Interspeech 2018
2-6 September 2018, Hyderabad

Integrating Spectral and Spatial Features for Multi-Channel Speaker Separation

Zhong-Qiu Wang[1], *DeLiang Wang*[1, 2]

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

Abstract

This paper tightly integrates spectral and spatial information for deep learning based multi-channel speaker separation. The key idea is to localize individual speakers so that an enhancement network can be used to separate the speaker from an estimated direction and with specific spectral characteristics. To determine the direction of the speaker of interest, we identify time-frequency (T-F) units dominated by that speaker and only use them for direction of arrival (DOA) estimation. The speaker dominance at each T-F unit is determined by a two-channel permutation invariant training network, which combines spectral and interchannel phase patterns at the input feature level. In addition, beamforming is tightly integrated in the proposed system by exploiting the magnitudes and phase produced by T-F masking based beamforming. Strong separation performance has been observed on a spatialized reverberant version of the wsj0-2mix corpus.

Index Terms: spatial features, permutation invariant training, deep neural networks, cocktail party problem.

1. Introduction

Riding on the tide of deep learning, monaural speaker-independent speaker separation, or the cocktail party problem, has made major advances since the introduction of deep clustering [1], [2], [3], [4], deep attractor networks [5] and permutation invariant training (PIT) [6]. These algorithms address the label permutation problem in the challenging monaural speaker-independent setup, demonstrating much better separation performance over conventional algorithms such as spectral clustering [7], computational auditory scene analysis [8], and speaker- or target-dependent systems [9].

When multiple microphones are available, spatial information can be leveraged for better separation, as speaker sources are directional and usually spatially separated in real-world environments. One stream of research to exploit this information is focused on spatial clustering [10], [11], [12], [13], [14], which clusters individual T-F units according to their spatial origins under the speech sparsity assumption [8], [15], using spatial cues such as interchannel time, phase or level differences (ITDs/IPDs/ILDs) and directional statistics. However, these approaches typically only consider spatial information, which is insufficient for separation in reverberant environments or when sound sources are close to one another. In contrast, recent developments in deep learning based monaural speech separation have shown that even with spectral information alone, remarkable separation performance can be achieved [16], [17].

One promising research direction is thus to combine the merits of these two streams of research so that spectral and spatial processing can be tightly integrated to improve separation. In [18] and [19], estimated masks or embeddings from monaural deep clustering are utilized to construct a beamformer in each frequency for separation. Their studies follow the recent development of T-F masking based beamforming in the CHiME challenges [20]. The performance is however largely limited by beamforming, which cannot produce sufficient separation when room reverberation is strong and when the speakers are close to one another. In such cases, performing further spectral masking would be very helpful. A recent study [21] applies monaural deep attractor networks on the outputs of a number of fixed beamformers. However, their approach requires the knowledge of microphone geometry to manually design the fixed beamformers for a single fixed device and such fixed beamformers are typically not as powerful as data-dependent beamformers, which can lead to significant noise reduction based on signal statistics. Different from the above approaches, which apply deep clustering or its variants only on monaural spectral features, our recent study [22] proposes a multi-channel deep clustering algorithm, which utilizes IPDs as additional features for DNN training. Although IPDs are inherently ambiguous across frequencies, experimental results suggest that spectral features can help to resolve this ambiguity. However, this approach does not exploit beamforming, which can produce phase enhancement and is known to perform very well in less reverberant conditions or when the number of microphones is large.

Following [22], this study utilizes IPD features as additional inputs for PIT, as the PIT approach is more end-to-end than deep clustering. The PIT network is used to resolve the permutation problem as well as for DOA estimation. With the permutation issue resolved and target direction estimated, the problem becomes how to separate the speaker of interest with specific spectral characteristics and arriving from a particular direction. We address this problem by using an enhancement network, where the input is a combination of spectral features, initial mask estimates from the PIT network, and directional features indicating whether the signal is from the estimated direction. Spectral and spatial information are hence tightly integrated at the input level to leverage the representational power of deep learning for better mask and magnitude estimation. In addition, the phase estimate produced by data-dependent beamforming is utilized as the enhanced phase.

Previous studies have utilized spatial features for DNN training [23], [24], [25]. However, they are designed for speech enhancement tasks (*i.e.* speech vs. noise) and assume that the target speech is in the front direction in the binaural setup. In more general cases, the target speaker may originate in any direction and the spatial features proposed in those studies would no longer work well. We evaluate the proposed algorithms on a spatialized reverberant version of the wsj0-2mix corpus. Much better separation results have been observed over the oracle multi-channel Wiener filter, MESSL [12], GCC-NMF [26] and multi-channel deep clustering [3].

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

10.21437/Interspeech.2018-1940

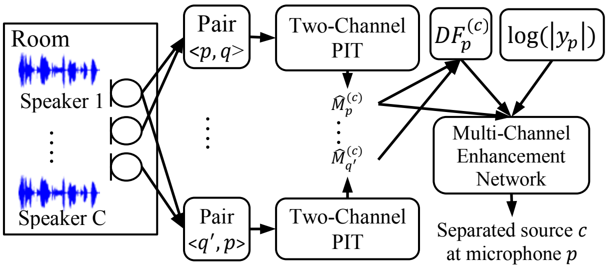

Figure 1. Illustration of overall system.

2. System Description

Our system contains two neural networks, with a two-channel PIT network (depicted in Figure 2) trained for initial mask and DOA estimation, and a multi-channel enhancement network trained to refine the initial mask estimate of each speaker. The overall system is depicted in Figure 1.

2.1. Single-Channel Permutation Invariant Training

A permutation invariant objective function was first proposed in [1] while later reported in [2], [27] to work as comparably well as deep clustering. The key idea is to train a neural network to minimize the minimum utterance-level loss of all the permutations. The phase-sensitive mask (PSM) [28] is typically used as the training target. The loss function to minimize is:

$$\mathcal{L}_{PIT} = \min_{\pi\epsilon\Psi} \sum\nolimits_c \left\| \hat{M}_p^{\pi(c)} |y_p| - T_0^{|y_p|} \left(|s_p^{(c)}| \cos(\angle s_p^{(c)} - \angle y_p) \right) \right\|_1, \tag{1}$$

where p indexes microphone channels, $\boldsymbol{s}^{(c)}$ and $\boldsymbol{y} = \sum_c \boldsymbol{s}^{(c)}$ are the STFT representation of source c and the mixture, Ψ is a set of permutations on C sources, $T_0^{|y_p|}(\cdot) = \max(0, \min(|y_p|, \cdot))$ truncates the PSM to the range [0,1], $\hat{M}$ denotes the estimated masks, $|\cdot|$ computes magnitude, and $\angle(\cdot)$ extracts phase. Following [29] and [3], the L_1 loss is used for training and sigmoidal units are used in the output layer.

A recurrent neural network with bi-directional long short-term memory (BLSTM) cells is commonly used in PIT. Log magnitude is used as the input feature. The network architecture is shown in Figure 2.

2.2. Two-Channel Permutation Invariant Training

Our recent study [22] found that simply including the cosine and sine of IPDs for DNN training leads to significant improvements. In this study, we include these two features for PIT. As PIT network is more capable of end-to-end optimization [3], the estimated mask is expected to be better than the binary mask produced by multi-channel deep clustering [22].

Given a microphone pair $\langle p, q \rangle$, we extract spectral features $\log(|y_p|)$ from microphone p, cosIPD $\cos(\angle y_p - \angle y_q)$, and sinIPD $\sin(\angle y_p - \angle y_q)$ to train our PIT network, where the labels are computed using the source images captured at microphone p. The network architecture is illustrated in Figure 2. At run time, the separation results are obtained as $\hat{s}_p^{(c)} = \hat{M}_p^{(c)} y_p$, where $\hat{M}_p^{(c)}$ denotes the estimated mask of source c at microphone p. The rationale [22] of using cosIPD and sinIPD is that $y_p/y_q = |y_p|/|y_q| e^{j(\angle y_p - \angle y_q)}$ should naturally form clusters within each frequency for spatially separated speakers with difference time delays. As $|y_p|$ and $|y_q|$ are very similar in far-field conditions, we only use the real and imaginary parts of $e^{j(\angle y_p - \angle y_q)}$ as the additional features. Although cosIPD and sinIPD are ambiguous across frequencies, the spectral features could help to resolve this ambiguity [22].

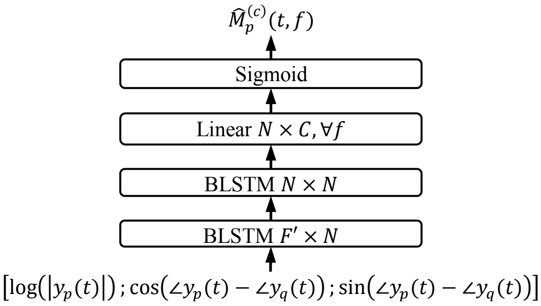
$\hat{M}_p^{(c)}(t,f)$
Sigmoid
Linear $N \times C, \forall f$
BLSTM $N \times N$
BLSTM $F' \times N$
$[\log(|y_p(t)|); \cos(\angle y_p(t) - \angle y_q(t)); \sin(\angle y_p(t) - \angle y_q(t))]$

Figure 2. Illustration of two-channel PIT.

2.3. Multi-Channel Speech Enhancement

The enhancement network takes in spectral features, initial mask estimates by the two-channel PIT network, and directional features to improve the mask estimation of target speakers. We introduce two types of directional features, one based on compensating IPDs and the other based on beamforming.

Suppose that there are $D(\geq 2)$ microphones and microphone p is designated as the reference microphone, we consider D microphone pairs: one pair $\langle p, q \rangle$, where q is a randomly-chosen non-reference microphone, and $D-1$ pairs $\langle q', p \rangle$ for any non-reference microphone q'. We apply the two-channel PIT network on each of the D pairs to obtain an estimated mask of each source at each microphone. With these estimated masks, we first compute the speech covariance matrix for source c following [30], [31], [32], [33]:

$$\hat{\Phi}^{(c)}(f) = \frac{1}{T} \sum\nolimits_t \eta^{(c)}(t,f) \boldsymbol{y}(t,f)\, \boldsymbol{y}(t,f)^H \tag{2}$$

where $(\cdot)^H$ stands for conjugate transposition and $\eta^{(c)}(t,f)$ is the weight denoting the importance of each T-F unit for the computation of $\hat{\Phi}^{(c)}(f)$. It is computed using the median of the estimated masks, following [31].

$$\eta^{(c)}(t,f) = \text{median}(\hat{M}_1^{(c)}(t,f), \ldots, \hat{M}_D^{(c)}(t,f)) \tag{3}$$

The steering vector $\hat{\boldsymbol{r}}^{(c)}(f)$ is then computed as the principal eigenvector of $\hat{\Phi}^{(c)}(f)$ [30]. The rationale is that if $\hat{\Phi}^{(c)}(f)$ is well estimated, it would be close to a rank-one matrix, as the target speaker is a directional source [30], [15]. The estimation of steering vectors is essentially similar to DOA estimation.

Following our recent study [34], one way to compute directional features (DF) is to compensate the IPD features using the estimated phase difference:

$$DF_p^{(c)}(t,f) = \frac{1}{D-1} \sum_{<q',p>\epsilon\Omega} \cos\left\{ \angle y_{q'}(t,f) - \angle y_p(t,f) - \left(\angle \hat{r}_{q'}^{(c)}(f) - \angle \hat{r}_p^{(c)}(f) \right) \right\} \tag{4}$$

where Ω contains all the considered $D-1$ microphone pairs. The key idea is that for a T-F unit dominated by source c, the observed IPD should be aligned with the estimated phase difference, only if the steering vector is well estimated. The phase compensation term is used to establish the consistency of directional features along frequency such that at any frequency, a value close to one in the derived directional feature would indicate that the T-F unit is likely dominated by the target source, while dominated by other sources otherwise. This property makes the directional features useful for DNN based T-F masking to enhance the signal from a specific direction.

An alternative is to use beamforming results as directional features. We here consider the multi-channel Wiener filter (MCWF) computed as:

$$\hat{\boldsymbol{w}}_p^{(c)}(f) = \left(\hat{\Phi}^{(y)}(f) \right)^{-1} \hat{\Phi}^{(c)}(f) \boldsymbol{u}, \tag{5}$$

where $\hat{\Phi}^{(y)}(f) = \frac{1}{T}\sum_t \boldsymbol{y}(t,f)\, \boldsymbol{y}(t,f)^H$ is the observed mixture covariance matrix and $\boldsymbol{u}$ a one-hot vector with u_p being one. The feature is then computed as:

$$DF_p^{(c)}(t,f) = \log\left(\left|\hat{\boldsymbol{w}}_p^{(c)}(f)^H \boldsymbol{y}(t,f)\right|\right) \tag{6}$$

Clearly, using the directional features alone is not sufficient enough for separation, as the underlying sources could be spatially close to each other and reverberation components of interfering sources could arrive the array from the estimated direction. We hence also utilize the spectral feature of the mixture and the initial mask estimate as the inputs to train the enhancement network. This way, only the signal with specific spectral characteristics and in a particular direction is enhanced while suppressed otherwise. More specifically, the enhancement network is trained using a combination of $\log(|y_p|)$, $\hat{M}_p^{(c)}$ and $DF_p^{(c)}$ to estimate the PSM of source c at microphone p. The loss function is:

$$\mathcal{L}_{Enh} = \left\| \hat{R}_p^{(c)}|y_p| - T_0^{|y_p|}\left(|s_p^{(c)}|\cos\left(\angle s_p^{(c)} - \hat{\theta}_p^{(c)}\right)\right)\right\|_1, \tag{7}$$

where $\hat{R}_p^{(c)}$ is the estimated mask produced by the enhancement network and $\hat{\theta}_p^{(c)}(t,f) = \angle(\hat{\boldsymbol{w}}_p^{(c)}(f)^H \boldsymbol{y}(t,f))$ is the enhanced phase produced by the MCWF beamformer. At run time, we run the enhancement network once for each source and the separation result is obtained as $\hat{s}_p^{(c)} = \hat{R}_p^{(c)}|y_p|e^{j\hat{\theta}_p^{(c)}}$. We emphasize that the phase produced by beamforming is employed as the enhanced phase, since $\hat{\theta}_p^{(c)}$ is expected to be better than the mixture phase $\angle y_p$ if the distortion introduced by beamforming is minimal. We also point out that we do not train a network to estimate a mask that will be, at run time, applied to $\hat{\boldsymbol{w}}_p^{(c)}(f)^H \boldsymbol{y}(t,f)$. Instead, the magnitude produced by beamforming is used as directional features to improve the magnitude estimation of source c at microphone p, as $s_p^{(c)}$, rather than fluid $\hat{\boldsymbol{w}}_p^{(c)}(f)^H \boldsymbol{s}^{(c)}(t,f)$, is considered as the reference for metric computation.

Our models, once trained, can be directly applied to microphone arrays with various numbers of microphones arranged in diverse geometry. We can first apply the well-trained two-channel PIT network on each microphone pair, then use Eq. (4) or (6) to constructively combine all the microphones, and finally apply the well-trained enhancement network on the derived features for each source for separation. Note that we can replace two-channel PIT with single-channel PIT in this pipeline. However, we found that the former produces better initial mask estimates, simply because the permutation problem can be alleviated by exploiting IPDs. The two-channel PIT network, however, can only utilize pairwise spatial information and its extension to multi-channel cases is not straightforward. In contrast, the directional features constructively combine the spatial information of all the microphones and hence are expected to better encode spatial information than the ambiguous pairwise cosIPDs and sinIPDs.

3. Experimental Setup

We train our models using simulated room impulse responses (RIR) and test on simulated as well as real-recorded RIRs. To create reverberant multi-channel speaker mixtures, we convolve the RIRs with the utterances in the open wsj0-2mix data [1], which contains 20,000, 5,000 and 3,000 single-channel anechoic two-speaker mixtures in its 30-hour training, 10-hour validation and 5-hour test set. The speakers in the validation set are seen during training, while the test speakers are unseen. The task is hence speaker-independent. In wsj0-2mix, the SNR of one source with respect to the other is uniformly drawn from -5 dB to 5dB. The sampling rate is 8 kHz.

```
**Input**: wsj0-2mix;
**Output**: spatialized reverberant wsj0-2mix;
**For** each source *s1*, source *s2* in wsj0-2mix **do**
    Sample room length $r_x$ and width $r_y$ from [5,10] m;
    Sample room height $r_z$ from [3,4] m;
    Sample mic array height $a_z$ from [1,2] m;
    Sample displacement $n_x$ and $n_y$ of mic array from $[-0.2,0.2]$ m;
    Place array center at $\left[\frac{r_x}{2}+n_x, \frac{r_y}{2}+n_y, a_z\right]$ m;
    Sample microphone spacing $a_r$ from [0.02,0.09] m;
    **For** $p = 1{:}D(=8)$ **do**
        Place mic $p$ at $\left[\frac{r_x}{2}+n_x-\frac{D-1}{2}a_r+(p-1)a_r, \frac{r_y}{2}+n_y, a_z\right]$ m;
    **End**
    Sample speaker locations in the frontal plane:
        $s_x^{(1)}, s_y^{(1)}, s_z^{(1)} = a_z; s_x^{(2)}, s_y^{(2)}, s_z^{(2)} = a_z$; such that any two
        speakers are at least 15° apart from each other with respect
        to the array center, and the distance from each speaker
        to the array center is in between [0.75,2] m;
    Sample T60 from [0.2,0.7] s;
    Generate impulse responses using RIR generator and convolve them with
    *s1* and *s2*;
    Concatenate channels of reverberated *s1* and *s2*, scale them to match SNR
    between original *s1* and *s2*, and add them to obtain reverberated mix;
**End**
```

Algorithm 1. Data spatialization process (simulated RIRs).

We employ the RIR generator[1], which is based on the classic image method, to generate simulated RIRs. The spatialization process is detailed in Algorithm 1. An illustration of the setup is depicted in Figure 3(a). The overall guideline is to make the setup as random as possible while still subject to realistic constraints. For each mixture in wsj0-2mix, we randomly generate a room with random room characteristics, speaker locations and array spacing. Here, we consider a linear array setup with speakers randomly located in the front plane. We generated 20,000, 5,000 and 3,000 eight-channel utterances for training, validation and testing, respectively. The average speaker-to-microphone distance is 1.38 m with 0.37 m standard deviation and the average direct-to-reverberant energy ratio (DRR) is 0.49 dB with 3.92 dB standard deviation.

We generated another 3,000 eight-channel utterances for testing using the Multi-Channel Impulse Responses Database[2] [35] recorded using eight-microphone linear arrays with three different microphone spacing (i.e. 3-3-3-8-3-3-3, 4-4-4-8-4-4-4 and 8-8-8-8-8-8-8 cm). The RIRs are measured in a room of size 6x6x2.4 m in steps of 15° from $-90°$ to 90°, at a distance of 1 m and 2 m, and at three T60s (0.16, 0.36 and 0.61 s). We randomly place each speaker in each test utterance of wsj0-2mix at a randomly-chosen direction and distance, using a randomly-chosen linear array and a randomly-chosen reverberation time. Note that for any two speakers, we constrain them to be at least 15° apart. See Figure 3(b) for an illustration. The average DRR is 2.8 dB with 3.8 dB standard deriva-tion. We emphasize that this setup is a very realistic one, as it is speaker-independent and we only use simulated RIRs for training, while real RIRs for testing.

The PIT and enhancement networks respectively contain four and three BLSTM layers, each with 600 units in each direction. Both networks are trained on 400-frame segments of the 20,000 eight-channel utterances generated using the simulated RIRs. Adam is utilized for optimization. The window size is 32 ms and the hop size is 8 ms. We apply 256-point FFT to extract 129-dimensional log magnitudes and spatial features for model training. We emphasize that the enhance-

[1] Available at https://github.com/ehabets/RIR-Generator.

[2] Available at http://www.eng.biu.ac.il/~gannot/RIR_DATABASE/.

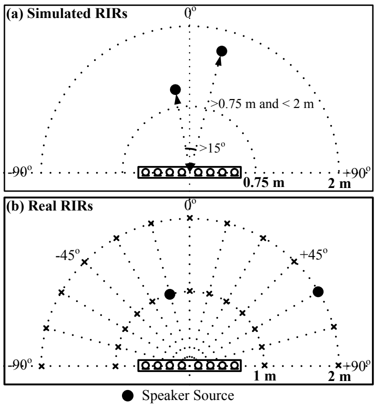

Figure 3. Illustration of experimental setup.

Table 1. SDRi (dB) results on spatialized wsj0-2mix (simulated RIRs).

Approaches	$DF^{(c)}(t,f)$	Simu RIRs	Real RIRs
1ch PIT	-	7.5	7.3
2ch PIT	-	9.9	9.1
+ enhancement network	Eq. (4)	10.6	10.3
+ enhancement network	Eq. (6)	10.9	10.9

Table 2. SDRi (dB) comparison with other approaches using various numbers of microphones on spatialized wsj0-2mix (real RIRs).

#mics	MESSL [12]	GCC-NMF [26]	eMCWF	MCDC [22]	Proposed		tPSM-MCWF	Oracle Masks			
					$DF_p^{(c)}(t,f)$			IRM	IBM	tPSM	MC-tPSM
					Eq.(4)	Eq.(6)					
2	4.1	5.0	6.5	9.2	10.3	10.9	7.1	12.1	13.0	14.1	14.1
3	-	-	7.8	9.6	10.8	11.7	8.6				14.8
4	-	-	8.7	9.8	11.1	12.3	9.6				15.3
5	-	-	9.4	9.9	11.4	12.8	10.4				15.8
6	-	-	9.8	10.0	11.6	13.2	11.0				16.2
7	-	-	10.2	10.0	11.8	13.4	11.5				16.5
8	-	-	10.5	10.0	11.9	13.6	11.9				16.7

ment network needs to be trained on the directional feature computed from various numbers of microphones, as its quality varies with the numbers of microphones. At run time, for each utterance, we randomly select a subset of microphones for testing. The aperture size can be 3 cm at minimum and 56 cm at maximum for the real RIRs, and 2 cm and 63 cm for the simulated RIRs. SDR improvement (SDRi) computed using the *bss_eval_images* software is used as the evaluation metric. The reverberant image of each source at the reference microphone, i.e. $s_p^{(c)}$, is used as the reference for metric computation.

4. Evaluation Results

The second last column of Table 1 presents the results on the simulated RIRs. The performance of single-channel PIT in the reverberant condition (7.5 dB) is much lower than the 10.0 dB SDRi [6], [3] obtained on the original anechoic wsj0-2mix corpus, likely because of the smearing of spectral features and the breaking of the speech sparsity property due to reverberation. Although cosIPDs and sinIPDs are ambiguous across frequencies and the microphone geometry is completely unknown, adding them as additional features for model training leads to large improvement (from 7.5 to 9.9 dB). Then, we use the enhancement network to further improve the performance, based on the masks estimated from the two-channel PIT network. This improves the performance from 9.9 to 10.6 and 10.9 dB for the directional features computed using Eq. (4) and Eq. (6), respectively. The last column of Table 1 reports the performance on the test data spatialized by the real RIRs. The results hold up reasonably well, although the models are trained only on the simulated RIRs. In addition, similar trends as in the second last column are observed.

Table 2 shows the results with up to eight microphones along with the comparison with other systems. Even with random microphone geometry, adding more microphones gradually improves the separation performance (from 10.9 dB for two microphones to 13.6 dB for eight microphones). Recent studies [18], [19] apply single-channel deep clustering on each microphone signal to derive a T-F masking based beamformer for each source for separation. To compare with their approaches, we supply the truncated PSM (tPSM), computed as $T_0^{1.0}(|s_p^{(c)}|\cos(\angle s_p^{(c)} - \angle y_p)/|y_p|)$, to Eq. (3) to compute oracle $\hat{\Phi}^{(c)}(f)$ and report oracle MCWF results (denoted as tPSM-MCWF). We also report the performance of estimated MCWF (eMCWF) obtained using the estimated masks $\hat{M}_p^{(c)}$ computed from the two-channel PIT network. In addition, we compare our algorithm with MESSL[3] [12], a popular Gaussian mixture model based wideband spatial clustering algorithm proposed for two-microphone array, and GCC-NMF[4] [26], where dictionary atoms obtained from non-negative matrix factorization are assigned to individual sources over time according to their time difference of arrival estimates obtained from GCC-PHAT. The recently-proposed multi-channel deep clustering (MCDC) [22] algorithm combines deep clusteirng with conventional spatial clustering. Its extension to multi-channel cases is done by first applying two-channel deep clustering on each microphone pair, then stacking the embeddings from each pair, and finally performing kmeans clustering on the stacked embeddings. Our approach is consitently better than MCDC, likely because our approach is more end-to-end and better integrates spatial information. We also list the performance of various ideal masks, such as the ideal binary mask (IBM), ideal ratio mask (IRM) and tPSM, which are computed based on the source images captured at the reference microphone. Compared with such monaural ideal masks that use mixture phase for re-synthesis, the multi-channel tPSM (MC-tPSM), computed as $T_0^{1.0}(|s_p^{(c)}|\cos(\angle s_p^{(c)} - \hat{\theta}_p^{(c)})/|y_p|)$ where $\hat{\theta}_p^{(c)}$ here is obtained from tPSM-MCWF and used as the phase for re-synthesis, is clearly better and becomes even better when more microphones are available. Note that MC-tPSM represents the upper bound performance of the proposed approach and shows the effectiveness of using $\hat{\theta}_p^{(c)}$ as the phase estimate.

By exploiting spatial information, we obtain 10.9 dB SDRi using two microphones and 13.6 dB SDRi using eight mcirophones, which are much better than the 7.3 dB SDRi obtained by single-channel PIT. The 13.6 dB result is better than the 12.1 and 13.0 dB results obtained using the monarual IBM and IRM (with mixture phase).

5. Concluding Remarks

We have proposed a novel and effective deep learning based approach for BSS, where complementary spectral and spatial information are integrated as input features to improve mask estimation. The trained models are flexible enough to be applied to arrays with various numbers of microphones arranged in diverse geometry. Future research would include the joint training of the PIT and the enhancement network, exploring other types of spatial features and tighter integration with beamforming algorithms.

[3] Available at https://github.com/mim/messl.

[4] Available at https://github.com/seanwood/GCC-nmf.

6. References

[1] J. R. Hershey, Z. Chen, J. L e Roux, and S. Watanabe, “Deep Clustering: Discriminative Embeddings for Segmentation and Separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 31–35.

[2] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-Channel Multi-Speaker Separation using Deep Clustering,” in *Proceedings of Interspeech*, 2016, pp. 545–549.

[3] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, “Alternative Objective Functions for Deep Clustering,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[4] Z.-Q. Wang, J. Le Roux, D. Wang, and J. R. Hershey, “End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction,” in *arXiv preprint arXiv:1804.10204*, 2018.

[5] Z. Chen, Y. Luo, and N. Mesgarani, “Speaker-Independent Speech Separation with Deep Attractor Network,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Jul. 2018.

[6] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multi-Talker Speech Separation with Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, pp. 1901–1913, Mar. 2017.

[7] F. Bach and M. Jordan, “Learning Spectral Clustering, with Application to Speech Separation,” *The Journal of Machine Learning Research*, vol. 7, pp. 1963–2001, 2006.

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

[9] X.-L. Zhang and D. L. Wang, “A Deep Ensemble Learning Method for Monaural Speech Separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 967–977, 2016.

[10] N. Q. K. Duong, E. Vincent, and R. Gribonval, “Under-Determined Reverberant Audio Source Separation using A Full-Rank Spatial Covariance Model,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[11] H. Sawada, S. Araki, and S. Makino, “Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[12] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, “Model-Based Expectation-Maximization Source Separation and Localization,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[13] M. I. Mandel and J. P. Barker, “Multichannel Spatial Clustering using Model-Based Source Separation,” in *New Era for Robust Speech Recognition Exploiting Deep Learning*, 2017, pp. 51–78.

[14] N. Ito, S. Araki, and T. Nakatani, “Recent Advances in Multichannel Source Separation and Denoising Based on Source Sparseness,” in *Audio Source Separation*, 2018, pp. 279–300.

[15] 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.

[16] Y.-M. Qain, C. Weng, X. Chang, S. Wang, and D. Yu, “Past Review, Current Progress, and Challenges Ahead on the Cocktail Party Problem,” *Frontiers of Information Technology & Electronic Engineering*, pp. 40–63, 2018.

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

[18] L. Drude; and R. Haeb-Umbach, “Tight Integration of Spatial and Spectral Features for BSS with Deep Clustering Embeddings,” in *Proceedings of Interspeech*, 2017, pp. 2650–2654.

[19] T. Higuchi, K. Kinoshita, M. Delcroix, K. Zmolkova, and T. Nakatani, “Deep Clustering-Based Beamforming for Separation with Unknown Number of Sources,” in *Proceedings of Interspeech*, 2017, pp. 1183–1187.

[20] 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.

[21] Z. Chen, J. Li, X. Xiao, T. Yoshioka, H. Wang, Z. Wang, and Y. Gong, “Cracking the Cocktail Party Problem by Multi-Beam Deep Attractor Network,” in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2017.

[22] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, “Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent Speech Separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[23] 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.

[24] 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.

[25] 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.

[26] S. U. N. Wood, J. Rouat, S. Dupont, and G. Pironkov, “Blind Speech Separation and Enhancement with GCC-NMF,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 4, pp. 745–755, 2017.

[27] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 241–245.

[28] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, “Phase-Sensitive and Recognition-Boosted Speech Separation using Deep Recurrent Neural Networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 708–712.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] E. Hadad, F. Heese, P. Vary, and S. Gannot, “Multichannel Audio Database in Various Acoustic Environments,” in *International Workshop on Acoustic Signal Enhancement*, 2014, pp. 313–317.