ON SPATIAL FEATURES FOR SUPERVISED SPEECH SEPARATION AND ITS APPLICATION TO BEAMFORMING AND 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

This study integrates complementary spectral and spatial information to elevate deep learning based time-frequency masking and acoustic beamforming. Coherence and directional features are designed as additional input features for deep neural network training to remove diffuse noise and other directional interferences pervasive in real-world recordings. The diffuse and directional features are designed to be relatively invariant to the underlying target direction, number of microphones and microphone geometry. The estimated masks are then utilized to compute steering vectors and spatial covariance matrices for beamforming and robust ASR. Experiments on the CHiME-4 dataset demonstrate the effectiveness of the proposed approach.

Index Terms—beamforming, time-frequency masking, recurrent neural networks, robust ASR, CHiME-4

1. INTRODUCTION

Riding on the tide of deep learning, monaural (single-channel) speech separation and enhancement have made dramatic advance in recent years [1]. It has been suggested in many studies that deep learning based time-frequency (T-F) masking is capable of accurately determining speech or noise dominance at each T-F unit. Such masking resulted in, for the first time, substantial speech intelligibility improvements for hearing-impaired listeners [2], [3]. In the recent CHiME-3 and 4 challenges [4], [5], deep learning based T-F masking has been prominently employed for acoustic beamforming and robust ASR. The key idea is to use estimated T-F masks produced by deep neural networks (DNNs) to compute the speech and noise covariance matrices [6], [7] or steering vectors [8] that are necessary for accurate beamforming. Remarkable improvements in terms of robust ASR performance have been observed over conventional beamforming techniques [5], which typically use voice activity detection for covariance matrix estimation and direction of arrival estimation for steering vector computation.

In previous studies [6], [7], [8], [9], DNNs rely only on single-channel spectral information to estimate one T-F mask from every microphone signal. The independently estimated masks are then combined into a single mask, which is used as weights to compute spatial covariance matrices for beamforming. An advantage of using only single-channel information for T-F masking is that the DNN model trained this way is applicable regardless of the number of microphones and microphone geometry.

Different from these studies, we incorporate spatial features as extra inputs for model training in order to complement the spectral information for more accurate mask estimation. Through pooling spatial features over microphone pairs, the applicability of the proposed approach is also not impacted by the number of microphones and microphone geometry.

A key observation motivating our study is that a real-world auditory scene is usually comprised of one directional target speaker, a number of directional interference sources, and diffuse noise or room reverberation coming from many various directions. To distinguish the target directional source from the other directional sources, robust speaker localization is needed to determine the direction that contains the target speech. If the target direction is known, directional features indicating whether the signal at each T-F unit is from that direction can be utilized to extract the target speech from that direction, and filter out the noise and reverberation from other directions. In addition, diffuse noises and reflections caused by room reverberation reach microphones from various directions. This property can be exploited to derive interchannel coherence based features to indicate whether a T-F unit is dominated by a directional source. We emphasize that spectral information is still indispensable to suppressing noise or reverberation coming from directions around the target direction. To take all these considerations into account, we simply encode them as discriminative input features for mask estimation. This way, complementary spectral and spatial information can be utilized to boost speech separation.

There are previous efforts employing directional features for DNN based mask estimation. Most of the earlier studies assume that the target speech comes from a fixed direction, typically the front direction. In [10], interaural time differences (ITD), interaural level differences (ILD) and entire cross-correlation coefficients are used as primary features for sub-band ideal binary mask estimation in the cochleagram domain. Subsequently, Zhang and Wang [11] propose to combine ITD, ILD, and spectral features derived from a fixed beamformer for mask estimation. In [12], Araki *et al*. use ILD and interchannel phase differences (IPD) for de-noising auto-encoder training. Although these approaches show good performance when the target is in the front, they would likely not perform well when the target speech is from other directions. Other studies perform single-channel post-filtering or spatial filtering on beamforming outputs for further noise reduction [13], [14], [15]. For coherence-based features, previous attempts [16], [17], [18] in robust ASR are mainly focused on using them as post-filters for beamforming. Different from the previous studies, we incorporate spatial and spectral features as extra input for DNN based mask estimation. Doing it this way, we find that DNN can exploit the complementary nature of spectral and spatial information, leading to better mask estimation and subsequent covariance matrices estimation. This results in better beamforming and robust ASR performance.

2. SYSTEM DESCRIPTION

We first review the beamforming techniques based on T-F masking and DNNs, and then present two proposed spatial features for better mask estimation. The diffuse feature is designed to suppress diffuse noise sources, and the directional feature is designed to suppress interference sources not coming from the estimated target direction. We discuss mask estimation in Section 2.4. An example

of the diffuse and directional feature is shown in Fig. 1. As can be seen from Fig. 1.(c) and 1.(d), they are both very discriminative to the IRM depicted in Fig. 1.(b).

2.1. MVDR Beamforming based on T-F Masking

Suppose that there is no or little room reverberation, the physical model can be modeled in the following way in the STFT domain.

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

where $s(t,f)$ represents the STFT value of the sound source signal at time t and frequency f, and $\boldsymbol{y}(t,f)$, $\boldsymbol{c}(f)$, and $\boldsymbol{n}(t,f)$ stand for the STFT vector of the observed noisy signal, acoustic transfer function, and received noise.

Recent studies suggest that the spatial covariance matrices that are critical for beamforming [6], [19], [8], [20] can be accurately estimated using DNN based T-F masking. The key idea is that DNNs are capable of accurately determining the speech and noise dominance at each T-F unit, and therefore speech covariance matrices can be estimated from speech-dominant T-F regions, and noise covariance matrices from noise-dominant T-F regions. Remarkable improvement has been observed over conventional beamforming methods, which are commonly based on direction of arrival estimation and voice activity detection [21].

Following [6], [8], the speech and noise covariance matrices in our study are estimated in the following way.

$$\widehat{\boldsymbol{\Phi}}_s(f) = \frac{\sum_t \eta(t,f)\boldsymbol{y}(t,f)\,\boldsymbol{y}(t,f)^H}{\sum_t \eta(t,f)} \tag{2}$$

$$\widehat{\boldsymbol{\Phi}}_n(f) = \frac{\sum_t \xi(t,f)\boldsymbol{y}(t,f)\boldsymbol{y}(t,f)^H}{\sum_t \xi(t,f)} \tag{3}$$

where $(\cdot)^H$ represents conjugate transpose, and $\eta(t,f)$ and $\xi(t,f)$ are the weights denoting the importance of each T-F unit for the computation of the speech and noise covariance matrices. They are calculated using the products of multiple estimated masks.

$$\eta(t,f) = \prod_{i=1}^{D} \widehat{M}_i(t,f) \tag{4}$$

$$\xi(t,f) = \prod_{i=1}^{D} \left(1 - \widehat{M}_i(t,f)\right) \tag{5}$$

where D represents the number of microphones and $\widehat{M}_i(t,f)$ denotes the estimated speech portion within each T-F unit using deep learning. Eq. (5) means that only the T-F units strongly dominant by target speech across all the microphone signals would be used to compute the speech covariance matrix. The noise covariance matrix is computed in a similar way, where we use one minus the speech mask to obtain the noise mask.

The steering vector at each frequency, $\hat{\boldsymbol{c}}(f)$, is then estimated as the principal eigenvector of $\widehat{\boldsymbol{\Phi}}_s(f)$ [19]. The rationale is that $\widehat{\boldsymbol{\Phi}}_s(f)$ would be close to a rank-one matrix if it is well estimated, as the target speech is from a directional source. With the estimated $\widehat{\boldsymbol{\Phi}}_n(f)$ and $\hat{\boldsymbol{c}}(f)$, an MVDR beamformer is constructed:

$$\widehat{\boldsymbol{w}}(f) = \frac{\widehat{\boldsymbol{\Phi}}_n(f)^{-1}\hat{\boldsymbol{c}}(f)}{\hat{\boldsymbol{c}}(f)^H\widehat{\boldsymbol{\Phi}}_n(f)^{-1}\hat{\boldsymbol{c}}(f)} \tag{6}$$

and the enhancement result is obtained using

$$\hat{\boldsymbol{y}}(t,f) = \widehat{\boldsymbol{w}}(f)^H\boldsymbol{y}(t,f) \tag{7}$$

Log Mel filterbank feature is then extracted from $\hat{\boldsymbol{y}}(t,f)$ and directly fed into backend acoustic models for decoding.

2.2. Magnitude Squared Coherence

If a T-F unit is dominated by directional target speech, it would be coherent. Similarly, for a T-F unit dominated by diffuse noises or room reverberations, it would be non-coherent. This coherence property can be utilized to design spatial features that can differentiate directional and non-directional sources. Our study employs the magnitude squared coherence (MSC) [22] as additional features for DNN based T-F masking.

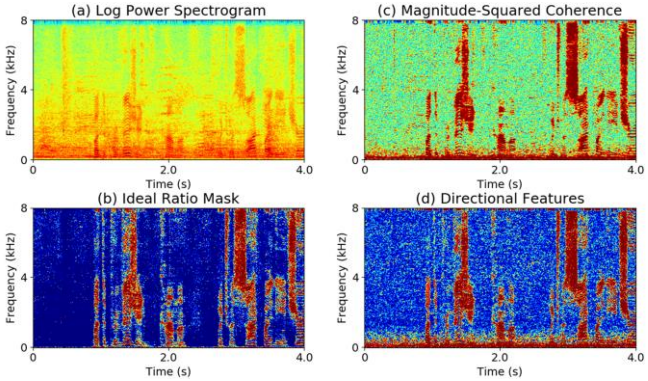

Fig. 1. Illustration of the spectral and spatial features using a simulated utterance (*029_029O0306_CAF*, first 4.0s) in the CHiME-4 dataset. (a) and (b) are obtained using its first channel, and (c) and (d) are computed using all the six microphone signals. In (d), the ideal ratio mask is used to derive the directional features.

To compute the MSC features, we first calculate the spatial covariance matrix of the noisy speech $\widehat{\Phi}_y(t,f)$ as

$$\widehat{\boldsymbol{\Phi}}_y(t,f) = \frac{1}{2w+1}\sum_{t'=t-w}^{t+w} \boldsymbol{y}(t',f)\boldsymbol{y}(t',f)^H \tag{8}$$

where w(=1 in this study) is the half-window length. Then we calculate the inter-channel coherence (ICC) between microphone signal i and j using

$$\mathrm{ICC}(i,j,t,f) = \frac{\widehat{\boldsymbol{\Phi}}_y(t,f,i,j)}{\sqrt{\widehat{\boldsymbol{\Phi}}_y(t,f,i,i)}\sqrt{\widehat{\boldsymbol{\Phi}}_y(t,f,j,j)}} \tag{9}$$

Finally, we pool over the ICCs of all the microphone pairs to obtain the MSC features:

$$\mathrm{MSC}(t,f) = \frac{1}{P}\sum_{i=1}^{D}\sum_{j=i+1}^{D} |\mathrm{ICC}(i,j,t,f)| \tag{10}$$

where $P = D(D-1)/2$ is the total number of microphone pairs and $|\cdot|$ extracts the magnitude. Note that the pooling operation here is a straightforward way to combine multiple microphone signals, and would significantly improve the quality of the MSC features.

Intuitively, if a T-F unit is strongly dominated by a directional source across all the microphone channels, the $|\mathrm{ICC}(i,j,t,f)|$ would be approximately equal to one. In contrast, if the T-F unit is strongly dominated by diffuse noises or room reverberations, the $|\mathrm{ICC}(i,j,t,f)|$ would be similar to a sinc function [21], which would become close to zero in high-frequency bands or when the microphone distance is large. An example is depicted in Fig. 1.(c). In low frequencies, the MSC features is not good enough, while very discriminative to the IRM in high frequencies.

In our study, we use the MSC features as extra input to our neural networks for mask estimation. It should be emphasized that the noise could also come from a directional source. Therefore it is beneficial to combine the MSC features with spectral features and the later introduced directional features for mask estimation. Note that one nice thing about the MSC feature is that it can be derived from noisy signals directly.

There are recent studies employing various coherence features for robust ASR. In [16], [17], [18], coherence based features are directly formulated as a post-filter to an MVDR beamformer for further noise reduction. In contrast, our study utilizes the MSC

feature for learning based time-frequency masking. This approach can leverage the representational power of DNN to improve mask estimation, and therefore benefit later beamforming.

2.3. Direction-Invariant Directional Features

Suppose that the true time delay between two microphone signals is known in advance, the observed phase difference at each T-F unit should be aligned with the time delay if the T-F unit is speech dominant. Based on this observation, the difference between the observed phase difference and the hypothesized phase difference would be indicative about whether the T-F unit is dominated by the speech from the hypothesized direction, or noises and inferences from the other directions [14], [17]. We use the following equation to derive the directional features for model training.

$$DF(t,f) = \frac{1}{P}\sum_{i=1}^{D}\sum_{j=i+1}^{D}\cos\left(\angle y_i(t,f) - \angle y_j(t,f) - \frac{2\pi f}{N} f_s \hat{\tau}_{i,j}\right) \quad (11)$$

where $\angle y_i(t,f) - \angle y_j(t,f)$ is the observed phase difference between microphone signal i and j at a specific time t and frequency f, and $\frac{2\pi f}{N} f_s \hat{\tau}_{i,j}$ is the hypothesized phase difference given the estimated time delay $\hat{\tau}_{i,j}$ in seconds. The 2π-periodic cosine operation can properly deal with potential phase-wrapping effects. If the time delay $\hat{\tau}_{i,j}$ is accurately estimated, the resulting feature would be close to one for speech dominant T-F units, while much smaller than one for noise-dominant T-F units. When there are more than two microphones ($D > 2$), we simply pool all the microphone pairs together to get the final feature.

Although recent studies suggested that time delay of arrival (TDOA) can be robustly estimated using time-frequency masking [23], our studies does not explicitly estimate TDOAs. Instead, we use the estimated steering vector from $\widehat{\mathbf{\Phi}}_s(f)$ to derive the spatial features, as the steering vector itself contains all the information regarding time delays and gain differences [21]. This strategy removes the need for a separate sound localization module and thus makes the system more simplified. In addition, it avoids the linear phase and planar wave assumption, which may not hold in practice. The spatial feature is computed as follows:

$$DF(t,f) = \frac{1}{P}\sum_{i=1}^{D}\sum_{j=i+1}^{D}\cos\left\{\angle y_i(t,f) - \angle y_j(t,f) - \left(\angle \hat{c}_i(f) - \angle \hat{c}_j(f)\right)\right\} \quad (12)$$

where $\angle \hat{c}_i(f)$ is the phase term extracted from the estimated steering vector, and therefore $\angle \hat{c}_i(f) - \angle \hat{c}_j(f)$ represents the estimated phase difference at the frequency f of microphone signal i and j. Eq. (12) measures whether the signal is from the estimated location. By using the spatial features for DNN training, we could extract the signal out from the estimated target direction.

There are previous efforts applying directional features for DNN training. Their directional features however are mainly designed for fixed target directions, and therefore are not invariant to target directions. In [12], the target speaker is assumed to be right in the front, so the phase difference for T-F units dominated by the target speech should be close to zero and $\cos\left(\angle y_i(t,f) - \angle y_j(t,f)\right)$ is directly used as the features to build an auto-encoder based speech enhancement system. Different from these studies, the features derived in this study is location-invariant. The invariance is achieved by subtracting the estimated phase difference from the observed phase difference so that a high value in the derived directional feature of a T-F unit would always indicate that the T-F unit is probably dominated by target speech.

Obviously, the directional features in Eq. (12) need an accurate estimation of the steering vector, $\hat{\boldsymbol{c}}(f)$, to yield high-quality and discriminative features. We use the principal eigenvector of $\widehat{\mathbf{\Phi}}_s(f)$ as the steering vector estimate. This strategy has been found to yield accurate steering vector estimates in many studies [19], [8].

2.4. Mask Estimation

Clearly, the performance of mask estimation plays a central role in the proposed algorithm. Our study trains a bi-directional long short-term memory (BLSTM) network to estimate the ideal ratio mask (IRM) [1], defined as the speech energy over the sum of speech energy and noise energy within each T-F unit, by minimizing the mean square error:

$$Loss = \sum_{t,f}\left\|\widehat{M}_i(t,f) - \frac{|c_i(f)s(t,f)|^2}{|c_i(f)s(t,f)|^2 + |n_i(t,f)|^2}\right\|^2 \quad (13)$$

It has been suggested in many studies that the mask estimator constructed using DNNs is capable of accurately determining the speech or noise dominance within each T-F unit, and yields remarkable speech intelligibility and quality improvements over conventional algorithms in speech enhancement [2], [3], [24], [25], and word error rates (WER) improvements in robust ASR [26], [27]. The estimated mask, $\widehat{M}_i(t,f)$, is used to derive spatial covariance matrices for beamforming as in Eq. (2). Even if the BLSTM only uses energy based features, it is still powerful enough to identify T-F units where the phase is much less contaminated.

3. EXPERIMENTAL SETUP

We evaluate our algorithms on the six-channel task of the recently-proposed CHiME-4 dataset [4]. The six microphones are mounted on a tablet, with the second one on the rear and the other five facing front. It contains simulated utterances, and real recordings from four real-world environments (street, pedestrian areas, cafeteria and bus), and exhibits strong mismatches between training and testing conditions. The training data includes 7,138 simulated and 1,600 real utterances, the validation data consists of 1,640 simulated and 1,640 real utterances, and the test data consists of 1,320 simulated and 1,320 real utterances.

Our acoustic model is a feed-forward DNN with seven hidden layers, each with 2,048 exponential linear units. The input feature is 40-dimensional log Mel filterbank feature together with its deltas and double deltas, and an 11-frame symmetric context window. The input dimension is therefore 1,320. Sentence-level mean-variance normalization is performed on the input features before global mean-variance normalization. The dropout rate is set to 0.3. Batch normalization and AdaGrad are used to speed up training. Our acoustic model is trained on all the unprocessed simulated and real training data, except the utterances from the second channel of the real recordings. The total number of utterances for training is therefore 7,138*6+1,600*5 (~104h). The senone labels are generated from the GMM-HMM system provided in the challenge. Note that the beamformed signal is directly fed into the acoustic model for decoding. To facilitate the comparison with other systems, the task-standard five-gram and RNN language model are employed here for lattice re-scoring. We use our recently-proposed unsupervised speaker adaptation algorithm [28] for speaker adaptation.

Multiple BLSTMs taking in different features are trained for mask estimation using the 7,138*6 utterances (~90h) in the simulated training data. The BLSTMs contain three hidden layers, each with 600 hidden units in each direction. Sigmoidal units are used in the output layer. The window size is 32ms and the hop size is 8ms.

512-point FFT is performed to extract 257-dimensional log power spectrogram features for BLSTM training. We apply 0.1 dropout rate to the output of each BLSTM layer. Sentence-level mean normalization is performed on the spectral features, while no sentence-level normalization is performed on spatial features. All of the features are then globally normalized to zero mean and unit variance. During training, we use the ideal speech covariance computed directly from clean speech to derive the $\hat{\mathbf{c}}(f)$ in Eq. (12). At the running time, we use the model trained using the log power spectrogram feature together with the MSC feature to get an estimated $\hat{\mathbf{c}}(f)$. When using the spatial features for training, we found that it is very helpful to initialize the corresponding parts of the network using a well-trained model built by only using the log power spectrogram features, likely because spectral information itself is very important for mask estimation.

To address microphone failures, we first select a signal that is most correlated with the rest five signals, and then throw away the signals with less than 0.3 correlation coefficients with the selected signal. The rest signals, except the one from the second channel, are used to derive the diffuse or directional features.

4. EVALUATION RESULTS

The ASR results are presented in Table 1, where we use our DNN based acoustic model after sMBR training for decoding, and the task-standard tri-gram language model for decoding if not specified. The BeamformIt and MVDR via SRP-PHAT are the two official baseline systems provided in the challenge. They are representative baselines of conventional approaches. Their performance on the real test set is however not impressive.

As a comparison, we use the MSC feature, $\mathrm{MSC}(t,f)$, as the $\eta(t,f)$ in Eq. (2) and $1-\mathrm{MSC}(t,f)$ as the $\xi(t,f)$ in Eq. (3) to construct an MVDR beamformer using Eq. (6) for enhancement. Note that the range of $\mathrm{MSC}(t,f)$ has been linearly mapped to $[0,1]$ within each utterance. Surprisingly, this simple approach, which does not even require any training or spatial clustering, achieves 9.91% WER on the real test set. This is probably because the real noises recorded in the CHiME-3/4 dataset are mostly diffuse noises. This makes sense as in practice the acoustic scene in a bus, cafeteria, pedestrian area, and on the street would contain noises or interferences from many directions, such as engine noises, background speakers, wind noises or room reverberations. Even if there are directional sources present, they are commonly much weaker than the target speaker when the SNR is not very low[1]. In such cases, the speech covariance matrix computed via weighted pooling in Eq. (2) would still be dominated by target speech.

Using the log power spectrogram feature to train a BLSTM to predict the IRM, we get to 7.28% WER. Adding the MSC features for BLSTM training pushes the performance to 6.92% WER. For the model trained with the log power spectrogram and directional features, we first use the model trained with the log power spectrogram and MSC features to get an estimated $\hat{\mathbf{c}}(f)$ and then use it to compute the directional features using Eq. (12). The result is further pushed to 6.70% WER. The directional features yield better results over the MSC features. This is reasonable as noises or inferences could also be directional. Note that after adding the spatial features, the performance on the simulated data however becomes worse, although consistent improvement is observed on the real data. This is likely because of the specific data simulation procedure[2] adopted in the CHiME-4 corpus, which uses the least square algorithm to estimate the speech and noise images from a far-field recording and its corresponding close-talk recording. This procedure could introduce some artifacts in the simulated data, especially on the fragile phase information that are important for spatial feature derivation.

Table 1. Comparison of the ASR performance (%WER) with other approaches.

Approaches	Dev. set		Test set	
	SIMU	REAL	SIMU	REAL
BeamformIt! [32], [33]	8.62	7.28	12.81	11.72
MVDR via SRP-PHAT [4]	6.32	9.38	7.05	14.60
MSC as the Estimated Mask (no training)	6.49	6.16	9.77	9.91
Log Power Spectrogram	5.67	5.16	6.09	7.28
Log Power Spectrogram + MSC	5.63	5.08	6.31	6.92
Log Power Spectrogram + DF	5.82	5.06	6.49	6.70
+Five-gram LM and RNNLM	3.90	3.11	4.33	4.54
+Unsupervised speaker adaptation [28]	2.83	**2.54**	3.11	**3.08**
Du *et al*. [29] (with model ensemble)	2.61	2.55	3.06	3.24
Best single model of [29]	-	2.88	-	3.87
Heymann *et al*. [31]	2.75	2.84	3.11	3.85

Using the task-standard five-gram and RNNLM language model for lattice re-scoring, the result is pushed to 4.54% WER. Note that the system so far is fully speaker independent. Further applying the unsupervised speaker adaptation algorithm in our recent study [28] improves the performance to 3.08% WER. This result is slightly better than the 3.24% WER obtained in the winning solution of the CHiME-4 challenge by Du *et al*. [29]. Their acoustic model is a combination of one DNN-based acoustic model and four CNN-based acoustic models trained from augmented training data. The input feature is a combination of log Mel filterbank features, fMLLR features and i-vectors. Their T-F masking based MVDR beamformer is constructed using a complex GMM based spatial clustering algorithm [19], a DNN based IRM estimator, the silence frames determined by the backend ASR systems, and an iterative mask refinement strategy [30]. The runner-up system by Heymann *et al*. [31] uses a BLSTM to drive a T-F masking based generalized eigen-vector beamformer [6], and a complicated wide-residual BLSTM for acoustic modeling. Input-level linear transform is performed on each testing speaker for unsupervised speaker adaptation. Their best performance when using the task-standard RNNLM is 3.87% WER. Different from these competitive systems, our approach is focused on frontend beamforming. Even with a simple feed-forward DNN as the backend acoustic model, our system has shown better performance. This justifies the benefits of the proposed beamforming algorithm.

5. CONCLUDING REMARKS

This study has proposed a novel approach to integrate spectral and spatial features to improve time-frequency masking based beamforming. Consistent improvement has been observed on the six-channel task of the CHiME-4 challenge. Although the computation of the directional features requires a separate localization-like procedure, our results indicate that directional and diffuse features likely contain discriminative information for supervised mask estimation. Hence combining them with spectral features for DNN training would lead to better mask estimates. Future research would use deep learning based post-filtering to achieve further noise reduction. Replacing the DNN based acoustic model with an RNN based acoustic model may also yield better ASR results.

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]Users tend to not use speech recognizers in very noisy environments.
[2]See http://spandh.dcs.shef.ac.uk/chime_challenge/chime2015/data.html for more details.

7. REFERENCES

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

[2] 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.

[3] E. Healy, S. Yoho, Y. Wang, and D. L. Wang, "An Algorithm to Improve Speech Recognition in Noise for Hearing-Impaired Listeners," *The Journal of the Acoustical Society of America*, vol. 23, no. 6, pp. 3029–3038, 2013.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, "On Time-Frequency Mask Estimation for MVDR Beamforming with Application in Robust Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 3246–3250.

[10] Y. Jiang, D. L. Wang, R. Liu, and Z. Feng, "Binaural Classification for Reverberant Speech Segregation using Deep Neural Networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 22, no. 12, pp. 2112–2121, 2014.

[11] X. Zhang and D. L. Wang, "Deep Learning Based Binaural Speech Separation in Reverberant Environments," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 5, pp. 1075–1084, 2017.

[12] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring Multi-Channel Features for Denoising-Autoencoder-Based Speech Enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 116–120.

[13] I. Tashev and A. Acero, "Microphone Array Post-Processor using Instantaneous Direction of Arrival," in *Proceedings of IWAENC*, 2006.

[14] P. Pertil and Joonas Nikunen, "Microphone Array Post-Filtering using Supervised Machine Learning for Speech Enhancement," in *Proceedings of Interspeech*, 2014, pp. 2675–2679.

[15] A. Brutti, A. Tsiami, A. Katsamanis, and P. Maragos, "A Phase-Based Time-Frequency Masking for Multi-Channel Speech Enhancement in Domestic Environments," in *Proceedings of Interspeech*, 2016, pp. 2875–2879.

[16] H. Barfuss, C. Huemmer, A. Schwarz, and W. Kellermann, "Robust Coherence-Based Spectral Enhancement for Distant Speech Recognition," in *arXiv preprint arXiv:1509.06882*, 2015.

[17] Z. Pang and F. Zhu, "Noise-Robust ASR for the Third 'CHiME' Challenge Exploiting Time-Frequency Masking based Multi-Channel Speech Enhancement and Recurrent Neural Network," in *arXiv preprint arXiv:1509.07211*, 2015.

[18] H. Barfuss, C. Huemmer, A. Schwarz, and W. Kellermann, "Robust Coherence-Based Spectral Enhancement for Speech Recognition in Adverse Real-World Environments," *Computer Speech and Language*, pp. 388–400, Apr. 2017.

[19] 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.

[20] Z.-Q. Wang and D. Wang, "Mask Weighted STFT Ratios for Relative Transfer Function Estimation and its Application to Robust ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[21] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multi-Microphone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 692–730, 2017.

[22] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[23] P. Pertila and E. Cakir, "Robust Direction Estimation with Convolutional Neural Networks based Steered Response Power," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 6125–6129.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] Y. Tu, J. Du, L. Sun, F. Ma, and C. Lee, "On Design of Robust Deep Models for CHiME-4 Multi-Channel Speech Recognition with Multiple Configurations of Array Microphones," in *Proceedings of Interspeech*, 2017, pp. 394–398.

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

[32] 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.

[33] 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.