A probability weighted beamformer for noise robust ASR

Suliang Bu[1]*, Yunxin Zhao*[1]*, Mei-Yuh Hwang*[2]*, Sining Sun*[3]

[1]Dept. of Electrical Engineering and Computer Science, University of Missouri-Columbia, USA
[2]Mobvoi AI Lab, Redmond WA, USA
[3]Sch. of Computer Science, Northwestern Polytechnical University, Xi'an, China

sbkc6@mail.missouri.edu, zhaoy@missouri.edu, mhwang@mobvoi.com, snsun@nwpu-aslp.org

Abstract

We investigate a novel approach to spatial filtering that is adaptive to conditions at different time-frequency (TF) points for noise removal by taking advantage of speech sparsity. Our approach combines a noise reduction beamformer with a minimum variance distortionless response (MVDR) beamformer or Generalized Eigenvalue (GEV) beamformer through TF posterior probabilities of speech presence (PPSP). To estimate PPSP, we study both statistical model-based and neural network based methods, where in the former, we use complex Gaussian mixture modeling (CGMM) on temporally augmented spatial spectral features, and in the latter, we use neural network (NN) based TF masks to initialize speech and noise covariance matrices in CGMM. We have conducted experiments on CHiME-3 task. On its real noisy speech test set, our methods of feature augmentation, TF dependent spatial filter, and NN-based mask initialization on covariances for CGMM have yielded relative word error rate (WER) reductions cumulatively by 8%, 16%, and 25% over the original CGMM based MVDR. On the real test data, the three methods have also produced consistent WER reductions when replacing MVDR by GEV.

Index Terms: noise robust speech recognition, MVDR beamformer, GEV beamformer, noise reduction

1. Introduction

The performance of an automatic speech recognition (ASR) system may degrade significantly in noisy environments. Microphone array beamforming has shown great potential in improving ASR performance in noise [1, 2, 3, 4, 5]. In narrowband beamforming, to estimate a steering vector (SV) or a spatial filter, eigen analyses can be made on the spatial spectral covariance matrices of speech and noise. In the minimum variance distortionless response (MVDR) beamformer of [6, 4, 7], a SV was estimated as the eigenvector associated with the largest eigenvalue of the speech spatial covariance matrix in each frequency bin. In the Generalized Eigenvalue beamformer (GEV) of [8, 9], the filter was estimated as the generalized eigenvector with the largest eigenvalue involving both speech and noise spatial covariance to maximize signal-to-noise ratio (SNR). Speech and noise spatial covariance matrices are usually estimated by using time-frequency (TF) masks of speech. The TF masks can be obtained by methods of statistical models [10, 6, 11] or neural networks (NN) [9, 12, 13]. The former does not need stereo training data and it usually estimates masks independently for each TF point, while the latter may require stereo data and it jointly estimates masks over all frequency bins.

MVDR and GEV beamformers are well established as effective methods for enhancing speech from noise. However, in real conditions, noticeable noises may still exist in their enhanced signals. To further remove noise, an alternative method is the speech distortion weighted multichannel Wiener filtering (SDW-MWF) [14, 15, 16], a generalization of Multichannel Wiener filtering (MWF), which provides a tradeoff between noise reduction and speech distortion. SDW-MWF can be viewed as a MVDR followed by a time-invariant post-filter [16] that scales the MVDR output in each frequency bin, which may not be optimal for sparse signal and time-varying noise.

In this work, we investigate a TF-dependent spatial filtering approach and adapt the spatial filter design to speech and noise conditions at different TF points. To do so, we first derive separate filters with different aims: one aiming at capturing target speech in a desired direction, which can be accomplished by MVDR or GEV, and another aiming at maximally reducing noise, which can be accomplished by a linear filter derived from noise spatial covariance. We then combine the speech capture and noise reduction filters via the posterior probability of speech presence (PPSP) at each TF point to generate a TF-dependent spatial filter. Furthermore, to improve the estimation of statistical model based PPSP, we incorporate a differential temporal context to spatial spectral vectors in CGMM, and derive parameter updating formula based on Expectation-Maximization (EM) algorithm. Additionally, we investigate using the NN-based TF masks of [9] to improve the initialization of speech and noise covariance matrices for CGMM.

In Section 2, we briefly review MVDR, GEV, SDW-MWF, and two methods of TF mask estimation of [6, 9]. In Section 3, we describe the proposed TF-dependent filter and differential context features for CGMM. We present experimental results on CHiME-3 [17] in Section 4, and draw conclusions in Section 5.

2. MVDR, GEV, SDW-MWF, and TF masks

In this paper, we use bold font for vectors and regular font for scalars, with matrices specified explicitly.

2.1. MVDR and GEV beamforming

Let $\boldsymbol{y}_{f,t} = [y_{f,t,1}, ..., y_{f,t,M}]^T$ denote the signal vector from M microphones, where $y_{f,t,i}$ denotes the i-th microphone signal at frequency f and time t, and $(.)^T$ denotes transpose. MVDR minimizes the output energy while keeping a fixed gain in the direction of the desired signal [18], whereas GEV maximizes SNR in the output signal at the expense of speech distortion [8].

Given the spatial covariance matrices of speech and noise $\boldsymbol{\Phi}_{xx}(f)$ and $\boldsymbol{\Phi}_{nn}(f)$, the GEV filter is the eigenvector with the largest eigenvalue of $\boldsymbol{\Phi}_{nn}^{-1}(f)\boldsymbol{\Phi}_{xx}(f)$. On the other hand, given a unit gain on the desired signal and a SV $\boldsymbol{h}_f$, the MVDR filter is

$$\boldsymbol{w}_{\text{MVDR},f} = \frac{\boldsymbol{\Phi}_{nn}^{-1}(f)\boldsymbol{h}_f}{\boldsymbol{h}_f^H \boldsymbol{\Phi}_{nn}^{-1}(f)\boldsymbol{h}_f} \quad (1)$$

Given a spatial linear filter $\boldsymbol{w}_f$, the enhanced signal $\hat{x}_{f,t}$ is ob-

tained by

$$\hat{x}_{f,t} = \boldsymbol{w}_f^H \, \boldsymbol{y}_{f,t} \tag{2}$$

with $(.)^H$ the conjugate transpose.

2.2. SDW-MWF

Using Woodbury identity [19], SDW-MWF can be decomposed into MVDR followed by a postfiltering as [16]:

$$\boldsymbol{w}_{\text{SDWMWF},f} = \boldsymbol{w}_{\text{MVDR},f} \cdot \frac{\sigma_{dx,f}^2}{\sigma_{dx,f}^2 + \mu\sigma_{dn,f}^2} \tag{3}$$

where $\sigma_{dx,f}^2$ and $\sigma_{dn,f}^2$ denote the speech and noise variances in the MVDR output signal, respectively, and μ is the tradeoff parameter between speech distortion and noise reduction: larger μ results in more noise reduction at the cost of larger speech distortion.

2.3. CGMM-based mask estimation

For statistical model based TF mask estimation, we adopt the CGMM method in [6]. Let $\boldsymbol{y}_{f,t}$, $\boldsymbol{x}_{f,t}$ and $\boldsymbol{n}_{f,t}$ denote an observed signal, speech signal[1], and noise signal at (f,t), respectively, with $\boldsymbol{x}_{f,t} = s_{f,t}^x \boldsymbol{r}_f^x$ and $\boldsymbol{n}_{f,t} = s_{f,t}^n \boldsymbol{r}_f^n$, where $s_{f,t}^x$ is the speech component, and $\boldsymbol{r}_f^x$ is the acoustic transfer function vector (ATF) from the speech source to the M microphones, and $s_{f,t}^n$ and $\boldsymbol{r}_f^n$ are defined similarly.

The variables $s_{f,t}^x$ and $s_{f,t}^n$ are assumed to have zero-mean complex Gaussian distributions, i.e., $s_{f,t}^x \sim \mathcal{CN}(0, \phi_{f,t}^x)$ and $s_{f,t}^n \sim \mathcal{CN}(0, \phi_{f,t}^n)$, with $\phi_{f,t}^x$ and $\phi_{f,t}^n$ the variance of speech and noise, respectively. Thus, $\boldsymbol{x}_{f,t}$ and $\boldsymbol{n}_{f,t}$ are modeled as $\boldsymbol{x_{f,t}} \sim \mathcal{CN}(\boldsymbol{0}, \phi_{f,t}^x \boldsymbol{R}_f^x)$ and $\boldsymbol{n_{f,t}} \sim \mathcal{CN}(\boldsymbol{0}, \phi_{f,t}^n \boldsymbol{R}_f^n)$, where $\boldsymbol{R}_f^x = \boldsymbol{r}_f^x (\boldsymbol{r}_f^x)^H$ and $\boldsymbol{R}_f^n = \boldsymbol{r}_f^n (\boldsymbol{r}_f^n)^H$. Accordingly, $\boldsymbol{y}_{f,t}$ is modeled by a CGMM with two-components, one for speech and the other for noise. In practice, to accommodate for variations in speaker and microphone positions, $\boldsymbol{R}_f^x$ and $\boldsymbol{R}_f^n$ are treated as full-rank covariance matrices [20].

The CGMM parameters are estimated by EM algorithm. For the speech component, the model parameters are iteratively updated as:

$$\phi_{f,t}^x = \left[\boldsymbol{y}_{f,t}^H \left(\boldsymbol{R}_f^x \right)^{-1} \boldsymbol{y}_{f,t} \right] / M \tag{4}$$

$$\boldsymbol{R}_f^x = \frac{1}{\sum_t \lambda_{f,t}^x} \sum_t \frac{\lambda_{f,t}^x}{\phi_{f,t}^x} \boldsymbol{y}_{f,t} \, \boldsymbol{y}_{f,t}^H \tag{5}$$

$$\lambda_{f,t}^x = \frac{w_f^x \cdot p(\boldsymbol{y}_{f,t} | \boldsymbol{0}, \phi_{f,t}^x \boldsymbol{R}_f^x)}{w_f^x \cdot p(\boldsymbol{y}_{f,t} | \boldsymbol{0}, \phi_{f,t}^x \boldsymbol{R}_f^x) + w_f^n \cdot p(\boldsymbol{y}_{f,t} | \boldsymbol{0}, \phi_{f,t}^n \boldsymbol{R}_f^n)} \tag{6}$$

where w_f^x and w_f^n are the mixture weights. When EM converges, the posterior probability of speech, $\lambda_{f,t}^x$, is taken as the local PPSP in our proposed local filter method (Section 3). The noise parameters are updated similarly.

2.4. NN-based mask estimation

For NN based mask estimation, we review the method of bidirectional long-short term memory (BLSTM) network in [9, 12] due to its good ASR performance. During its noise-aware training, binary masks are used as training targets. The ideal binary mask for speech, IBM_X, and noise, IBM_N, are defined by

$$IBM_X(t,f) = \begin{cases} 1, & \|x\|/\|n\| > 10^{th_X(f)} \\ 0, & else, \end{cases} \tag{7}$$

$$IBM_N(t,f) = \begin{cases} 1, & \|x\|/\|n\| < 10^{th_N(f)} \\ 0, & else, \end{cases} \tag{8}$$

where $\|\cdot\|$ is the Euclidean norm, $th_X(f)$ and $th_N(f)$ are two different thresholds. During test, the masks obtained for each channel are then condensed to a single speech mask and a single noise mask using a median operation that reduces the effect of outliers, such as broken channels. The speech and noise masks are used as weights on spatial spectral vectors in computing the spatial covariance matrices of speech and noise.

3. Proposed methods

We first describe the proposed TF-dependent spatial filter, and then explain the proposed temporal augmentation to the spatial spectral features for CGMM. In the following, we deal with narrowband beamformers by default, so the frequency index f is omitted when no ambiguity occurs.

3.1. Speech probability weighted spatial filter

Conventional beamformers, like MVDR or GEV, often use a time-invariant filter in each frequency bin [5, 9, 4]. Such filters are desired if the target signal always exists in the frequency band. However, this is not true for speech as it is sparse in the TF domain. Therefore, the beamformed signals often require a followup postfiltering to reduce residue noise.

Here, to further reduce noises, we exploit the speech property of sparsity and investigate the following approach to spatial filter design: depending on the PPSP at a TF point, $\lambda_{f,t}^x$, we swing the spatial filtering objective between speech capture (like MVDR or GEV filters) and noise reduction. For noise reduction, we consider the following filter $\boldsymbol{w_n}$:

$$\boldsymbol{w_n} = \underset{\boldsymbol{w}}{\text{argmin}} \quad \boldsymbol{w^H \Phi}_{nn} \boldsymbol{w}, \quad st. \quad \boldsymbol{w^H w} = 1 \tag{9}$$

The eigenvector with the minimum eigenvalue of $\boldsymbol{\Phi}_{nn}$ is a solution to Eq. (9). Admittedly, we can switch between $\boldsymbol{w}_{\text{MVDR}}$/$\boldsymbol{w}_{\text{GEV}}$ and $\boldsymbol{w_n}$ with reference to a threshold of PPSP. However, to avoid tuning the threshold, we adopt a soft-switching approach. Specifically, we define the following spatial filter for each (f,t) point:

$$\boldsymbol{w_t} = (\boldsymbol{w}_*)^{p_t} \odot (\boldsymbol{w_n})^{1-p_t} \tag{10}$$

where $\boldsymbol{w}_*$ can be $\boldsymbol{w}_{\text{GEV}}$ or $\boldsymbol{w}_{\text{MVDR}}$, p_t is the PPSP at (f,t) point, and $(.)^p$ and $\odot$ denote element-wise power and multiplication operations, respectively. Clearly, if $p_t = 1$, $\boldsymbol{w_t}$ equals to $\boldsymbol{w}_*$; if $p_t = 0$, $\boldsymbol{w_t}$ is $\boldsymbol{w_n}$; for intermediate values of p_t's, the combined local filter would have a mixed effect on speech capture and noise reduction. For the i'th microphone channel, the filter's phase is a weighted linear interpolation of the phases of $\boldsymbol{w}_{*,i}$ and $\boldsymbol{w}_{n,i}$, and the magnitude is the weighed geometric average of the magnitudes of $\boldsymbol{w}_{*,i}$ and $\boldsymbol{w}_{n,i}$, with the weights being p_t and $1 - p_t$, respectively.

In narrowband beamformers, usually only single speech and noise covariance matrices are used to estimate spatial filters in each frequency bin [5, 9, 4]. In this case, the filters do not adapt to time-varying noises that often occur in real conditions. However, by using PPSP's as the combining weight, our

[1] In [6], an observed signal was defined to be consisted of noisy speech and noise. To avoid confusions, we use the term of speech and noise instead.

proposed spatial filter (10) is able to change its objective from speech capture to noise removal. Although it no longer guarantees distortionless response in the desired signal, this weakness is compensated for by more effective noise reduction.

3.2. Spatial spectral feature augmentation in CGMM

As the accuracy of local PPSP, $\lambda^x_{f,t}$, is important to the above filter composition method, it is desired to improve the speech-noise discriminative power of CGMM. In [6, 7], only TF-specific spatial spectral vectors $\boldsymbol{y}_t$ were used in local CGMMs. Since neighboring spectra may provide additional discriminative information, we augment each center spatial spectral vector by its temporal context. Specifically, a first-order time difference of $\boldsymbol{y}_t$ with the step size l, $\boldsymbol{\Delta y}_t = \boldsymbol{y_{t+l}} - \boldsymbol{y_{t-l}}$, is also considered as a feature:

$$\boldsymbol{\Delta y}_t = \boldsymbol{\Delta x}_t + \boldsymbol{\Delta n}_t = \Delta s_t \boldsymbol{r}^x + \Delta n_t \boldsymbol{r}^n$$

We see that in $\boldsymbol{\Delta y}_t$, the ATFs of speech and noise remain unchanged. Therefore, $\boldsymbol{\Delta x}_t$ and $\boldsymbol{\Delta n}_t$ can be modeled by $\mathcal{CN}(\boldsymbol{0}, \Delta\phi^x_t \boldsymbol{R}^x_f)$, and $\mathcal{CN}(\boldsymbol{0}, \Delta\phi^n_t \boldsymbol{R}^n_f)$, respectively. For computational convenience, we adopt block-diagonal covariance matrices for speech and noise to model the augmented feature vector, $[\boldsymbol{y}^T_t \ \boldsymbol{\Delta y}^T_t]^T$, in CGMM. Specifically, for the speech component, its covariance matrix becomes:

$$\begin{bmatrix} \phi^x_{1,t}\boldsymbol{R}^x & \boldsymbol{0} \\ \boldsymbol{0} & \phi^x_{2,t}\boldsymbol{R}^x \end{bmatrix}$$

and its CGMM parameter update formulas are derived as :

$$\phi^x_{1,t} = \left[\boldsymbol{y}^H_t \left(\boldsymbol{R}^x\right)^{-1} \boldsymbol{y}_t\right] / M \tag{11}$$

$$\phi^x_{2,t} = \left[\boldsymbol{\Delta y}^H_t \left(\boldsymbol{R}^x\right)^{-1} \boldsymbol{\Delta y}_t\right] / M \tag{12}$$

$$\boldsymbol{R}^x = \frac{1}{2\sum_t \lambda^x_t} \sum_t \lambda^x_t \left(\frac{\boldsymbol{y}_t \boldsymbol{y}^H_t}{\phi^x_{1,t}} + \frac{\boldsymbol{\Delta y}_t \boldsymbol{\Delta y}^H_t}{\phi^x_{2,t}}\right) \tag{13}$$

For noise, its covariance matrix and its formulas for parameter updates are similarly defined and derived.

4. Experiments and Results

The CHiME-3 task covered four noisy environments: cafe (CAF), street (STR), public transport (BUS) and pedestrian area (PED). Real noisy speech data had 1600 utterances which were supplemented by 7138 simulated noisy speech utterances for acoustic model training. Test data also had real and simulated noisy speech and consisted of the 330 sentences as in the WSJ0 5k task. Data details are described in [17].

4.1. Experiment Setup

In the setup for speech recognition, we used the CHiME-3 baseline backend in Kaldi [21] without any modification.

In the setup for beamforming, we evaluated our proposed methods in two cases: one only used CGMM to estimate PPSP, the other used NN-based masks to initialize the speech and noise covariances for CGMM and the converged CGMM was used to estimate PPSP.

When CGMM alone was used to estimate PPSP, we largely followed the setting in [6] but added the following refinements. Before CGMM initialization, the microphone channel with the highest SNR was determined. Within the first and last 25 frames of this microphone signal and in each frequency bin, noise TF points were detected and used to initialize noise covariance, and these TF points were fixed as noise during EM iterations, while the rest points were used to initialize speech covariance. In feature augmentation, the step size l was set to 2 to avoid temporal overlap between contextual vectors (the frame shift was 25% of frame size). Note that the augmented features were not used in Eq.(2) or parameter initialization.

We adopt the NN-based masks in [9, 12] due to its reported good performance for ASR. Empirically, we found that these masks could not be used directly as PPSP in (10). This might be due to the 0-1 binary target setting in NN training (Eq. (7, 8)), and the separate estimation of speech and noise masks that does not guarantee the sum-to-one constraint. Although the mask scores computed during test were continuous and normalized within [0,1], they did not work well as PPSP's. On the other hand, to take advantage of the smooth TF masks produced by NN due to its estimating masks jointly across frequency bins, we investigated using the NN-based masks to initialize the speech and noise covariances for CGMM. Specifically, the NN-based masks were first used to detect noise TF points at the two ends of each utterance: points with a noise score larger than 0.9 were fixed as noise in EM iterations of CGMM. These noise points together with the rest NN-based masks were used for noise and speech covariance initialization in CGMM.

4.2. Experiment Results

Our ASR results are summarized in word error rate (WER) for the simulated and real test data. We first evaluated the proposed feature augmentation when CGMM-based masks were used alone, and compared it with the CHiME-3 baseline BeamformIt [22]. These results are given in Table 1, where "MVDRΔ" and "GEVΔ" denote using the augmented features.

Table 1: *WERs (%) of baseline, MVDR, GEV, with and without feature augmentation*

	eval simu					eval real				
	BUS	CAF	PED	STR	AVG	BUS	CAF	PED	STR	AVG
baseline	8.7	13.1	12.9	14.9	12.4	18.8	10.5	10.3	9.8	12.4
MVDR	4.8	6.5	**5.4**	7.7	6.1	16.6	8.4	6.8	8.3	10.0
MVDRΔ	4.4	5.4	5.6	8.2	5.9	15.6	**6.9**	**6.2**	8.1	9.2
GEV	4.6	5.3	5.6	7.8	5.8	14.0	7.4	7.0	**7.5**	9.0
GEVΔ	**4.2**	**4.8**	5.5	**7.4**	**5.5**	**12.2**	7.8	6.8	7.8	**8.6**

In Table 1, the average WERs by our MVDR on simulated and real data was 6.1% and 10.0%, respectively, which greatly lowered the baseline WERs, indicating the effectiveness of the approach of [6]. These two figures were better than the corresponding figures of 6.96% and 10.37% of [6] with five microphones used in beamforming. A possible reason was that our noise spatial covariance initialization was more informative than the identity matrix based initialization in [6]. On the other hand, GEV performed better than MVDR. One likely reason was the better numerical stability of GEV over MVDR [12]: while MVDR needed matrix inversion, GEV did not. In addition, we observed that in the GEV beamformed signals, the lower frequency components appeared to be attenuated appreciably, which was beneficial to conditions with strong low-frequency noise, like BUS. This might be another reason why GEV was better than MVDR in this task.

On the other hand, comparing MVDR with MVDRΔ or GEV with GEVΔ, feature augmentation further reduced average WERs, suggesting its benefit in boosting the discriminative power of CGMM. In the subsequent experiments, feature augmentation was used in CGMM by default.

In Table 2, we provide WER for the proposed TF-dependent filters, where MVDRΔ* and GEVΔ* indicate that MVDR and GEV filter were used in (10), respectively. In addition, we compared MVDRΔ* with SDW-MWF. Based on Eq. (3), our implementation of SDW-MWF was actually MVDRΔ followed by a post-filter, where the tradeoff parameter μ was set to 0.5 and 1, respectively, denoted as MWF0, and MWF1. In the post-filter, the speech and noise power spectrum density (PSD) estimation was based on [23]. For convenience of comparison, results for MVDRΔ and GEVΔ were repeated.

Table 2: *WER (%) of proposed local filtering and SDW-MWF*

	eval simu					eval real				
	BUS	CAF	PED	STR	AVG	BUS	CAF	PED	STR	AVG
MVDRΔ	4.4	5.4	5.6	8.2	5.9	15.6	6.9	6.2	8.1	9.2
MWF0	4.5	5.4	5.0	7.5	5.6	16.0	6.8	6.2	7.6	9.1
MWF1	4.2	5.4	5.9	7.9	5.9	16.8	6.8	6.2	8.2	9.5
MVDRΔ*	5.4	5.9	**5.3**	7.7	6.1	13.8	**6.1**	6.5	**7.2**	8.4
GEVΔ	**4.2**	**4.8**	5.5	**7.4**	**5.5**	**12.2**	7.8	6.8	7.8	8.6
GEVΔ*	5.0	6.1	6.0	7.6	6.2	12.4	6.7	**5.7**	7.7	**8.1**

Comparing MWF0, MWF1 with MVDRΔ, it appeared that SDW-MWF did not reduce WER significantly. A possible reason was that the noise being non-stationary and hence the noise PSD was difficult to model. On the other hand, comparing MVDRΔ with MVDRΔ*, or GEVΔ with GEVΔ*, we found that our TF-dependent filters worked effectively on real data. In comparison with MVDR in Table 1, MVDRΔ* got 16% relative WER reduction on real test data. To better understand the positive effect of the noise reduction filter $\boldsymbol{w}_n$ in WER reduction, we examined the estimated filter component values and found that the magnitude of the individual components correlated with the relative noise strength in the multi-channels: a larger magnitude was correlated with a lower level of noise in a channel. As the result, according to (10), a cleaner channel would make a larger contribution to the beamformed signal. That said, more careful analysis is still needed in a future study.

On the other hand, the local filter methods slightly increased WER on simulated data. Clearly, better performance on real data is more valuable for real applications. A further examination revealed that MVDR and GEV tended to remove noises much better in simulated data than in real data: the beamformed simulated data tended to have larger SNR than beamformed real data. As the result, noise corruption in simulated data was less an issue after MVDR or GEV, but speech distortion due to the TF-dependent filtering became noticeable, which might be significant if the PPSP's were inaccurate at some TF points. This points to the need for further improving the robustness and design of our filter in Eq. (10). One possibility along this line is to derive more accurate PPSP from the beamformed signal of MVDR or GEV, and then use the PPSP in Eq. (10).

Table 3: *WER (%) of local filtering with NN-based masks for CGMM initialization*

	eval simu					eval real				
	BUS	CAF	PED	STR	AVG	BUS	CAF	PED	STR	AVG
MVDRΔn	4.4	6.2	5.6	6.6	5.7	13.0	7.3	6.9	7.4	8.6
MVDRΔn*	**4.2**	6.2	5.9	6.7	5.7	11.8	**6.0**	5.9	6.5	7.5
GEVΔn	4.2	**5.4**	**5.2**	**5.9**	**5.2**	10.4	6.9	**5.5**	6.6	7.4
GEVΔn*	4.4	6.2	5.4	6.8	5.7	**9.5**	6.4	5.8	**6.4**	**7.0**

In Table 3, the NN-based TF masks were used to initialize the speech and noise covariance matrices for CGMM and all methods were tagged by "n" to indicate this setting. Comparing results of Table 3 with Table 2, we found that all methods gained benefit from this initialization as it led to better PPSP estimates. Compared with MVDR in Table 1, MVDRΔn* obtained 25% relative WER reduction on real test data. On the other hand, our GEVΔn on real test data had 7.4% WER, while in [12] GEV had a WER of 7.45%, which directly used the NN-based masks to calculate its filters. Although using NN-based masks to initialize CGMM did not affect WER of GEV, the probability weighted beamformer further reduced the WER to 7.0%.

Finally, we summarize the performance in WER on the real test data in Fig. 1 for MVDR/GEV and our proposed methods. With the successive introduction of feature augmentation, TF dependent filter, and NN mask based initialization, the WER is decreased progressively for both MVDR and GEV, and the total relative WER reduction was 25% for MVDR and 22% for GEV.

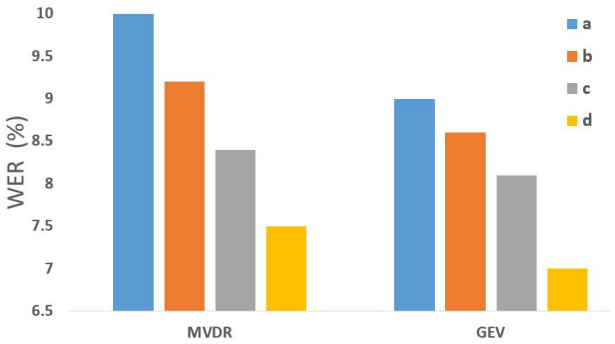

Figure 1: *WER comparison of (a) CGMM based MVDR/GEV; (b) with feature augmentation in CGMM; (c) TF dependent filter with feature augmentation; and (d) TF dependent filter with feature augmentation and NN mask based initialization*

It is worth noting that presumably, using soft masks as targets in NN training might produce mask scores more compatible with the PPSP's for filter composition in Eq. (10). On the other hand, unlike CGMM, NN-based methods such as [12] do not exploit the ATF model that may facilitate discrimination between speech and noise. Our approach of using NN masks to initialize CGMM provides a way to utilize both the spectral-temporal context-dependent scores provided by NN and the explicit ATF modeling by CGMM.

5. Conclusions

In this paper, we have introduced a TF-dependent spatial filter that focuses on speech capture or noise reduction dynamically according to PPSP at different TF points. This method takes into consideration of speech sparsity in TF domain and attempts to remove noise more aggressively than MVDR or GEV alone. To better estimate PPSP under CGMM, we have augmented spatial spectral vectors by their contextual vectors. We have further investigated using the NN-based TF masks to initialize the speech and noise covariance matrices for CGMM. We have achieved word error reductions with each of these methods. On the real test set of ChiME-3 task, our methods of feature augmentation, local spatial filter, and NN-based mask initialization on covariances for CGMM have cumulatively yielded relative word error rate reductions of 8%, 16%, and 25% over our implementation of CGMM based MVDR of [6]. The three methods have also produced consistent word error rate reductions when GEV was used in place of MVDR on real test data. In a future work, we plan to further improve the probability weighted beamformer and investigate its performance in heavier reverberation conditions than those of CHiME-3.

6. References

[1] K. Kumatani, T. Arakawa *et al.*, "Microphone array processing for distant speech recognition: Towards real-world deployment," in *APSIPA ASC*, 2012, pp. 1–10.

[2] L. Pfeifenberger, T. Schrank *et al.*, "Multi-channel speech processing architestures for noise robust speech recognition: 3-rd CHiME chanllenge results," in *Interspeech*, 2015.

[3] T. Menne, J. Heymann *et al.*, "The RWTH/UPB/FORTH system combination for the 4th CHiME challenge evaluation," in *The 4th IWSPEE*, 2016.

[4] T. Yoshioka, N. Ito *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *ASRU*, 2015.

[5] H. Erdogan, T. Hayashi *et al.*, "Multi-channel speech recognition: Lstms all the way through," in *CHiME-4 workshop*, 2016.

[6] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *ICASSP*, 2016.

[7] T. Higuchi, N. Ito *et al.*, "Online MVDR beamformer based on complex gaussian mixture model with spatial prior for noise robust asr," *IEEE/ACM Trans. ASLP*, vol. 25, no. 4, pp. 780–793, 2017.

[8] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE/ACM, Trans. ASLP*, vol. 15, no. 5, pp. 1529–1539, 2007.

[9] J. Heymann, L. Drude *et al.*, "BLSTM supported gev beamformer front-end for the 3rd CHiME challenge," in *ASRU*, 2015.

[10] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with dirichlet prior," in *ICASSP*, 2009.

[11] D. Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *ICASSP*, 2010.

[12] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, 2016.

[13] X. Xiao, S. Zhao *et al.*, "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition," in *ICASSP*, 2017.

[14] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

[15] S. Doclo, A. Spriet *et al.*, "Speech distortion weighted multichannel wiener filtering techniques for noise reduction," *Speech enhancement*, pp. 199–228, 2005.

[16] S. Gannot, E. Vincent *et al.*, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. ASLP*, vol. 25, no. 4, pp. 692–730, 2017.

[17] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *ASRU*, 2015.

[18] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer, Berlin-Heidelberg-New York, 2008.

[19] G. H. Golub and C. F. V. Loan, "Matrix computations," 1996.

[20] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE/ACM, Trans. ASLP*, 2010.

[21] P. Daniel, G. Arnab *et al.*, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.

[22] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE/ACM, Trans. ASLP*, vol. 15, no. 7, pp. 2011–2022, 2007.

[23] M. Souden, J. Chen *et al.*, "Gaussian model-based multichannel speech presence probability," *IEEE/ACM, Trans. ASLP*, vol. 18, no. 5, pp. 1072–1077, 2010.