SINGLE-CHANNEL SPEECH EXTRACTION USING SPEAKER INVENTORY AND ATTENTION NETWORK

Xiong Xiao, Zhuo Chen, Takuya Yoshioka, Hakan Erdogan
Changliang Liu, Dimitrios Dimitriadis, Jasha Droppo, Yifan Gong

Microsoft, Redmond, WA, USA

[xioxiao, zhuc, tayoshio, haerdoga, chanliu, didimit, jdroppo, ygong]@microsoft.com

ABSTRACT

Neural network-based speech separation has received a surge of interest in recent years. Previously proposed methods either are speaker independent or extract a target speaker's voice by using his or her voice snippet. In applications such as home devices or office meeting transcriptions, a possible speaker list is available, which can be leveraged for speech separation. This paper proposes a novel speech extraction method that utilizes an inventory of voice snippets of possible interfering speakers, or speaker enrollment data, in addition to that of the target speaker. Furthermore, an attention-based network architecture is proposed to form time-varying masks for both the target and other speakers during the separation process. This architecture does not reduce the enrollment audio of each speaker into a single vector, thereby allowing each short time frame of the input mixture signal to be aligned and accurately compared with the enrollment signals. We evaluate the proposed system on a speaker extraction task derived from the Libri corpus and show the effectiveness of the method.

Index Terms— speaker extraction, speech separation, attention, speaker profile.

1. INTRODUCTION

Single channel speech separation is a challenging task for speech processing, where a system is required to separate one speaker's signal from other competing speakers given only their mixture. The separation performance has been largely and continuously improving in the past couple of years thanks to the advances in deep learning-based algorithms [1–11] and has many applications such as robust speech recognition [12].

Deep learning based speech separation systems can be classified into two groups: blind speech separation and informed speaker extraction. For the first approach, the system attempts to recover individual speakers' signals all at once by decomposing the mixture signal into its constituents. For this setup, the model has no preference on the speaker order, i.e. each separated speaker can be assigned to an arbitrary output stream. Thus, the training objective only measures the overall separation quality, which can be measured by averaging the separation scores for all the speaker signals. Because of this order invariant nature, the neural network based blind separation often suffers from the so-called "permutation problem" [1], where the permutation ambiguity prevents gradients from being consistent during training. A solution to this problem is the use of a permutation invariant objective. Representative methods include deep clustering [1], deep attractor network [2] and permutation invariant training (PIT) [3]. In [1, 2], a clustering-oriented objective is used to remove the label ambiguity, while in [3], the best permutation is estimated for every training sample and used to compute the cost.

In contrast, with the informed speaker extraction approach, only a target speaker is extracted, while all other speakers are considered the interference. A bias signal is required to differentiate the target speaker from the rest in the mixture. Different bias types have been proposed in previous works. For example, a previously collected utterance for the target speaker is used as the bias signal in [4, 13–15]. In this paper, the bias utterance is called the profile utterance of the target speaker. It may be obtained from explicit enrollment or learned automatically from previous user interactions. In [16–19], a location based speaker bias is applied under a multi-channel setting. In [20–22], image or video based biases are extracted for the target speaker. In this work, we propose an improvement to the speaker extraction approach with speakers' profile utterances.

Compared with the blind separation approach, the speaker extraction approach is more straightforward as it doesn't suffer from the permutation problem, allowing easier integration with downstream modules such as automatic speech recognition. In addition, the overall performance upper bound is potentially higher since the network only focuses on the reconstruction of the target speaker.

Existing speaker extraction systems have two major limitations [4, 13–15]. Firstly, the speaker bias is provided as a fixed-dimensional speaker vector, obtained by averaging speaker vectors over the profile utterances, as in most systems. Although the speaker vector is able to provide a global bias for the target speaker, e.g., in speaker identification tasks [23, 24], the local dynamics and the temporal structure in the profile utterances, which may be helpful for accurate separation, is now lost due to the averaging. Secondly, current speaker extraction systems usually adopt a "one vs. all" strategy, where only audio snippets of the target speaker are used to bias the system. This approach is reasonable for applications such as public speech denoising, where the interfering speaker identities are unknown. On the other hand, in scenarios such as voice-enabled home devices and office meeting transcription, the audio data for other participating speakers can be available and used for improving the speaker extraction performance.

In this work we propose a novel architecture for the speaker extraction network to address these limitations. The proposed architecture does not reduce the profile audio of each speaker into a single vector. Instead, an attention module is introduced, computing the local similarity between the profile utterances and the mixed speech, thus allowing each short time segment of the input mixture signal to be aligned and accurately compared with the bias signals. This framework also takes advantage of the profile audio data for both the target and the interfering speakers.

The rest of the paper is organized as follows: The proposed

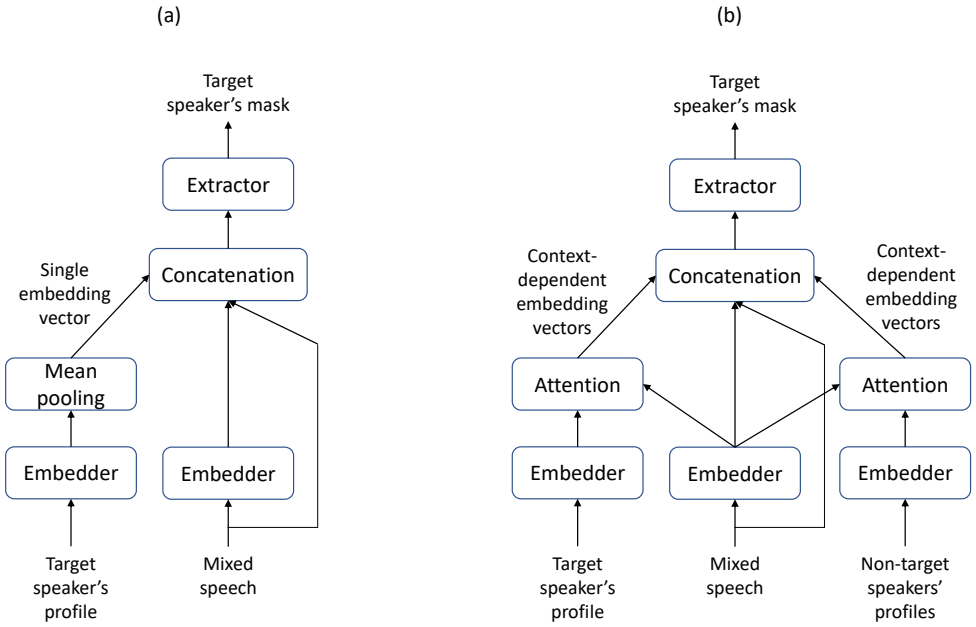

Fig. 1. Model structures: (a) baseline structure with a single pooled embedding vector of the target speaker being concatenated to every frame of the mixed speech; (b) proposed structure context-dependent embedding vectors of both target and non-target speakers are generated through attention.

model is described in Section 2. The experimental setup is presented in Section 3. The results are discussed in Section 4, followed by the conclusions in Section 5.

2. SPEAKER EXTRACTION USING ATTENTION AND PROFILES

2.1. Task Definition

The investigated task is to extract the target speaker's voice from a mixed speech waveform. We know the characteristics of the target speaker through its profile, which is a set of utterances collected either in an enrollment process or from a dataset where the target speaker's label is available. Regarding our knowledge about the competing speaker, there are three scenarios: 1) we don't know the identity of the competing speaker or there is no profile for the speaker; 2) we know the identity of the competing speaker and have his/her profile; 3) we know the competing speaker is a member of a small set of speakers where all profiles are available. The third scenario is common in a gathering that includes a few people and we have the profiles of every speaker. In this paper, our focus is on improving speaker extraction performance in the second and third scenarios.

2.2. Baseline Model

Fig. 1(a) shows the baseline model structure used in this study. A single bias vector is learned for the target speaker and there is no bias vector for the competing speaker. The embedding vector is concatenated to every frame of the mixed speech feature vectors and used as the input of the extraction network. The network only depends on the information of the target speaker bias vector to differentiate the target speaker from the competing speaker. Similar structures have been used recently in [4] and [15], where a global speaker vector or d-vector is used to guide the extraction network. We also apply the embedding network to the mixed speech, making it more comparable to the model in Fig. 1(b).

2.3. Context-Dependent Bias Through Attention

A limitation of the baseline model in Fig. 1(a) is that it summarizes the information about the target speaker in a single bias vector, independent of the amount and content of the profile data. There are two drawbacks of using single speaker bias vector. First, a single vector does not allow a rich representation of all the information in the profile data. Second, the speaker embedding network has no interaction with the mixed speech, meaning that no matter what is spoken in the mixed speech, we always use the same speaker bias for all frames of the mixed speech.

In this paper, we propose to use time-varying, context-dependent bias vectors to represent the characteristics of the target speaker. We choose to use an attention mechanism to generate the bias vectors as follows. Let the target speaker's profile audio and the mixed speech both go through an embedding network, and let $\mathbf{X} \in R^{D \times T_s}$ and $\mathbf{Y} \in R^{D \times T_m}$ denote the target and mixed embedding matrices, respectively. D is the dimension of the embedding vectors, and T_s and T_m are the number of frames of the target speaker's profile audio and the mixed speech, respectively. The context-dependent bias vectors $\mathbf{B} \in R^{D \times T_m}$ are computed from the embedding vectors as

$$B_t = \sum_{i=1}^{T_s} w_{t,i} X_i \quad (1)$$

$$w_{t,i} = \frac{\exp(d_{t,i})}{\sum_{j=1}^{T_s} \exp(d_{t,j})} \quad (2)$$

$$d_{t,i} = Y_t^T X_i \quad (3)$$

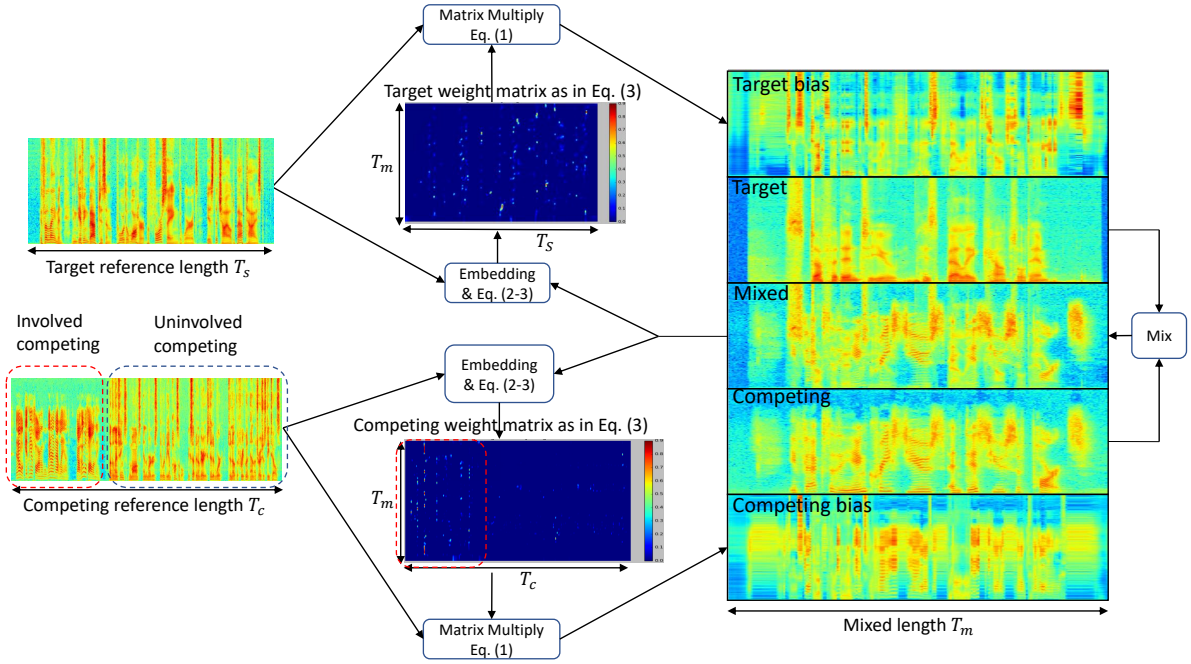

Fig. 2. Effect of attention mechanism. Target and competing speech signals are mixed to obtain the mixed signal in the time domain. Target bias and competing bias are both computed using Eq. (1-3), but with target and competing speakers' profile audio, respecitively.

where $d_{t,i}$ is the inner product of Y_t and X_i and measures their similarity, the weight $w_{t,i}$ is the softmax of $d_{t,i}$ over $i \in [1, T_s]$, and the bias vector at frame t is B_t which is a weighted sum of the embedding vectors of the target speaker's profile data.

The equation in (2) assigns more weight to the target speaker's frames that are more similar to the current mixed speech frame. In this way, the bias vectors will be from similar phonetic/acoustic contexts as the mixed speech frames, and hence context-dependent bias can be achieved. After training, we observed that the weights tend to be very sparse, and the weighted sum of the original target speaker's profile spectrum mimics that of the mixed speech as shown in Fig. 2. This is an indication the attention mechanism is able to align the profile data to the mixed speech, allowing more accurate comparison between them. We will discuss more about the effect of attention in Section 4.1.

2.4. Contrast Bias From Competing Speakers

With the knowledge of the target speaker, the network can accurately extract the target voice from the mixture if the competing speaker has significantly different characteristics. However, it may be challenging for the network to differentiate the target and competing speakers, when the competing speaker's voice is similar to that of the target speaker.

In some applications, we not only have information about the target speaker, but also about the competing speaker, as discussed in 2.1. In such cases, it is beneficial to use the competing speaker's profile to help extract the target speaker's voice. For example, assume we want to extract the voice of speaker A from a mixture which contains the voice of both A and another speaker X, where X could be either B or C. If we have an inventory of sentences from B and C, we can use both of them as the contrast information for extracting the voice of A. By providing the extraction network what kind of voice we want to extract and what kinds of voices we don't want, the network is be able to better extract the target voice. The competing speakers' voices can be processed in the same way as that of the target speaker using equations (1)-(3), as shown in Fig. 2(b). After we obtain the context-dependent bias vectors from both the target and the competing speakers, we can concatenate them with the embedding vectors of the mixed speech and the mixed log spectrum, now feeding the extraction network. This network predicts the time-frequency (TF) mask of the target speaker, or both the target and the competing speaker's masks. We don't need to use the PIT cost function when predicting the masks, as the prediction-target pairing ambiguity is already resolved by the target and competing biases. That is, we already know which predicted mask is corresponding to which source in the mixture.

3. EXPERIMENTAL SETTINGS

3.1. Training and Test Data

We use the Libri speech corpus [25] for model training and testing. Specifically, the two clean training sets of Libri corpus, including both 100 hours and 360 hours sets, are used for training, while the official test set is used for testing. There are 1172 speakers in the training set and 40 speakers in the test set.

The training samples are generated as following:

1. Randomly sample two speakers from the speaker list.
2. For each speaker, randomly sample an utterance longer than the minimum length of 5s.
3. Mix the two sentences using a signal-to-interference ratio (SIR) uniformly sampled from [-2.5dB, 2.5dB]. The length of the mixed speech is set to that of the longer source. The shorter source's starting point is randomly sampled.

4. For each speaker, randomly sample a profile sentence which is longer than the minimum length of 10s and different from the sentences used for generating the mixed speech.
5. Randomly sample two other speakers and one sentence from each speaker. These sentences will be used as the extra competing speakers that do not contribute to the mixed speech.

The training samples are generated on-the-fly, so in every epoch, the network is trained by different training samples. We generated 3000 test samples in the same way as the training samples, but from the test speakers. The test samples are saved to ensure that we always test on the same data. For models not requiring competing speakers' information, these information are not used. For efficiency, we always use 5s mixed segments and 10s profile segments during training. We also don't use the extra competing speakers generated in step 5 during training in this study.

Log scaled magnitude spectrum is used as features for the model. The analysis frame is 32ms long (512 samples for a sampling rate of 16kHz) and shifts by 16ms. The FFT length is 512.

3.2. Model Settings

The embedding network contains 2 bidirectional LSTM (BLSTM) layers[26], with 512 cells in each direction of each layer. The last layer's hidden activations are projected to 512D embedding vectors. The same network is used to generate the embedding vectors for the target/competing speakers and the mixed speech.

The extraction network contains 3 BLSTM layers, each with 512 cells in each direction of each layer. The output is TF masks of 257 dimensions. When competing speakers are used during training, the mask of the competing speaker is also predicted by the model.

The masks are multiplied with the linear magnitude of the mixed speech to generate the extracted magnitude of the sources. The mean squared error between the extracted source magnitude and the true source magnitude is used as the cost function. The networks are trained with the Adam optimizer [27] jointly. The learning rate starts at 1e-4 and exponentially decayed by 0.999 after every 10 hours of training data. The training is stopped after about 30,000 hours of mixed speech have been observed.

4. RESULTS

The speech extraction performance is evaluated using the signal-to-distortion ratio (SDR) [28]. The results are shown in Table 1. The first row is the baseline model that does not use attention and competing speaker's inventory. By using one sentence of the target speaker to generate the speaker bias, a SDR of 9.4dB is obtained. When 5 sentences are used as the target profile, the SDR is improved to 9.8dB. If both target and competing speaker's profiles are used, and attention is used to generate the context-dependent bias vectors, the SDR is improved to 11.5dB (third row). This shows the effectiveness of the proposed model.

In case the exact identity of the competing speaker is not known, we can use all possible competing speakers' profiles. The fourth and fifth rows show the case when 1 and 2 extra speakers are used as competing speakers, respectively. It is observed that the SDR is degraded to 11.1dB and 10.9dB, respectively. The degradation could be due to that the extra speakers may be similar to the target speaker. Also note that the current models are not trained with extra speakers. Despite the degradation, the SDR is still significantly higher than the baseline. When multiple sentences for both the target and competing speakers are available, the SDR is improved further up to 12.1dB as shown in the rest of the table.

Table 1. Results of speaker extraction. "# target sent." and "# compete sent." are the number of sentences in the target and competing speaker's inventories, respectively. "# other compete speaker / sent." is the number speakers and sentences added to the competing inventory who do not contribute to the mixed speech.

Systems	# target sent.	# compete sent.	# other compete speaker / sent.	SDR (dB)
Baseline	1	0	0	9.4
	5	0	0	9.8
Proposed	1	1	0	11.5
	1	1	1-1	11.1
	1	1	2-2	10.9
	2	2	0	11.9
	3	3	0	12.0
	5	5	0	12.1
	5	5	1-1	12.1
	5	5	2-2	12.0

4.1. Insights to context-dependent bias

To gain a deeper insight of the attention-based, context-dependent speaker bias vectors, we show the attention weights and the "synthesized target and competing spectrograms" in Fig. 2. On the left side of the figure, there are the profile spectrogram of the target speaker (top) and concatenated profile spectrograms of two competing speakers (bottom). Only the first competing speaker (circled by red dotted line) contributed to the mixed speech.

On the right hand side of the figure, there are five spectrograms. The second and the fourth are the target and the competing sources, respectively, and the mixed spectrogram is the third. After applying the attention module in Fig. 1, two weight matrices are generated according to equation (2). It is observed that the weights are in general very sparse. In addition, for the competing speakers, the weights of significant values are concentrated in the first part of the weight matrix that belongs to the speaker contributing to the mixed speech (circled by the red dotted lines). This shows that the proposed attention mechanism is able to select relevant speakers from the profiles.

If we multiply the weight matrices with their corresponding profile spectrograms, we can generate "synthesized spectrograms" that tell us what regions of the profile are used by the network as bias. These are the first and fifth spectrograms on the right hand side of the figure. It is interesting to see that these synthesized spectrograms highly resemble the shape of their corresponding source spectrograms. This suggests that the proposed attention mechanism also has the ability of selecting relevant contexts from the inventory. In this way, for every frame of the mixed speech, speaker and content dependent biases are generated from the inventories for the extraction network to separate the underlying source signals.

5. CONCLUSIONS

In this paper, we proposed a novel neural network based speaker extraction model that uses context dependent speaker biases of both the target and competing speakers. An attention mechanism is introduced to select speaker and context relevant examples from the speaker inventory for guiding the extraction network. Experimental results on single channel synthesized mixed speech shows that better extraction performance can be obtained than the baseline model that uses global speaker bias of the target speaker only.

6. REFERENCES

[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

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

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

[4] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Interspeech*, 2017.

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

[6] 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.

[7] J. L. Roux, G. Wichern, S. Watanabe, A. Sarroff, and J. R. Hershey, "Phasebook and friends: Leveraging discrete representations for source separation," *arXiv preprint arXiv:1810.01395*, 2018.

[8] Y. Liu and D. Wang, "A casa approach to deep learning based speaker-independent co-channel speech separation," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5399–5403.

[9] Z. Chen and J. Droppo, "Sequence modeling in unsupervised single-channel overlapped speech recognition," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[10] L. Drude, T. von Neumann, and R. Haeb-Umbach, "Deep attractor networks for speaker re-identification and blind source separation," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 11–15.

[11] T. Tan, Y. Qian, and D. Yu, "Knowledge transfer in permutation invariant training for single-channel multi-talker speech recognition," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[12] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[13] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5554–5558.

[14] J. Wang, J. Chen, D. Su, L. Chen, M. Yu, Y. Qian, and D. Yu, "Deep extractor network for target speaker recovery from single channel speech mixtures," *arXiv preprint arXiv:1807.08974*, 2018.

[15] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, J. Weiss, R, Y. Jia, and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," *arXiv preprint arXiv:1810.04826*, 2018.

[16] Z. Chen, X. Xiao, T. Yoshioka, J. Li, H. Erdogan, and Y. Gong, "Multi-channel multi-speaker overlapped speech recognition with location guided speech extraction network," in *Spoken Language Technology Workshop (SLT)*, 2018.

[17] Z. Chen, T. Yoshioka, X. Xiao, J. Li, M. L. Seltzer, and Y. Gong, "Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5384–5388.

[18] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Multi-channel speech separation with recurrent neural networks from high-order ambisonics recordings," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[19] Q. Liu, Y. Xu, P. Jackson, W. Wang, and P. Coleman, "Iterative deep neural networks for speaker-independent binaural blind speech separation," *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[20] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv preprint arXiv:1804.03619*, 2018.

[21] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," *arXiv preprint arXiv:1804.03160*, 2018.

[22] A. Owens and A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," *arXiv preprint arXiv:1804.03641*, 2018.

[23] S.-X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 171–178.

[24] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[26] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *ICANN '99*, 1999.

[27] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] 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.