INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Improved Speech Separation with Time-and-Frequency Cross-domain Joint Embedding and Clustering

Gene-Ping Yang[1]*, Chao-I Tuan*[2]*, Hung-yi Lee*[3]*, Lin-shan Lee*[4]

Graduate Institute of Networking and Multimedia, National Taiwan University

r06944010@ntu.edu.tw[1], chaoi111.t@gmail.com[2], tlkagkb93901106@gmail.com[3], lslee@gate.sinica.edu.tw[4]

Abstract

Speech separation has been very successful with deep learning techniques. Substantial effort has been reported based on approaches over magnitude spectrogram, which is well known as the standard time-and-frequency cross-domain representation for speech signals. It is highly correlated to the phonetic structure of speech, or "how the speech sounds" when perceived by human, but primarily frequency domain features carrying temporal behaviour. Very impressive work achieving speech separation over time domain was reported recently, probably because waveforms in time domain may describe the different realizations of speech in a more precise way than magnitude spectrogram lacking phase information. In this paper, we propose a framework properly integrating the above two directions, hoping to achieve both purposes. We construct a time-and-frequency feature map by concatenating 1-dim convolution encoded feature map (for time domain) and magnitude spectrogram (for frequency domain), which was then processed by an embedding network and clustering approaches very similar to those used in time and frequency domain prior works. In this way, the information in time and frequency domains, as well as the interactions between them, can be jointly considered during embedding and clustering. Very encouraging results (state-of-the-art to our knowledge) were obtained with WSJ0-2mix dataset in preliminary experiments.

Index Terms: Speech separation, Cocktail party problem, deep clustering

1. Introduction

Human beings are able to focus on the voice produced by a single speaker when conversing with a particular individual in a crowded and noisy environment. This requires the ability to extract the desired voice from some mixed audio signal. This so-called cocktail party problem has been shown to be difficult for computers. Very impressive results have been achieved when the speakers are known in advance, but the task remains challenging when the speakers in the mixed voice are not known, referred as the speaker-independent source separation problem. Substantial effort has been made on this problem, obviously because good solutions to it may lead to good contributions to many downstream tasks such as speech recognition [1], speaker identification [2], and audio classification [3] in noisy environment.

Deep learning techniques have accomplished a big step forward on this speech separation task. The Deep Clustering (DPCL) technique [4] was a good such example. It successfully coped with this problem to a good extent by projecting each element in the mixture magnitude spectrogram to a high-dimensional embedding space which is more discriminative for speaker partitioning. Various related deep learning approaches have been proposed and showed great success in enhancement [5, 6, 7] and separation [8, 9, 10, 11, 12] tasks, although many techniques reported at that time consisted of multiple stages separately optimized under different criteria, such as signal representation and embeddings, embedding clustering for speaker assignments [4, 13], mask generation over mixture magnitude spectrogram [8, 14], and phase approximation approaches [15, 16, 17]. End-to-end approaches then became popular later on, in which all different stages with different functions were jointly trained [18, 19].

For most methods performed over the magnitude spectrogram, the ignorance of the phase of the individual sources inevitably distorted the time domain signals. Moreover, predicting masks for each source also caused mismatch for the individual signals. These problems remained even with great effort made. A very impressive phase estimation approach was proposed recently [17] based on a trigonometric perspective over the magnitude spectrogram. This approach achieved great progress over previous methods, offering a signal-to-distortion-ratio improvement $\text{SDR}_i = 15.6$ on the publicly available datasets WSJ0-2mix [4], which seems to be the recent state-of-the-art on the task.

On the other hand, different from the above mentioned methods operated over the spectrograms, a surprisingly successful approach was reported recently called TasNet [20], which directly handled the problem over the time domain signals and achieved superior performance with $\text{SDR}_i = 15.0$ dB on WSJ0-2mix dataset. It contained an encoder module, a separation module and a decoder module, where the separation module consisted of multiple blocks of dilated convolutional layers similar to Wavenet [21], but with fewer parameters and less computation due to the adoption of depthwise separable convolution previously proposed [22].

To the best of our knowledge, existing approaches for the considered problem have taken either time or frequency domain representations as input, both achieving very good and very close performance. Obviously, both representations possess their respective advantages: the robustness of magnitude spectrograms in frequency domain, and the sophisticated but fine structures of time domain signals. The magnitude spectrogram is highly correlated to the phonetic structure of speech, or "how the speech sounds" when perceived by human. But the waveform in time domain describes the various realizations of the sound in a more precise way.

In this paper, we try to integrate both time and frequency domain features together, with the hope to take the advantages of both domains. We construct a time-and-frequency feature map by concatenating features for both time and frequency domains, and perform cross-domain joint embedding and clustering over this feature map, so the model can learn signal behavior in both domains as well as the cross-domain correlations. We make part of the approach similar to TasNet [20], which has fewer parameters with large receptive field due to the dilated convolutional layers. We also adopt the previously proposed clustering method for mask estimation [14], which directly predicts masks for each source in the mixture from feature embeddings. Such a model structure is also in good parallel to the insight offered by a recently reported

http://dx.doi.org/10.21437/Interspeech.2019-2181

work [23]. As will be shown below, very encouraging performance (state-of-the-art to our knowledge) was obtained on WSJ0-2mix dataset in preliminary experiments.

2. Proposed Approach

2.1. Overview of the Proposed Approach

The overview of the proposed approach is in section 2.1, while the details are in sections 2.2-2.5 . The proposed approach consists of three processing modules as shown in Figure 1: an encoder on the left, a separator in the middle, and a decoder on the right.

- The **Encoder** on the left of Figure 1 encodes the input mixture x into a hybrid-domain 2-dim feature map H with $F = F_{conv} + F_{spec}$ channels and T time frames, where F_{conv} and F_{spec} are respectively the dimensionality of the time and frequency domain features, both of which at each time frame correspond to features extracted from a given small segment of the mixture signal.

$$H = Encode(x) \tag{1}$$

- The **Separator** in the middle of Figure 1 consists of two parts, an Embedding Network and Clustering plus Mask Estimation. The Embedding Network projects each element in the feature map H onto a D-dimensional space, forming the embeddings V. Clustering plus Mask Estimation then follows, from which the masks M for the different speakers are generated, and each element of H is assigned to the speakers based on these masks.

$$V = Embed(H) \tag{2}$$

$$M = Clust\text{-}Mask(V) \tag{3}$$

- The **Decoder** decodes the masked feature map into the estimated time domain signals $\hat{s}$, which is the weighted sum of two signal components respectively obtained from the estimated time and frequency domain features, and $\odot$ denotes element-wise multiplication.

$$\hat{s} = Decode(M \odot H) \tag{4}$$

So in this approach we aim to reconstruct time domain signals from both domain features, and directly optimize the signal-to-distortion ratio on the estimated waveform.

2.2. Encoder

In our cross-domain setting, we utilize both time domain signals and the magnitude spectrogram jointly. The input is the mixture signal $x(t)$ produced by N speakers $s_1(t), ..., s_N(t)$, and the corresponding frequency domain representations are obtained by Short-Time Fourier Transform.

$$x(t) = \sum_{i=1}^{N} s_i(t) \tag{5}$$

$$X(f,t) = \sum_{i=1}^{N} S_i(f,t) \tag{6}$$

As shown in the left part of Figure 1, the encoder is composed of two parallel procedures: a 1-dim convolutional block and the Short-Time Fourier Transform. To properly integrating the two extracted features from different domains, we set the same window size and the same striding for both domains. This gives a 2-dim feature map M_{conv} with F_{conv} channels from the convolutional block and a spectrogram M_{spec} of F_{spec} frequency channels. We concatenate M_{conv} and M_{spec} along the channel/frequency axis while aligning the time frames, giving a hybrid-domain feature map H with $F = (F_{conv} + F_{spec})$ channels and T time frames.

2.3. Separator

The separator has two parts, an embedding network and clustering plus mask estimation.

2.3.1. Embedding Network

In order to estimate the speaker assignment for each T-F index on the hybrid-domain feature map H, we seek a D-dimensional embedding representing each T-F index. We project the elements in the hybrid-domain feature H to D-dimension embeddings through multiple layers of 1-d dilated convolutional blocks, as shown in the middle block in Figure 1. Here the "1-d Conv" block in the figure is actually a residual block consisting of a 1x1-conv, a dilated depthwise convolution and a 1x1-conv module, following the prior work [20]. We stack B residual blocks with the dilation factors of 1,2,...,2^{B-1} and repeat these blocks for R times, followed by a linear layer with a D-dimension vector output for each T-F index on the hybrid-domain feature map H. This gives the embeddings V for H.

2.3.2. Clustering and Mask Estimation

We follow the clustering algorithm previously proposed [14] as shown in Figure 2, starting with K initial centers $e_1, e_2, ..., e_K$ (Figure 2(a)). By arbitrarily choosing N (number of speakers in the mixture) out of the K initial centers (Figure 2(b)), we can get a set of N new centroids by performing k-means clustering for I iterations on the embeddings V. Considering all $\binom{K}{N}$ possible selections out of the K initial centers, we can obtain a total of $\binom{K}{N}$ sets of centroids after performing k-means (Figure 2(c)), among which we choose the set of centroids A with the largest in-set distance (Figure 2(d)). The in-set distance is the minimum distance among all pairs of centroids if $N > 2$. The masks for each speaker $M \in \mathbb{R}^{TF \times N}$ is then estimated by the dot product of the chosen centroids $a_i \in A$ and the embeddings V.

$$M_{i_{t,f}} = V_{t,f} \cdot a_i \qquad \text{for } a_i \in A \tag{7}$$

2.4. Decoder

After multiplying the hybrid-domain feature map H by masks M, we disassemble the masked encoded features into their original components: convolutional feature map $\hat{M}_{conv}$ and frequency domain spectrogram $\hat{M}_{spec}$. As shown in Figure 1, we reconstruct the original signal waveforms from each individual domain: $\hat{s}_{deconv}$ from convolutional feature map and $\hat{s}_{istft}$ from the spectrogram, the former through a deconvolution layer followed by overlap-add method to reconstruct the signals $\hat{s}_{deconv}$; the latter taking the phase of the mixture signal for inverse Fourier Transform to derive $\hat{s}_{istft}$. Weighted sum of the two components with a weight parameter α is then taken as the estimated separated signals $\hat{s}$.

$$\hat{s} = \alpha * \hat{s}_{deconv} + (1 - \alpha) * \hat{s}_{istft} \tag{8}$$

2.5. Training Objective

We take the negative signal-to-distortion ratio as our training objective.

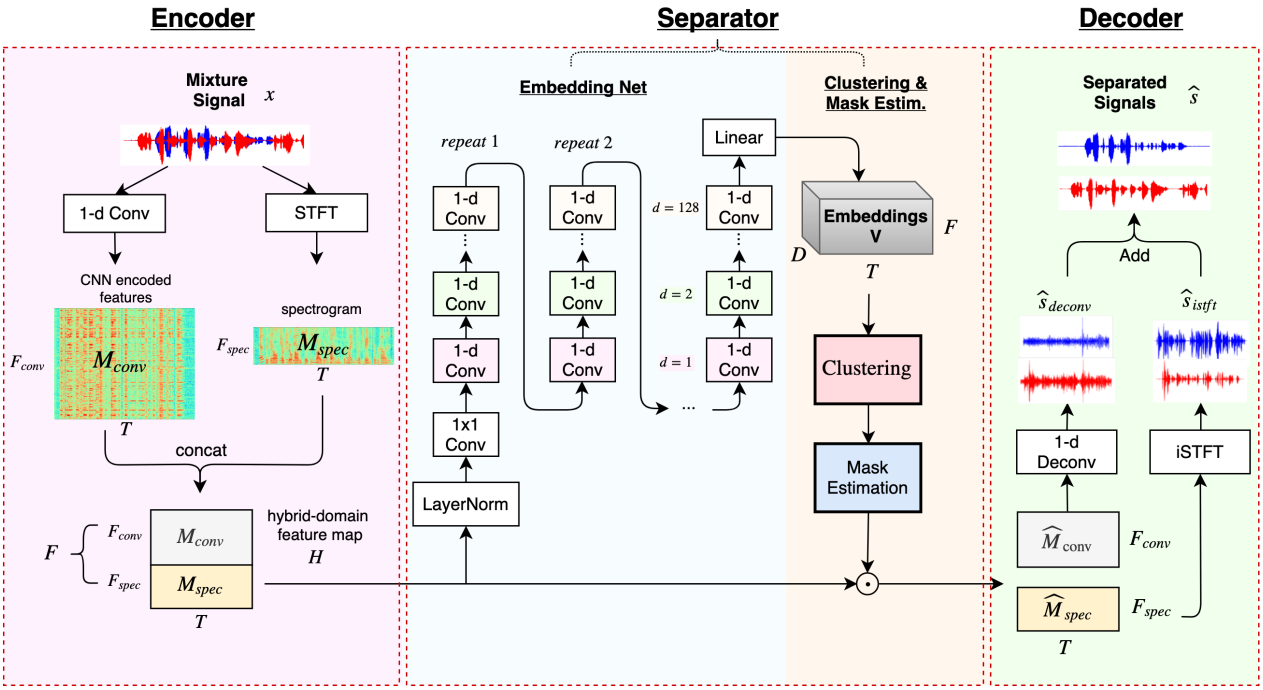

Figure 1: *The proposed approach. The encoder extracts features in both time and frequency domains from input signals. The separator includes an embedding network projecting each element of the feature map to a D-dim vector, followed by clustering and mask estimation for each source speaker. The decoder reconstructs the signal waveforms of each source from the masked features.*

$$\mathcal{L}_{oss} = -SDR(s, \hat{s}) = -10 \log_{10} \frac{\langle s, \hat{s} \rangle^2}{||s||^2 ||\hat{s}||^2 - \langle s, \hat{s} \rangle^2} \quad (9)$$

where $\langle \cdot, \cdot \rangle$ represents dot product and $||s||^2 = \langle s, s \rangle$ denotes the signal power. The training is performed end-to-end, so all components are jointly learned.

3. Experiments

3.1. Datasets

We evaluated the proposed method on publicly avaliable dataset WSJ0-2mix [4], which was derived from WSJ0 corpus. The 30hrs of training set and 10hrs of validation set consisted of two-speaker mixtures generated by different speakers from WSJ0 training set si_tr_s mixed at various signal-to-noise ratio between -2.5 dB and 2.5 dB. The 5hrs of testing set was similarly generated from WSJ0 validation set si_dt_05 and evaluation set si_et_05 produced by 18 speakers. All waveforms were resampled to 8 kHz.

We used in addition environmental sounds from Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [24] in the test. We resampled all types of background noise from 16kHz to 8kHz, and mixed one arbitrarily chosen type of noise into the mixtures in WSJ0-2mix test set with given signal-to-noise ratio (SNR).

3.2. Experimental Setup

The window size of the Short-time Fourier Transform (STFT) and the kernel size for the first convolution layer were both 2.5ms, and the square root Hann window was used for STFT. 20-point DFT was performed to extract the 11-dimensional log magnitude feature, combined with the 256-dimensional feature extracted by 1-d conv, and formed 267-dimensional features in the feature map H.

For the separator, the feature map H first went through a 1x1 Conv block with 256 filters, followed by 8 residual 1-d Conv blocks, with dilated rate of 1,2,...,128, repeated for 4 times. $D = 20$ was chosen to be the embedding dimension following the prior works [4, 14, 15, 17] working on speech separation for better comparison. We set $N = 4$ initial centers and $I = 1$ iteration for k-means following the prior work [14]. We also tested the approach without clustering by passing the output of the last 1-d conv in the embedding net through an additional 1x1 convolution block to estimate N masks without Softmax function as done similar in TasNet [20]. Since the weight α in (8) show minor difference in preliminary experiments, we set $\alpha = 1$ in most of our experiments. The networks are trained from scratch on 4-second segments for 100 epochs using Adam algorithm with permutation invariant training [10, 25].

We evaluate the approach by the signal-to-distortion ratio improvement (SDR_i) [26] and the scale-invariant signal-to-noise ratio improvement (SI-SNR_i) [14].

3.3. Results

We report the performance tested on the WSJ0-2mix test set compared to prior works in Table 1. We see the approach proposed here (row (g)) offered significant improvement with SDR_i = 16.9 dB and SI-SNR_i = 16.6 dB compared to all prior works including the previous state-of-the-art methods on time domain SDR_i = 15.0 dB (TasNet [20] in row (e)) and frequency domain SDR_i = 15.6 dB (Sign Prediction Net [17] in row (f)). In addition, it is worth mentioning that in contrast to the bidirectional LSTM consisting of 600 units on each direction used in the prior works [15, 17], our implementation of depthwise dilated convolution significantly lowered the parameter size by a factor near 1/3.

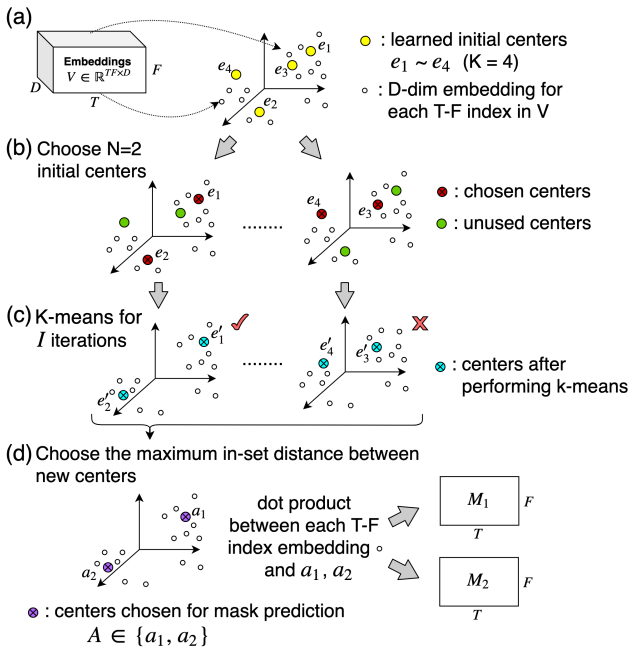

Figure 2: *Clustering and mask estimation procedure ($K = 4$ initial centers and $N = 2$ speakers shown here).*

Table 2 shows the ablation studies for the proposed approach. The upper section (I) is for the separator without clustering, so the masks were directly generated from the embeddings. Here in row (1) the Encoder generated only the time domain features (very similar to TasNet [20] in row (e) of Table 1), while in rows (2)(3) the spectrogram was included in addition. From rows (1)(2) we see the spectrogram improved SDR$_i$ by 0.45 dB (from 15.82 to 16.27 dB, first column of rows (1)(2)). Comparing rows (2)(3) we see the value of α didn't make too much difference, or $\alpha = 1$ is about good enough. This implied the cross-domain features were very useful in jointly learning the various modules, but the time domain features alone were about adequate to generate the precise waveforms.

The lower Section (II) of Table 2 is for the separator including clustering, rows (4)(5) with time domain features only, and rows (6)(7) with cross-domain features. We see adding the clustering module offered improvement in SDR$_i$ of 0.46 dB (from 15.82 to 16.28 dB, rows (4)(5) vs. (1)) for time domain feature alone, and 0.23 to 0.60 dB (from 16.27 to 16.50 or 16.87 dB, rows (6)(7) vs. (2)) for cross-domain features. The spectrogram was certainly useful here too (rows (6)(7) vs. (4)(5)). We also see the choice of the parameter K made the difference for cross-domain features (rows (6)(7)), although this was not clear for time domain features alone (rows (4)(5)). With joint learning including clustering on cross-domain features, we achieve the best result (state-of-the-art to our knowledge) of SDR$_i$ = 16.87 dB in row (7), or 16.9 dB in row (g) of Table 1.

The results for noisy mixture test data (clean data training) with SNRs = 20, 15 dB are listed in the middle and the right columns of Table 2. We see all trends observed above remained true, and a degradation of roughly 1.7 dB or less for 20 dB of SNR, and roughly extra 2 dB for extra 5 dB (20-15) of noise. This showed the robustness of the proposed approach. These results are also shown in Figure 3, for all rows (1)-(7) of Table 2, plus results for SNR = 10 dB which we were not able to include in Table 2.

Table 1: *SI-SNR$_i$ and SDR$_i$ comparison to different prior works tested on WSJ0-2mix dataset. "*" indicates our estimation not written in the original paper.*

	Approaches	Params	SI-SNR$_i$	SDR$_i$
prior works	(a) DPCL++ [13]	13.6M	10.8 dB	-
	(b) uPIT-ST [25]	92.7M	-	10.0 dB
	(c) ADANet [14]	9.1M	10.4 dB	10.8 dB
	(d) Chimera++ [15]	32.9M	11.5 dB	12.0 dB
	(e) TasNet [20]	8.8M	14.6 dB	15.0 dB
	(f) Sign Pred Net [17]	36.8M*	15.3 dB	15.6 dB
	(g) Proposed	10M	16.6 dB	16.9 dB

Table 2: *SDR$_i$ (dB) performance of the proposed approach when the separator included clustering or not (section (I)(II)) and the encoder generated time domain features alone or cross-domain features, with clean or noisy input for different parameters (K: number of cluster centers in Fig 2(a), α: weight in (8)).*

Encoder (with Spectr. or not)	K	α	SDR$_i$		
			Clean Data	Noisy Data	
				20dB	15dB
(I) No clustering in Separator					
(1) Time	-	1	15.82	14.34	12.40
(2) Time + Freq	-	1	16.27	14.69	12.65
(3) Time + Freq	-	0.5	16.28	14.69	12.68
(II) With clustering in Separator					
(4) Time	2	1	16.28	14.71	12.72
(5) Time	4	1	16.28	14.71	12.74
(6) Time + Freq	2	1	16.50	14.88	12.83
(7) Time + Freq	4	1	16.87	15.12	13.00

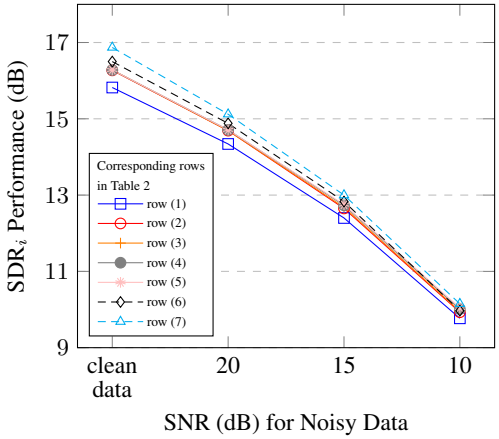

Figure 3: *Performance degradation with noise level for different configurations in Table 2.*

4. Conclusions

In this paper, we propose to integrate the time and frequency domain features and perform cross-domain joint learning for speech separation. State-of-the-art performance of $SDR_i = 16.9$ dB was achieved on the WSJ0-2mix dataset. This verified that the different advantages of the two domains can be well taken, not to mention the correlations between them are useful.

5. References

[1] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[2] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[3] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 31–35.

[5] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[6] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[7] H.-S. Choi, J. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=SkeRTsAcYm

[8] Y. Wang, A. Narayanan, and D. 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.

[9] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[10] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[11] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 61–65.

[12] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[13] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[14] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[15] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 686–690.

[16] Z.-Q. Wang, J. L. Roux, D. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," *arXiv preprint arXiv:1804.10204*, 2018.

[17] Z.-Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," *arXiv preprint arXiv:1811.09010*, 2018.

[18] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.

[19] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[20] Y. Luo and N. Mesgarani, "Tasnet: Surpassing ideal time-frequency masking for speech separation," *arXiv preprint arXiv:1809.07454*, 2018.

[21] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[22] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[23] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," *arXiv preprint arXiv:1808.00158*, 2018.

[24] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. ASA, 2013, p. 035081.

[25] M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen, M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.