Listen and Look: Audio–Visual Matching Assisted Speech Source Separation

Rui Lu, *Student Member, IEEE*, Zhiyao Duan, *Member, IEEE*, and Changshui Zhang, *Fellow, IEEE*

***Abstract*—Source permutation, i.e., assigning separated signal snippets to wrong sources over time, is a major issue in the state-of-the-art speaker-independent speech source separation methods. In addition to auditory cues, humans also leverage visual cues to solve this problem at cocktail parties: matching lip movements with voice fluctuations helps humans to better pay attention to the speaker of interest. In this letter, we propose an audio–visual matching network to learn the correspondence between voice fluctuations and lip movements. We then propose a framework to apply this network to address the source permutation problem and improve over audio-only speech separation methods. The modular design of this framework makes it easy to apply the matching network to any audio-only speech separation method. Experiments on two-talker mixtures show that the proposed approach significantly improves the separation quality over the state-of-the-art audio-only method. This improvement is especially pronounced on mixtures that the audio-only method fails, in which the speakers often have similar voice characteristics.**

***Index Terms*—Audio–visual matching, deep learning, speaker-independent speech source separation.**

I. Introduction

SPEECH separation, also known as the "cocktail-party problem" [1], has long been a challenging task. Its goal is to recover clean source signals from mixed utterances. Although a broad range of methods have been designed [2]–[8], existing techniques are still far behind human capability.

Most speech separation methods only take the speech signal into consideration. Traditional methods, such as computational auditory scene analysis (CASA) [9], [10], nonnegative matrix factorization [11], [12], and hidden Markov model [13], [14], are shallow models and lack the ability to learn from large corpora. Since deep learning was introduced to this problem, performance of speech separation has been significantly improved [3], [4], [6], [8], [15], [16]. Deep methods usually approach the problem through spectrogram mask estimation. Classification-based methods [15]–[17] classify time-frequency (TF) bins to distinct speakers, hence often fail under speaker-independent cases due to the *source permutation problem* [3], [7]: source signals may be well separated at each time frame, but their assignment to sources are inconsistent over time. In auditory scene analysis and CASA [9], this is also called *sequential grouping* or *streaming*. To address this problem, recent methods introduce the concept of clustering to the supervised training paradigm. Among these methods, deep clustering (DC) [3]–[6] and utterance-level permutation invariant training (uPIT) [8] are the current state of the art. Their ideas are either to approximate the clustering affinity matrix [3] or to minimize the separation error [8] by enumerating permutations at the utterance level. During inference, DC and uPIT predict the assignments of all TF bins at the utterance level at once, without the need of frame-by-frame assignment, which is the major cause of the permutation problem. However, when vocal characteristics of the speakers are similar, these methods still suffer from the permutation problem. This is shown in the experiments in [7] and [8], where the optimal permutation is applied at the frame level, separation quality can be further improved.

At a cocktail party, humans employ multiple cues to solve the permutation problem to pay attention to the speaker of interest. These cues include the consistency and continuity of timbre, semantics, and location of the speaker's voice. While timbre has been modeled by DC and uPIT, semantic modeling of the speaker's voice is difficult in the audio mixture and location cues are not captured in single-channel audio recordings. However, in addition to acoustic cues, people also leverage the visual cues to better attend to the voice of the speaker of interest (e.g., lip reading). It is thus a natural idea to exploit the audio–visual correspondence between lip movements and speech utterances [18], [19] to improve speech source separation performance for machines.

This idea has been explored by nondeep models [20]–[25]: in [12] and [21], motion information is exploited to indicate whether speakers are silent; and in [22], [23], and [26], methods are proposed to model audio–visual data joint distributions. These nondeep algorithms often fail to learn from large datasets and the generalization ability is limited. Deep approaches [27]–[29] have been proposed in recent years to reduce noise by exploiting the visual information. However, they are mainly designed for speaker-dependent settings and the separation results are still not satisfactory.

In this letter, we continue the idea of leveraging visual information for speech separation. Specifically, we design a neural network to learn speaker-independent audio–visual matching and use this matching to address the source permutation

Manuscript received May 14, 2018; revised June 29, 2018; accepted July 2, 2018. Date of publication July 5, 2018; date of current version July 19, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61473167 and Grant 61751308, in part by the German Research Foundation (DFG) in Project Crossmodal Learning, in part by the National Natural Science Foundation of China under Grant 61621136008/DFG TRR-169, and in part by the U.S. National Science Foundation under Grant 1741472. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Woon-Seng Gan. *(Corresponding author: Changshui Zhang.)*

R. Lu and C. Zhang are with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: lur13@mails.tsinghua.edu.cn; zcs@mail.tsinghua.edu.cn).

Z. Duan is with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14623, USA (e-mail: zhiyao.duan@rochester.edu).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2018.2853566

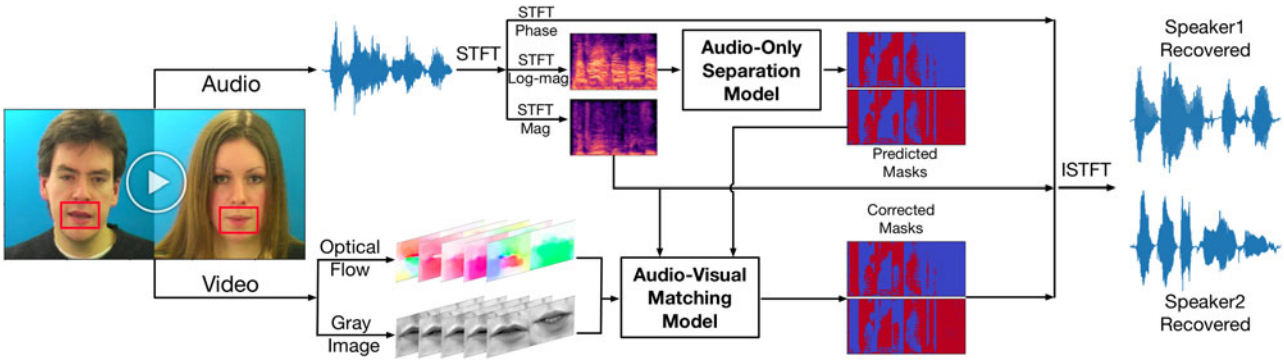

Fig. 1. Proposed audio–visual speech separation framework: The audio–visual matching model integrates both motion (optical flow) and appearance (gray image) information of the lip region to correct spectrogram masks predicted by the audio-only separation model for better separation. As shown in the figure, the permutation problem primarily exists in the beginning quarter time frames of the masks predicted by the audio-only model, and the proposed audio–visual matching network can correct this problem by assigning the predicted masks to the correct speakers.

problem. We carry out experiments on two-speaker mixtures of a publicly available audio–visual dataset to verify the effectiveness of our method. As shown in Fig. 1, the proposed audio–visual matching model successfully corrects the permutation problems in the masks predicted by the audio-only separation model. Our contributions in this letter are threefold: First, we explicitly solve the permutation problem with a deep audio–visual matching network. Second, the proposed approach is especially effective when the performance of audio-only separation is poor. Third, the training procedure of the audio–visual matching model is independent of the audio-only separation model, allowing it to be combined with any mask-estimation-based audio-only separation methods (e.g., [3]–[8]) under the proposed framework.

II. Method

As shown in Fig. 1, we employ DC [3], [4] as the "audio-only separation model" of our system and the state-of-the-art baseline. Input to the DC model is the log-amplitude spectrogram of the input mixture $\mathbf{X}_{\log} \in \mathbb{R}^{F\times T}$. Training procedure of the DC model is to approximate the affinity matrix of the ideal binary mask (IBM) $\mathbf{Y} \in \mathbb{R}^{F\times T\times 2}$ [3], where $\mathbf{Y}_{f,t,1} = 1$ if speaker 1 dominates the TF bin, and $\mathbf{Y}_{f,t,1} = 0$ otherwise. With the predicted mask $\hat{\mathbf{Y}} \in \mathbb{R}^{F\times T\times 2}$ from the DC model and the linear-amplitude spectrogram $\mathbf{X} \in \mathbb{R}^{F\times T}$, we can calculate similarities between the separated audio and visual streams with our proposed "audio–visual matching model," then use these similarities to correct masks predicted by the audio-only model.

A. Audio–Visual Matching Model

The proposed audio-visual matching network aims to learn a shared embedding space [18] where audio and visual frames belonging to the same speaker have higher similarities.

Audio Network: As shown in Fig. 2, given the mixture spectrogram $\mathbf{X} \in \mathbb{R}^{F\times T}$ (upper left) and a mask $\mathbf{Y}$ or $\hat{\mathbf{Y}} \in \mathbb{R}^{F\times T\times 2}$ (upper right), we can recover the source spectrograms $\mathbf{X}^{(i)} = \mathbf{X}\cdot\mathbf{Y}^{(i)}$ or $\hat{\mathbf{X}}^{(i)} = \mathbf{X}\cdot\hat{\mathbf{Y}}^{(i)}$ for $i = 1, 2$. We feed its tth frame (of size $\mathbb{R}^F$) to the audio network (layers $fc1$, $fc2$, $fc3$ with output sizes of 256, 128, and D, and with rectified linear unit (ReLU) nonlinearity) to get an embedding $\mathbf{a}_t$ or $\hat{\mathbf{a}}_t \in \mathbb{R}^D$ $(t = 1, \ldots, T)$.

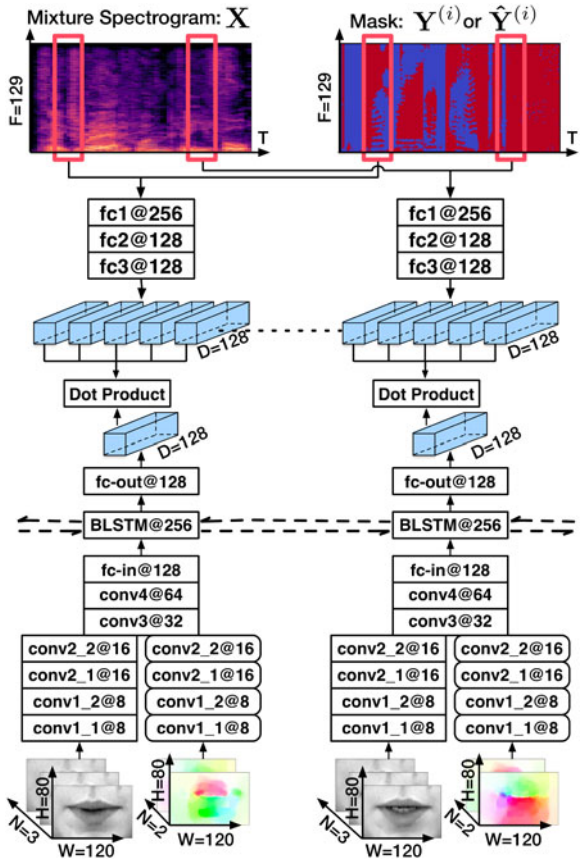

Fig. 2. Audio–visual matching network: Audio and visual streams are encoded as framewise embeddings, we compute inner products of temporally aligned audio and visual embeddings as similarity measure. Every five audio frames correspond to one video frame. All convolution kernels have a size of 3×3.

Visual Network: We adopt two visual geometry group (VGG)-style [18], [30]–[33] convolutional neural networks (CNNs) with the same structure but untied weights (conv1_1 – conv2_2) for early stages of gray image and optical flow inputs. As shown in Fig. 2, N consecutive stacked gray images $(H \times W \times N)$ and optical flow data $(H \times W \times 2)$ of lip regions form the two input branches at each

frame. Outputs of the layers conv2_2 from both streams are concatenated to go through layers conv3 and conv4. All CNN layers use 3×3 filters with a stride of 1, followed by batch normalization [34] and ReLU nonlinearity. There are max-pooling layers with a kernel size of 2 and stride of 2 after conv1_2, conv2_2, conv3, and conv4. Such a "separate-merge strategy" [35], [36] works best in our task. "fc-in" is a fully connected layer with an output size of 128 and ReLU nonlinearity. The feedforward part of our visual network executes on each visual frame separately to capture local invariant features of lip regions [37]. We add a single layer of bidirectional long short-term memory (BLSTM) with a hidden size of 256 on top of the feedforward network to model the contextual information. Outputs of the BLSTM are fed to another fully connected layer "fc-out" with ReLU to get the visual embedding vectors $\mathbf{v}_t \in \mathbb{R}^D$. As the video frame rate is one-fifth of the audio frame rate, here $t = 1, \ldots, \lceil T/5 \rceil$.

Audio–Visual Similarity: We compute similarities of temporally aligned audio and visual embeddings by inner product: similarities between speaker$_i$'s audio frame and visual frames of both speakers at time t are $s_{it}^{+} = \langle \mathbf{a}_{it}, \mathbf{v}_{i\lceil t/5 \rceil} \rangle$ and $s_{it}^{-} = \langle \mathbf{a}_{it}, \mathbf{v}_{j\lceil t/5 \rceil} \rangle$ ($s_{it}^{+}, s_{it}^{-} \in \mathbb{R}$ and $i \neq j$). Similarities obtained by the predicted masks are denoted as $\hat{s}_{it}^{+}, \hat{s}_{it}^{-}$ ($i = 1, 2$). We exploit the relative similarity of audio and visual streams [18], [19], [33], [38] by applying the triplet loss for training and set $m = 1$ empirically

$$L = \frac{1}{2T} \sum_{t=1}^{T} \sum_{i=1}^{2} \max\{s_{it}^{-} - s_{it}^{+} + m, 0\}. \tag{1}$$

B. Mask Correction With Audio–Visual Matching

During the test phase, we first predict masks $\hat{\mathbf{Y}} \in \mathbb{R}^{F \times T \times 2}$ with DC and then compute the audio–visual similarity scores $\hat{s}_{1t}^{+}, \hat{s}_{1t}^{-}, \hat{s}_{2t}^{+}, \hat{s}_{2t}^{-}$. A binary sequence $\{\hat{I}_t\}_{t=1}^{T}$ to decide whether we permute masks predicted by the audio-only model can be obtained by the following equation:

$$\hat{I}_t = \mathbb{1}\{\hat{s}_{1t}^{+} + \hat{s}_{2t}^{+} > \hat{s}_{1t}^{-} + \hat{s}_{2t}^{-}\} \tag{2}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. $\hat{I}_t = 1$ indicates that $\hat{\mathbf{Y}}_t^{(1)}, \hat{\mathbf{Y}}_t^{(2)} \in \mathbb{R}^F$ are in the right order. Otherwise, the permutation problem exists and we swap $\hat{\mathbf{Y}}_t^{(1)}, \hat{\mathbf{Y}}_t^{(2)}$ to correct the predictions. Fig. 3(b) is the sequence predicted by the proposed audio–visual network. Original $\{\hat{I}_t\}_{t=1}^{T}$ can be noisy and we apply a median filter to smooth the results in Fig. 3(c). After we obtain the corrected masks as shown in Fig. 1, we apply the phase of the mixture and use inverse short-time Fourier transform (STFT) to obtain the time-domain signals [3], [4], [7].

III. Experiments

A. Dataset and Setup

We perform experiments on the two-speaker mixtures of the WSJ0 [3] and GRID [39] datasets and report results in terms of delta signal-to-distortion ratio (ΔSDR), signal-to-artifacts ratio (SAR), and signal-to-interference ratio (SIR) [40].

WSJ0 Dataset: WSJ0 is an audio-only dataset [3], [5], [7]. We follow [3] to generate a training set of 30 h, a validation set of 10 h, and a test set of 5 h. We implement the DC model [3] and achieve similar performance (overall ΔSDR = 9.7) with other state of the arts when exploiting a similar network structure. This dataset is used to pretrain the audio-only model in some experiments.

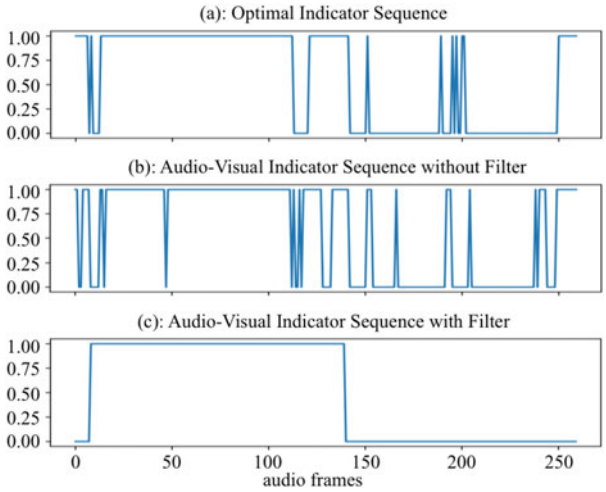

Fig. 3. Permutation indicator sequences for correcting permutation problems in DC separation. (a) Optimal sequence derived from the IBM. (b) Sequence predicted by the audio–visual matching network. (c) Sequence predicted by the audio–visual matching network followed by median filtering.

GRID Dataset: The GRID [39] dataset contains 34 speakers and each of them has 1000 frontal face video recordings. Each video has a duration of 3 s at 25 FPS and resolution of 720×576 pixels. The corpus has a much smaller vocabulary (51 words) compared to WSJ0. We randomly select three males and three females to construct a validation set of 2.5 h and another three males and three females for a test set of 2.5 h. The rest of the speakers form the training set of 30 h.

GRID-Extreme Dataset: We select the same testing speakers used in the GRID dataset, and create a corpus of 1.5 h mixtures that the DC model fail to operate properly. This dataset is to show the benefits of the proposed audio–visual separation method in challenging situations.

Preprocessing and Setup: All audio recordings are downsampled to 8 kHz for STFT with a window size of 32 and 8 ms hop size [3], [5] ($F = 129$ in Section II-A). Linear-amplitude spectrogram (X) and log-amplitude spectrogram ($X_{\log}$) are fed into the "audio–visual" and "audio-only" networks, respectively. Lip regions are detected with the Dlib library [41] as 80×120 patches ($H = 80$, $W = 120$). The stack of three consecutive gray images ($N = 3$) achieves the best performance. We set $D = 128$ as the dimension of the audio–visual embedding space to reduce the dimensionality of the outputs from both modalities. We follow [4] to exploit the optimal structure for DC: 4 BLSTM layers with 300 hidden units for each layer. All models are trained by Adam [42] with a learning rate of $\lambda = 0.001$; we stop training if the validation loss does not decrease for five consecutive epochs.

B. Results on GRID Dataset

We show the separation results in Table I. "GRID" means that the DC model is trained on GRID dataset, whereas "GRID + WSJ0" means that the DC model is pretrained on

TABLE I
COMPARISON OF THE SEPARATION QUALITY ON THE GRID DATASET WITH DIFFERENT MODELS

Models	ΔSDR				SIR				SAR			
	F-F	M-M	F-M	overall	F-F	M-M	F-M	overall	F-F	M-M	F-M	overall
$GRID$	6.23	6.45	9.96	7.89	11.85	13.10	16.60	14.25	9.40	8.77	11.97	10.32
$GRID^{AV}$	7.30	6.23	9.49	7.93	13.33	12.89	16.11	14.40	9.72	8.39	11.51	10.11
$GRID^{AV}_{35}$	**8.40**	**7.02**	**9.88**	**8.64**	**14.66**	**13.91**	**16.54**	**15.25**	**10.72**	**9.06**	**11.89**	**10.75**
$GRID^{\star}$	8.68	7.39	9.90	8.83	14.96	14.27	16.58	15.46	10.89	9.39	11.88	10.88
$GRID + WSJ0$	11.46	7.25	10.48	9.83	17.95	14.24	17.23	16.57	13.59	9.36	12.41	11.87
$(GRID + WSJ0)^{AV}$	8.77	6.40	9.97	8.60	15.07	13.20	16.70	15.23	10.99	8.51	11.91	10.67
$(GRID + WSJ0)^{AV}_{115}$	**11.47**	**7.46**	**10.47**	**9.88**	**17.96**	**14.50**	**17.22**	**16.65**	**13.57**	**9.48**	**12.40**	**11.89**
$(GRID + WSJ0)^{\star}$	11.57	7.85	10.50	10.04	18.12	14.92	17.27	16.83	13.59	9.77	12.40	11.98
IBM	15.80	11.98	13.63	13.77	23.18	19.82	21.07	21.31	17.32	13.42	15.14	15.26

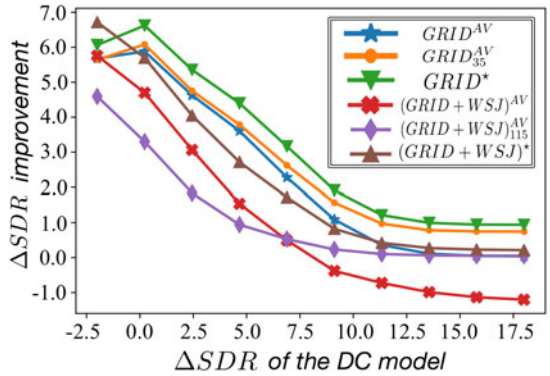

Fig. 4. Improvement on ΔSDR of different settings of the proposed approach against the state-of-the-art DC baseline.

TABLE II
ΔSDR ON GRID-EXTREME DATASET OF DIFFERENT MODELS

Models	F-F (370)	M-M (1202)	F-M (228)	Overall (1800)
$GRID$	2.56	2.31	5.04	2.71
$GRID^{AV}_{35}$	**5.84**	**3.80**	**4.91**	**4.36**
$GRID^{\star}$	6.75	4.62	5.13	5.12
$GRID + WSJ0$	2.22	1.84	2.91	2.05
$(GRID + WSJ0)^{AV}_{115}$	**5.58**	**3.24**	**3.33**	**3.73**
$(GRID + WSJ0)^{\star}$	7.12	4.35	4.28	4.91
IBM	14.41	10.71	11.87	11.61

Numbers in parentheses show the number of mixtures in each setting.

WSJ0 and fine-tuned on GRID. Superscript "AV" shows results by our proposed audio–visual method and "$\star$" shows results of the optimal permutation: For each time frame t, we select the permutation that shows the highest correlation between $\mathbf{Y}_t$ (IBM) and the audio-only predicted mask $\hat{\mathbf{Y}}_t$. Subscripts 35 and 115 indicate the lengths of median filters (in audio time frames), which are chosen empirically.

It is clear that the proposed method improves the separation quality by a large margin on the GRID dataset. We achieve an improvement of 0.75 dB on the overall ΔSDR with $\text{GRID}^{\text{AV}}_{35}$ over the audio-only DC model. The improvement is only the same gender mixtures: 2.17 dB for F–F and 0.57 dB for M–M. This is because in same-gender mixtures, vocal characteristics of the two speakers are similar, which is difficult for "audio-only" separation methods to decide the belongings of audio frames along the temporal axis. When visual cues are introduced, we can better trace the speakers given visual information of the lip regions, thus relieving the source permutation problem. We also observe that median filtering consistently improves the performance, suggesting that audio–visual matching in individual frames can be noisy while considering contextual information stabilizes the result.

When the DC model is pretrained on the WSJ0 dataset, it already achieves a high overall ΔSDR of 9.83 dB, thanks to the rich vocabulary. This makes the improvement of our proposed audio–visual approach (9.88 dB) limited. Even the upper bound (GRID + WSJ0)* (10.04 dB) is only 0.21 dB higher. Nevertheless, the proposed method still matters in cases when the "audio-only" method fails. In Fig. 4, we can see that as the performance of the DC model degrades, the improvement due to the audio–visual model is more pronounced, reaching 4.6 dB improvement on ΔSDR when the DC model's performance is -2.5 dB.

C. Results on GRID-Extreme Dataset

Table II shows the separation quality on GRID-Extreme dataset in terms of ΔSDR. We get overall improvements of 1.65 and 1.68 dB compared to the DC model under both cases. These results indicate that when the "audio-only" model fails to separate speech mixtures properly, the proposed "audio–visual" approach still maintains stable performance.

IV. CONCLUSION

In this letter, we proposed an audio–visual approach to speaker-independent speech source separation. It uses an audio–visual matching network to learn the correspondence between voice fluctuations and lip movements of human speech, and then applies this matching to correct source permutation problems encountered in audio-only separation. The modular design of this approach allows the matching network to work with any audio-only separation methods. Experiments showed significant improvements on separation quality over the state of the art of audio-only speech source separation.

REFERENCES

[1] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica United Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[2] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2014, pp. 577–581.

[3] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. 41st Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 31–35.

[4] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, 2016, pp. 545–549.

[5] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. 42nd Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 246–250.

[6] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 4, pp. 787–796, Apr. 2018.

[7] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. 42nd Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 241–245.

[8] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

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

[10] K. Hu and D. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 122–131, Jan. 2013.

[11] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[12] S. Parekh, S. Essid, A. Ozerov, N. Q. Duong, P. Pérez, and G. Richard, "Motion informed audio source separation," in *Proc. 42nd Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 6–10.

[13] T. Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space," in *Proc. Interspeech*, 2006, pp. 89–92.

[14] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 45–66, 2010.

[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[16] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[17] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. 39th Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 1562–1566.

[18] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3D convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22081–22091, 2017.

[19] R. Arandjelovi and A. Zisserman, "Objects that sound," in *Europ. Conf. Comput. Vision*, 2018.

[20] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 125–134, May 2014.

[21] B. Rivet, L. Girin, and C. Jutten, "Visual voice activity detection as a help for speech source separation from convolutive mixtures," *Speech Commun.*, vol. 49, no. 7–8, pp. 667–677, 2007.

[22] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 96–108, Jan. 2007.

[23] W. Wang, D. Cosker, Y. Hicks, S. Saneit, and J. Chambers, "Video assisted speech source separation," in *Proc. 30th Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. v/425–v/428.

[24] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 358–371, Aug. 2010.

[25] Q. Liu, W. Wang, P. J. Jackson, M. Barnard, J. Kittler, and J. Chambers, "Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5520–5535, Nov. 2013.

[26] Q. Liu, W. Wang, and P. Jackson, "Use of bimodal coherence to resolve the permutation problem in convolutive bss," *Signal Process.*, vol. 92, no. 8, pp. 1916–1927, 2012.

[27] J.-C. Hou *et al.*, "Audio-visual speech enhancement based on multimodal deep convolutional neural network," CoRR, vol. abs/1709.00944, 2017.

[28] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing through noise: Speaker separation and enhancement using visually-derived speech," in *Proc. 43rd Int. Conf. Acoust., Speech, Signal Process.*, 2018.

[29] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement using noise-invariant training," arXiv:1711.08789, 2017.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[31] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2405–2413.

[32] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3444–3453.

[33] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 609–617.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[35] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[36] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1933–1941.

[37] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 87–103.

[38] W. Yuan, S. Wang, S. Dong, and E. Adelson, "Connecting look and feel: Associating the visual and tactile properties of physical materials," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4494–4502.

[39] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[40] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[41] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, no. Jul., pp. 1755–1758, 2009.

[42] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–13.