-Cordovilla, S. Dalmia, I. Illina, A. Liutkus, Robust ASR using neural network based speech enhancement and feature simulation, in: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015, pp. 482–489.

[15] A. A. Nugraha, A. Liutkus, E. Vincent, Multichannel audio source separation with deep neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (9) (2016) 1652–1664.

[16] J. Heymann, L. Drude, A. Chinaev, R. Haeb-Umbach, BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge, in: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015, pp. 444–451.

[17] H. Erdogan, T. Hayashi, J. R. Hershey, T. Hori, C. Hori, W.-N. Hsu, S. Kim, J. Le Roux, Z. Meng, S. Watanabe, Multi-channel speech recognition: LSTMs all the way through, in: Workshop on Speech Processing in Everday Environments, 2016.

[18] H. Cox, R. M. Zeskind, M. Owen, Robust adaptive beamforming, IEEE Transactions on Acoustics, Speech, and Signal Processing 35 (10) (1987) 1365–1376.

[19] E. Warsitz, R. Haeb-Umbach, Blind acoustic beamforming based on generalized eigenvalue decomposition, IEEE Transactions on Audio, Speech, and Language Processing 15 (5) (2007) 1529–1539.

[20] S. Doclo, M. Moonen, GSVD-based optimal filtering for single and multimicrophone speech enhancement, IEEE Transactions on Signal Processing 50 (9) (2002) 2230–2244.

[21] A. Spriet, M. Moonen, J. Wouters, Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction, Signal Processing 84 (12) (2004) 2367–2387.

[22] S. Doclo, A. Spriet, J. Wouters, M. Moonen, Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction, Speech Communication 49 (7) (2007) 636–656.

[23] R. Serizel, M. Moonen, B. Van Dijk, J. Wouters, Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants, IEEE/ACM Transactions on Audio, Speech, and Language Processing 22 (4) (2014) 785–799.

[24] J. R. Jensen, J. Benesty, M. G. Christensen, Noise reduction with optimal variable span linear filters, IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (4) (2016) 631–644.

[25] J. Benesty, J. Chen, Y. Huang, Noncausal (frequency-domain) optimal filters, in: Microphone Array Signal Processing (2008) 115–137.

[26] M. Souden, J. Benesty, S. Affes, On optimal frequency-domain multichannel linear filtering for noise reduction, IEEE Transactions on Audio, Speech, and Language Processing 18 (2) (2010) 260–276.

[27] J. Benesty, M. Souden, J. Chen, A perspective on multichannel noise reduction in the time domain, Applied Acoustics 74 (3) (2013) 343–355.

[28] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, A consolidated perspective on multimicrophone speech enhancement and source separation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (4) (2017) 692–730.

[29] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, R. Marxer, An analysis of environment, microphone and data simulation mismatches in robust speech recognition, Computer Speech & Language, 2016.

[30] S. Braun, K. Kowalczyk, E. A. Habets, Residual noise control using a parametric multichannel Wiener filter, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 360–364.

[31] S. Markovich, S. Gannot, I. Cohen, Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals, IEEE Transactions on Audio, Speech, and Language Processing 17 (6) (2009) 1071–1086.

[32] D. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

[33] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167, 2015.

[34] J. Barker, R. Marxer, E. Vincent, S. Watanabe, The third CHiME speech separation and recognition challenge: analysis and outcomes, Computer Speech & Language, 2016.

[35] B. Cornelis, M. Moonen, J. Wouters, Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors, IEEE Transactions on Audio, Speech, and Language Processing 19 (5) (2011) 1368–1381.

[36] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech, The Journal of the Acoustical Society of America 130 (5) (2011) 3013–3027.