MASK WEIGHTED STFT RATIOS FOR RELATIVE TRANSFER FUNCTION ESTIMATION AND ITS APPLICATION TO ROBUST ASR

Zhong-Qiu Wang[♪], *DeLiang Wang*[♪,♫]

[♪]Department of Computer Science and Engineering, The Ohio State University, USA
[♫]Center for Cognitive and Brain Sciences, The Ohio State University, USA
{wangzhon, dwang}@cse.ohio-state.edu

ABSTRACT

Deep learning based single-channel time-frequency (T-F) masking has shown considerable potential for beamforming and robust ASR. This paper proposes a simple but novel relative transfer function (RTF) estimation algorithm for microphone arrays, where the RTF between a reference signal and a non-reference signal at each frequency band is estimated as a weighted average of the ratios of the two STFT (short-time Fourier transform) coefficients of the speech-dominant T-F units. Similarly, the noise covariance matrix is estimated from noise-dominant T-F units. An MVDR beamformer is then constructed for robust ASR. Experiments on the two- and six-channel track of the CHiME-4 challenge show consistent improvement over a weighted delay-and-sum (WDAS) beamformer, a generalized eigenvector beamformer, a parameterized multi-channel Wiener filter, an MVDR beamformer based on conventional direction of arrival (DOA) estimation, and two MVDR beamformers both based on eigendecomposition.

Index Terms - relative transfer function estimation, beamforming, deep neural networks, robust ASR, CHiME-4

1. INTRODUCTION

Modern electronic devices typically contain multiple microphones for speech applications. Acoustic beamforming techniques based on microphone arrays have shown to be quite beneficial for robust ASR [1], [2]. With the support of multiple microphones, spatial information can be exploited and corrupted signals can be reconstructed with high noise reduction and at the same time with low speech distortions [3], [4]. Conventionally, acoustic transfer functions are estimated via DOA estimation and the knowledge of microphone geometry, and the noise covariance matrices are commonly computed directly from the leading and ending frames of an utterance. Recently, acoustic beamforming algorithms based on deep learning and T-F masking have gained popularity and demonstrated their potential in the CHiME-3 and CHiME-4 challenges [5], [6]. The key idea is to estimate a monaural T-F mask using deep neural networks so that the spatial covariance matrices of speech and noise can be derived for beamforming (see [7] for a recent review). In the winning solution of CHiME-3 [8], a two-component complex Gaussian mixture model is devised to identify T-F units containing both noise and speech, and noise alone. Then, the steering vector is estimated as the principal eigenvector of the speech covariance matrix derived from the identified T-F units. An MVDR beamformer is built for robust ASR afterwards. In the same challenge, Heymann *et al*. [9] proposes a bi-directional LSTM for mask estimation. They estimate one mask for each signal using only single-channel information and then combine them into one mask by median pooling. The beamforming weights are computed as the principal generalized eigenvector of the speech and noise covariance matrices. In [10], Erdogan *et al.* utilize the covariance matrices of speech and noise estimated from a neural network to drive a parameterized multi-channel Wiener filter (PMWF) [4]. Later, in the CHiME-4 challenge, almost all the top teams adopt the T-F masking based techniques for beamforming [11], [12], [13], [14], [15], [7]. Remarkable improvement in terms of ASR performance has been reported over the official WDAS beamformer [16] and the default MVDR beamformer driven by SRP-PHAT [17]. The major advantages of the T-F masking based approaches are attributed to their versatility and flexibility, as the learning machine only needs to learn how to estimate a continuous T-F mask, or determine the speech or noise dominance at each T-F unit during training, which is a well-defined and well-studied task in single channel speech separation and enhancement [18]. In addition, the same learned model and algorithmic pipeline can be directly applied to microphone arrays with any number of microphones, without using any knowledge of the underlying microphone geometry.

This paper proposes a simple yet effective algorithm for RTF estimation, which is based on the direct utilization of speech-dominant T-F units, without computing speech covariance matrices, performing any eigendecomposition, or estimating any gains or time delays, and therefore makes fewer assumptions. Intuitively, for two corresponding T-F units between two signals with both T-F units strongly dominated by speech, the RTF can be reasonably estimated as the ratio of the two STFT coefficients. To improve the robustness, our system estimates one RTF for each frequency by weighted pooling, where the weights are devised in a way such that the T-F regions strongly dominated by target speech get more weights. The noise covariance matrix is obtained in a similar way. With these two, an MVDR beamformer is built for robust ASR tasks. One critical step here is the accurate identification of speech- and noise-dominant T-F units. We employ deep neural networks (DNN) for mask estimation, as they have shown state-of-the-art performance for single-channel speech enhancement in noisy and reverberant environments [19], [20]. We evaluate our algorithm on the two- and six-channel task of the CHiME-4 challenge. Consistent improvement is observed over other strong beamformers in terms of ASR performance.

2. SYSTEM DESCRIPTION

We first use a DNN to estimate a T-F mask for every microphone signal. With the estimated masks, speech-dominant T-F units can be selected for RTF estimation and noise-dominant T-F units for noise covariance matrix estimation. An MVDR beamformer is then constructed for enhancement and log Mel filterbank features are extracted for robust ASR. We first introduce MVDR beamforming, and then detail the proposed approach for RTF estimation. We discuss mask estimation in Section 2.3.

2.1. MVDR Beamforming

Assuming that there is no or little reverberation, the physical model in the STFT domain is formulated as

$$\boldsymbol{y}(t,f) = \boldsymbol{c}(f)s(t,f) + \boldsymbol{n}(t,f) \tag{1}$$

where $\boldsymbol{y}(t,f)$ is the STFT vector of the received speech, $s(t,f)$ is the STFT value of the target speaker, and $\boldsymbol{n}(t,f)$ is the STFT vector of the received noise at a specific T-F unit. $\boldsymbol{c}(f)$ is the so-called steering vector or acoustic transfer function between the target source and microphones at every frequency channel. In our study, we assume that there is only one target source and its position is fixed within each utterance.

The MVDR beamformer [21] is to find a weight vector for every frequency, $\boldsymbol{w}(f)$, such that the target speech along the look direction is maintained, while the interference or noise from other directions is suppressed. Mathematically,

$$\begin{gathered}\boldsymbol{w}^{*}(f) = argmin_{\boldsymbol{w}(f)}\ \boldsymbol{w}(f)^{H}\boldsymbol{\Phi}_{n}(f)\boldsymbol{w}(f) \\ \text{subject to}\ \ \boldsymbol{w}(f)^{H}\boldsymbol{c}(f) = 1\end{gathered} \tag{2}$$

where $\boldsymbol{\Phi}_n(f)$ is the noise covariance matrix and $(\cdot)^H$ stands for conjugate transpose. The close-form solution can be computed as

$$\boldsymbol{w}^{*}(f) = \frac{\boldsymbol{\Phi}_n(f)^{-1}\boldsymbol{c}(f)}{\boldsymbol{c}(f)^{H}\boldsymbol{\Phi}_n(f)^{-1}\boldsymbol{c}(f)} \tag{3}$$

The beamformed signal is then obtained as

$$\hat{y}(t,f) = \boldsymbol{w}^{*}(f)^{H}\boldsymbol{y}(t,f) \tag{4}$$

As we can see, the key for MVDR beamforming is the accurate estimation of $\boldsymbol{c}(f)$ and $\boldsymbol{\Phi}_n(f)$.

2.2. RTF and Noise Covariance Matrix Estimation

A key concept behind T-F masking is that the signal within a T-F unit may be assigned to the dominant source [22], [23]. For a speech-dominant T-F unit, the noise level is so low that the physical model becomes

$$\boldsymbol{y}(t,f) = \boldsymbol{c}(f)s(t,f) + \boldsymbol{\varepsilon} \tag{5}$$

where $\boldsymbol{\varepsilon}$ represents a negligibly small term. In this case, the RTF with respect to a reference microphone at a specific T-F unit, $\bar{\boldsymbol{c}}(t,f)$, can be reasonably estimated as:

$$\bar{\boldsymbol{c}}(t,f) = \frac{\boldsymbol{c}(f)}{c^{ref}(f)} = \frac{\boldsymbol{c}(f)s(t,f)}{c^{ref}(f)s(t,f)} \approx \frac{\boldsymbol{y}(t,f)}{y^{ref}(t,f)} \tag{6}$$

To improve the robustness, we normalize $\bar{\boldsymbol{c}}(t,f)$ to unit length and perform weighted pooling within each frequency to obtain $\bar{\boldsymbol{c}}(f)$:

$$\bar{\boldsymbol{c}}(t,f) = \frac{\bar{\boldsymbol{c}}(t,f)}{\|\bar{\boldsymbol{c}}(t,f)\|} \tag{7}$$

$$\bar{\boldsymbol{c}}(f) = \frac{\sum_t \eta(t,f)\bar{\boldsymbol{c}}(t,f)}{\sum_t \eta(t,f)} \tag{8}$$

where $\eta(t,f)$ is a weight denoting the importance of the T-F unit. It is defined as

$$\eta(t,f) = \prod_{i=1}^{D} \mathfrak{I}\{\widehat{M}_i(t,f) > \theta\}(\widehat{M}_i(t,f) - \theta) \tag{9}$$

where $\widehat{M}_i$ is the estimated mask representing the energy portion of speech for the signal at microphone i, θ is a manually set threshold to filter out non-reliable T-F units, D is the number of microphones, and $\mathfrak{I}\{\cdot\}$ is the indicator function. We emphasize that the normalization in Eq. (7) leads to more robust estimation of $\bar{\boldsymbol{c}}(f)$, as it can remove the influence of diverse energy levels at different T-F units, and in addition reduce the effects due to potential microphone failures. Eq. (9) means that only the T-F units dominated by speech across all D microphone channels would be considered for RTF estimation and the higher the values in the estimated masks are, the more weights are placed. Note that we need to estimate a mask for the signal at every microphone.

Finally, we normalize $\bar{\boldsymbol{c}}(f)$ to have unit length to get our estimated RTF for MVDR beamforming.

$$\hat{\boldsymbol{c}}(f) = \frac{\bar{\boldsymbol{c}}(f)}{\|\bar{\boldsymbol{c}}(f)\|} \tag{10}$$

It should be noted that for the frequencies with no predicted speech-dominant T-F units, the noisy speech at the reference microphone is used as the output directly. We summate over all the values in each estimated mask and choose the microphone with the largest summation as the reference microphone. The rationale is that for the signal with the highest input SNR, the largest percentage of energy would normally be retained by our DNN based mask estimator.

Following [8], [9], the noise covariance matrix is estimated as

$$\widehat{\boldsymbol{\Phi}}_n(f) = \frac{\sum_t \xi(t,f)\boldsymbol{y}(t,f)\,\boldsymbol{y}(t,f)^{H}}{\sum_t \xi(t,f)} \tag{11}$$

where $\xi(t,f)$ is the weight representing the importance of the T-F unit for the noise covariance matrix estimation. In our study, it is defined as

$$\xi(t,f) = \prod_{i=1}^{D} \mathfrak{I}\{[1 - \widehat{M}_i(t,f)] > \gamma\}([1 - \widehat{M}_i(t,f)] - \gamma) \tag{12}$$

where γ is a tunable threshold to select noise-dominant T-F units for noise covariance matrix computation. We obtain the noise mask by simply subtracting the speech mask from one.

With Eq. (10) and (11), an MVDR beamformer can be derived for enhancement using Eq. (3). After enhancement results are obtained using Eq. (4), log Mel filterbank features are extracted and fed into acoustic models for decoding.

Our approach makes fewer assumptions than the other popular beamformers. Compared with traditional TDOA estimation approaches or the WDAS beamformer [16], [17], our approach estimates a complex and continuous gain directly rather than separately estimates a time delay and a gain for steering vector derivation. In addition, our approach does not require any knowledge of microphone geometry or compute speech covariance matrices. There is no eigendecomposition or post-filtering involved. The amount of computation is hence considerably less compared with the GEV beamformer [9], [24] or the MVDR beamformer [14].

2.3. Mask Estimation

The goal of mask estimation is to identify the energy portion of speech at each T-F unit. It plays a central role in our algorithm. Many studies have shown the effectiveness of DNN-based T-F masking on single channel enhancement [19], [25], [26], [27] and robust ASR [28], [29], [30] tasks. In our study, we train a DNN to estimate the ideal ratio mask (IRM) [25] defined in the power domain:

$$\mathrm{IRM}_i(t,f) = \frac{|c_i(f)s(t,f)|^2}{|c_i(f)s(t,f)|^2 + |n_i(t,f)|^2} \tag{13}$$

where $|c_i(f)s(t,f)|^2$ and $|n_i(t,f)|^2$ represent the speech energy and noise energy at time t and frequency f of microphone signal i, respectively. The DNN is trained to estimate the IRM at the central frame from the log power spectrogram features with a large context window. We use the mean square error as the loss function for training.

3. EXPERIMENTAL SETUP

Our experiments are conducted on the six- and two-channel task of the CHiME-4 challenge [17]. The six-channel CHiME-4 dataset reuses the data in WSJ0-5k and CHiME-3, and features one-, two-,

Table 1. Comparison of the ASR performance (%WER) of different beamformers (sMBR training and tri-gram LM for decoding) on the six-channel track.

Approaches	Covariance matrices	Beamforming weights	Post-filters	Dev. set		Test set	
				SIMU	REAL	SIMU	REAL
BeamformIt [33], [16]	None	See [16]	None	8.62	7.28	12.81	11.72
MVDR via SRP-PHAT [17]	$\hat{\boldsymbol{\Phi}}_n(f)$ from 400-800ms context	$\hat{\boldsymbol{c}}(f)$ via SRP-PHAT, see [17] $\hat{\boldsymbol{w}}(f) = \frac{\hat{\boldsymbol{\Phi}}_n(f)^{-1}\hat{\boldsymbol{c}}(f)}{\hat{\boldsymbol{c}}(f)^H\hat{\boldsymbol{\Phi}}_n(f)^{-1}\hat{\boldsymbol{c}}(f)}$	None	6.32	9.38	7.05	14.60
GEV beamformer [9], [24], [12]	$\hat{M} = median(\hat{M}_1,\dots,\hat{M}_D)$ $\hat{\boldsymbol{\Phi}}_n(f) = \frac{\sum_t(1-\hat{M}(t,f))\boldsymbol{y}(t,f)\boldsymbol{y}(t,f)^H}{\sum_t(1-\hat{M}(t,f))}$ $\hat{\boldsymbol{\Phi}}_s(f) = \frac{\sum_t \hat{M}(t,f)\boldsymbol{y}(t,f)\boldsymbol{y}(t,f)^H}{\sum_t \hat{M}(t,f)}$	$\hat{\boldsymbol{w}}(f) = \mathcal{P}\{\hat{\boldsymbol{\Phi}}_n(f)^{-1}\hat{\boldsymbol{\Phi}}_s(f)\}$ $\mathcal{P}\{\cdot\}$ – principal eigenvector	$g_{BAN}(f) =$ $\frac{\sqrt{\hat{\boldsymbol{w}}(f)^H\hat{\boldsymbol{\Phi}}_n(f)\hat{\boldsymbol{\Phi}}_n(f)\hat{\boldsymbol{w}}(f)/D}}{\hat{\boldsymbol{w}}(f)\hat{\Phi}_n(f)\hat{\boldsymbol{w}}(f)}$	5.79	5.84	6.70	7.97
PMWF-0 [10], [13], [34]	Same as above	$\hat{\boldsymbol{w}}(f) = \frac{\hat{\boldsymbol{\Phi}}_n(f)^{-1}\hat{\boldsymbol{\Phi}}_s(f)}{trace(\hat{\boldsymbol{\Phi}}_n(f)^{-1}\hat{\boldsymbol{\Phi}}_s(f))}\boldsymbol{u_f}$	None	6.05	5.86	8.04	8.43
MVDR via eigendecomposition I	Same as above	$\hat{\boldsymbol{c}}(f) = \mathcal{P}\{\hat{\boldsymbol{\Phi}}_s(f)\}$ $\hat{\boldsymbol{w}}(f) = \frac{\hat{\boldsymbol{\Phi}}_n(f)^{-1}\hat{\boldsymbol{c}}(f)}{\hat{\boldsymbol{c}}(f)^H\hat{\boldsymbol{\Phi}}_n(f)^{-1}\hat{\boldsymbol{c}}(f)}$	None	5.91	5.62	7.20	8.30
MVDR via eigendecomposition II [14]	$\hat{M}, \hat{\boldsymbol{\Phi}}_n(f)$ as above $\hat{\boldsymbol{\Phi}}_{sn}(f) = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}(t,f)\boldsymbol{y}(t,f)^H$ $\hat{\boldsymbol{\Phi}}_s(f) = \hat{\boldsymbol{\Phi}}_{sn}(f) - \hat{\boldsymbol{\Phi}}_n(f)$	Same as above	None	6.13	5.65	6.98	8.07
Proposed	$\hat{\boldsymbol{\Phi}}_n(f)$ as above	See Section 2.2 (not using Eq. (7))	None	5.65	5.49	6.44	7.89
	$\hat{\boldsymbol{\Phi}}_n(f)$ as in Eq. (11) and (12)	See Section 2.2 (not using Eq. (7))	None	5.65	5.45	6.40	7.68
	Same as above	See Section 2.2 (using Eq. (7))	None	5.64	**5.40**	6.23	**7.30**

and six-channel tasks. The six microphones are mounted on a tablet, with the second one in the rear and the other five in the front. It incorporates simulated utterances and real recordings from four challenging daily environments, i.e. bus, pedestrian area, cafe, and street, exhibiting significant training and testing mismatches in terms of speaker, noise and spatial characteristics, and containing microphone failures in around 12% of the real recordings. The training data contains 7,138 simulated and 1,600 real utterances, the development set consists of 1,640 simulated and 1,640 real utterances, and the test set includes 1,320 simulated and 1,320 real utterances. Each of the three real subsets is recorded with four different speakers. For the two-channel task, only the signals from randomly selected two of the front five channels are provided in the development and test set.

Our acoustic model is trained on all the noisy signals from all the six microphones, i.e. 7,138*6+1,600*5 utterances (~104h), except the second microphone signals in the real training set. We follow the common pipelines in the Kaldi toolkit to build our ASR systems, i.e. GMM-HMM training, DNN training, sMBR training, language model rescoring, and speaker adaptation. Our DNN-based acoustic model has seven hidden layers, each with 2,048 exponential linear units. There are 3,161 senone states in our system. The input feature is 40-dimensional log Mel filterbank feature with deltas and double deltas, and an 11-frame symmetric context window. Sentence level mean-variance normalization is performed before global mean-variance normalization. The dropout rates are set to 0.3. Batch normalization [31] and AdaGrad are utilized to speed up training. To compare our overall ASR system with other systems, we apply the challenge-standard five-gram language model and the RNN language model for lattice rescoring. In addition, we apply the unsupervised speaker adaptation algorithm proposed in our recent study [32] for run-time adaptation. We use the word error rates (WER) on the real utterances of the development set for parameter tuning.

The DNN for mask estimation is trained using all the 7,138*6 simulated utterances (~90h) in the training set. It has four hidden layers, each with 2,048 exponential linear units. Sigmoidal units are used in the output layer. The log power spectrogram features are mean normalized at the sentence level before global mean-variance normalization. We symmetrically splice 19 frames as the input to the DNN. The dropout rates are set to 0.1. The window length is 25ms and the hop size is 10ms. Pre-emphasis and hamming window are applied before performing 512-point FFT. The input dimension is hence 257*19 and the output dimension is 257. For the six-channel task, θ in Eq. (9) and γ in Eq. (12) are both simply set to zero. It should be noted that proper strategies are necessary to avoid numerical underflow in the computation of Eq. (9) and (12) as D gets large. To deal with microphone failures in the six-channel task, we first select a microphone signal that is most correlated with the remaining five signals, and then throw away the signals with less than 0.3 correlation coefficients with the selected microphone signal. The signals left are utilized for beamforming. For the two-channel task, θ and γ are both set to 0.5.

4. EVALUATION RESULTS

We compare the performance of our system with several other beamformers, each of which is detailed in Table 1. These beamformers have been previously applied to the CHiME-4 corpus and demonstrated strong performance. We use the acoustic model after sMBR training and the trigram language model for decoding, setting aside the contributions of backend processing. For all the masking based beamformers listed in Table 1, we use the same estimated masks from our DNN for a fair comparison.

The BeamformIt represents the official WDAS beamformer implemented using the BeamformIt toolkit [33], [16]. It uses the GCC-PHAT algorithm for time delay estimation and the cross-correlation function for gain estimation. The MVDR via SRP-PHAT algorithm [17] is another official baseline provided in the challenge. It uses the conventional SRP-PHAT algorithm for DOA estimation. The gains are assumed to be equal across different microphone channels. With these two, a steering vector is derived for MVDR beamforming. The noise covariance matrix is estimated from 400-800ms context immediately before each utterance. For the GEV beamformer, following the original algorithms [9], [24], [12], we combine the estimated masks using median pooling before computing the speech and noise covariance matrices. After that, generalized eigendecomposition is performed to obtain beamforming weights. A post-filter based on blind analytic normalization is further appended to reduce speech distortions. The PMWF-0 approach [4] uses matrix operations on speech and noise covariance matrices to compute the weights. It is later combined with T-F masking based approaches in [10], [13], [34]. $\boldsymbol{u_f}$ here is a one-hot vector denoting the index of the reference microphone. Note that we use the method detailed in Section 2.2 for reference microphone selection. For the MVDR via eigendecomposition I, we use the principal eigenvector of the speech covariance matrix as the estimation of the steering vector, assuming that the speech covariance matrix is a rank-one matrix, although this assumption may not hold when there is room reverberation, e.g. in the bus or cafeteria environment. In MVDR via eigendecomposition II, we follow the algorithm for covariance matrix calculation proposed in [8], [14], where the speech covariance matrix is obtained by subtracting the noise covariance matrix from the covariance matrix of noisy speech. The motivation is that the noise covariance obtained via pooling would be more accurate, as there are normally many frames containing only noises, which would be easily detected by the DNN. As we can see from the last entry of Table 1, our approach consistently outperforms the competing approaches in all the simulated and real subsets, especially on the real test set. Another comparison is provided in the first two entries of the proposed beamformer in Table 1, where we use the same noise covariance matrix as in the other beamformers together with the proposed RTF estimation algorithm for MVDR beamforming. We can see that using Eq. (11) and (12) to estimate the noise covariance matrix leads to a slight improvement (from 7.89% to 7.68% WER). In the last entry of the proposed beamformer, we use Eq. (7) to normalize $\boldsymbol{y}(t,f)/y^{ref}(t,f)$ before weighted pooling. Consistent improvement has been observed (from 7.68% to 7.30% WER). This is likely because of the normalization of diverse energy levels, and better handling of extremely large or small ratios caused by microphone failures.

We then use the task-standard language models to re-score the lattices, and perform run-time unsupervised speaker adaptation using our recently proposed algorithm in [32]. The results are reported in Table 2. The best result we have obtained on the real test set is 3.65% WER. We compare our results with the results from other systems, which are obtained using the same constrained RNNLM for decoding[1]. The winning system by Du *et al.* [11] obtains 3.24% WER on the real test set, but their overall system is an ensemble of multiple DNN- and deep CNN-based acoustic models trained from augmented training data. Their best single model trained on the augmented training data obtains 3.87% WER (according to the Table 3 of [11]). Their solution combines a clustering method, a DNN based method, and the feedbacks from backend ASR systems for mask estimation. The RTF is obtained via eigendecomposition. Then, a general sidelobe canceller with post-filtering is constructed for beamforming. The runner-up system by Heymann *et al.* [12] utilizes a complicated wide-residual bidirectional LSTM network for acoustic modeling and a bidirectional LSTM model for GEV beamforming. Their best result is 3.85% WER. We emphasize that our system uses simple DNNs for both mask estimation and acoustic modeling, and does not use any data augmentation, or model or system ensemble, as we aim for a simple and readily reproducible algorithm for RTF estimation. The results presented in Table 1 and 2 clearly demonstrate the effectiveness of the proposed algorithm.

For the two-microphone task, the RTF at each T-F unit is estimated as the ratio between a signal and the corresponding reference signal as in Eq. (6). The ASR results on the two-channel task are reported in Table 3 and 4. In Table 3, we use the same mask estimator for beamforming, and the same acoustic model after sequence training and the tri-gram language model for decoding. The results in each entry of Table 3 are obtained using the same algorithm detailed in the corresponding entry of Table 1. The only difference is D=2 now. For all the matrix inversions, we use the close-form solution of two-by-two matrices to avoid numeric issues. Similar trends as in Table 1 are observed, indicating that our approach also performs well in the two-microphone case. Nonetheless, the relative improvement over other beamformers is slightly smaller than in the six-channel task. Finally, we apply language model re-scoring and speaker adaptation to our system. The results are presented in Table 4. Similar trends to Table 2 are observed.

Table 2. Comparison of the ASR performance (%WER) with other systems (using the constrained RNNLM for decoding) on the six-channel track.

Approaches	Dev. set		Test set	
	SIMU	REAL	SIMU	REAL
Proposed beamformer + sMBR and tri-gram LM	5.64	5.40	6.23	7.30
+Five-gram LM and RNNLM	3.77	3.43	4.46	5.24
+Unsupervised speaker adaptation	2.69	2.70	3.09	3.65
Du *et al.* [11] (with model ensemble)	2.61	2.55	3.06	3.24
Best single model of [11]	-	2.88	-	3.87
Heymann *et al.* [12]	2.75	2.84	3.11	3.85

Table 3. Comparison of the ASR performance (%WER) of different beamformers (using sMBR training and tri-gram LM for decoding) on the two-channel track.

Approaches	Dev. set		Test set	
	SIMU	REAL	SIMU	REAL
BeamformIt	10.56	8.68	15.83	15.30
MVDR via SRP-PHAT	9.22	9.52	11.37	16.29
GEV beamformer	9.09	7.64	10.97	12.55
PMWF-0	9.00	7.66	12.33	12.85
MVDR via eigendecomposition I	9.19	7.66	11.69	12.47
MVDR via eigendecomposition II	9.05	7.50	10.79	12.42
Proposed beamformer	8.90	7.32	10.58	12.00
	8.74	**7.29**	10.50	11.84
	8.75	7.32	10.36	**11.81**

Table 4. Comparison of the ASR performance (%WER) with other systems (using the constrained RNNLM for decoding) on the two-channel track.

Approaches	Dev. set		Test set	
	SIMU	REAL	SIMU	REAL
Proposed beamformer + sMBR and tri-gram LM	8.74	7.29	10.50	11.84
+Five-gram LM and RNNLM	6.60	4.98	7.77	8.81
+Unsupervised speaker adaptation	4.95	3.84	5.60	6.10
Du *et al.* [11] (with model ensemble)	4.89	3.56	7.30	5.41
Best single model of [11]	-	4.05	-	6.87
Heymann *et al.* [12]	4.45	3.8	5.38	6.44

5. CONCLUDING REMARKS

We have proposed a novel approach for RTF estimation, which is based on the STFT ratios weighted by speech dominance. Although mathematically and conceptually much simpler, our approach has shown consistent improvement over several competitive methods on the six- and two-channel tasks of the CHiME-4 challenge. Future work would include estimating the RTF and noise covariance matrix adaptively, and evaluating the performance in terms of speech enhancement.

The T-F masking based beamforming approaches rely heavily on the availability of strongly speech-dominant T-F units, where the phase information is much less contaminated. In daily recorded utterances, the number of such T-F units is commonly sufficient for RTF estimation, and the DNN performs well at identifying them, although only energy features are used. Future research would be analyze and improve the performance in extremely low-SNR and highly-reverberant environments. One possible way is to use the ratio of enhanced spectrogram in Eq. (6).

6. ACKNOWLEDGEMENTS

This research was supported in part by an AFRL contract (FA8750-15-1-0279), an NSF grant (IIS-1409431), and the Ohio Supercomputer Center.

[1] See http://spandh.dcs.shef.ac.uk/chime_challenge/results.html (in the last table) for the ranking of all the results obtained when using the baseline RNNLM for decoding – as of Jan. 2017. Note that all the teams in the challenge were requested to report the decoding results using the official RNNLM.

7. REFERENCES

[1] K. Kumatani, A. Takayuki, K. Yamamoto, J. McDonough, B. Raj, R. Singh, and I. Tashev, "Microphone Array Processing for Distant Speech Recognition: Towards Real-World Deployment," in *Annual Summit and Conference on Signal and Information Processing*, 2012, pp. 1–10.

[2] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments," *arXiv preprint arXiv:1705.10874*, May 2017.

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, vol. 1. 2008.

[4] M. Souden, J. Benesty, and S. Affes, "On Optimal Frequency-domain Multichannel Linear Filtering for Noise Reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 260–276, 2010.

[5] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An Analysis of Environment, Microphone and Data Simulation Mismatches in Robust Speech Recognition," *Computer Speech and Language*, pp. 535–557, 2017.

[6] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The Third 'CHiME' Speech Separation and Recognition Challenge: Analysis and Outcomes," *Computer Speech and Language*, vol. 46, pp. 605–626, 2017.

[7] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," in *arXiv preprint arXiv:1708.07524*, 2017.

[8] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 System: Advances in Speech Enhancement and Recognition for Mobile Multi-Microphone Devices," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 436–443.

[9] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM Supported GEV Beamformer Front-End for the 3rd CHiME Challenge," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 444–451.

[10] H. Erdogan, J. Hershey, S. Watanabe, and M. Mandel, "Improved MVDR Beamforming using Single-channel Mask Prediction Networks," in *Proceedings of Interspeech*, 2016, pp. 1981–1985.

[11] J. Du, Y. Tu, L. Sun, F. Ma, H. Wang, and J. Pan, "The USTC–iFlytek System for CHiME-4 Challenge," in *Proceedings of CHiME-4*, 2016, pp. 36–38.

[12] J. Heymann and R. Haeb-Umbach, "Wide Residual BLSTM Network with Discriminative Speaker Adaptation for Robust Speech Recognition," in *Proceedings of CHiME-4*, 2016.

[13] H. Erdogan, T. Hayashi, J. Hershey, T. Hori, and C. Hori, "Multi-Channel Speech Recognition: LSTMs All the Way Through," in *Proceedings of CHiME-4*, 2016.

[14] X. Zhang, Z.-Q. Wang, and D. L. Wang, "A Speech Enhancement Algorithm by Iterating Single- and Multi-Microphone Processing and its Application to Robust ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 276–280.

[15] Z.-Q. Wang and D. L. Wang, "On Spatial Features for Supervised Speech Separation and its Application to Beamforming and Robust ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[16] T. Hori, Z. Chen, H. Erdogan, J. R. Hershey, J. Le Roux, V. Mitra, and S. Watanabe, "The MERL/SRI System for the 3rd CHiME Challenge Using Beamforming, Robust Feature Extraction, and Advanced Speech Recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 475–481.

[17] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The Third 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 504–511.

[18] D. Wang; and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," in *arXiv preprint arXiv:1708.07524*, 2017.

[19] Y. Wang and D.L. Wang, "Towards Scaling Up Classification-Based Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[20] Y. Zhao, Z.-Q. Wang, and D. L. Wang, "A Two-Stage Algorithm for Noisy and Reverberant Speech Enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5580–5584.

[21] O. Frost, "An Algorithm for Linearly Constrained Adaptive Array Processing," *Proceedings of the IEEE*, vol. 60, pp. 926–935, 1972.

[22] Ö. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1846, 2004.

[23] D. L. Wang and G. J. Brown, *Eds., Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.

[24] J. Heymann and L. Drude, "Neural Network Based Spectral Mask Estimation for Acoustic Beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 196–200.

[25] Y. Wang, A. Narayanan, and D. L. Wang, "On Training Targets for Supervised Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[26] Z.-Q. Wang, Y. Zhao, and D. L. Wang, "Phoneme-Specific Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 146–150.

[27] Z.-Q. Wang and D. L. Wang, "Recurrent Deep Stacking Networks for Supervised Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 71–75.

[28] Z.-Q. Wang and D. L. Wang, "Joint Training of Speech Separation, Filterbank and Acoustic Model for Robust Automatic Speech Recognition," in *Proceedings of Interspeech*, 2015, pp. 2839–2843.

[29] D. Bagchi, M. Mandel, Z. Wang, Y. He, A. Plummer, and E. Fosler-Lussier, "Combining Spectral Feature Mapping and Multi-channel Model-based Source Separation for Noise-robust Automatic Speech Recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015.

[30] Z.-Q. Wang and D. L. Wang, "A Joint Training Framework for Robust Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, Apr. 2016.

[31] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *arXiv preprint arXiv:1502.03167*, 2015.

[32] Z.-Q. Wang and D. L. Wang, "Unsupervised Speaker Adaptation of Batch Normalized Acoustic Models for Robust ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 4890–4894.

[33] X. Anguera and C. Wooters, "Acoustic Beamforming for Speaker Diarization of Meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2011–2022, 2007.

[34] X. Xiao, C. Xu, Z. Zhang, S. Zhao, S. Sun, S. Watanabe, L. Wang, L. Xie, D. L. Jones, E. S. Chng, and H. Li, "A Study of Learning Based Beamforming Methods for Speech Recognition," in *Proceedings of CHiME-4*, 2016.