In: Proceedings of the 21th Annual Conference of the International Speech Communication Association
(INTERSPEECH 2020), Shanghai, China, Oct. 25-29, 2020

Multimodal Target Speech Separation with Voice and Face References

Leyuan Qu, Cornelius Weber, Stefan Wermter

Knowledge Technology, Department of Informatics, University of Hamburg
Vogt-Koelln-Str. 30, 22527 Hamburg, Germany
{qu,weber,wermter}@informatik.uni-hamburg.de

Abstract

Target speech separation refers to isolating target speech from a multi-speaker mixture signal by conditioning on auxiliary information about the target speaker. Different from the mainstream audio-visual approaches which usually require simultaneous visual streams as additional input, e.g. the corresponding lip movement sequences, in our approach we propose the novel use of a single face profile of the target speaker to separate expected clean speech. We exploit the fact that the image of a face contains information about the person's speech sound. Compared to using a simultaneous visual sequence, a face image is easier to obtain by pre-enrollment or on websites, which enables the system to generalize to devices without cameras. To this end, we incorporate face embeddings extracted from a pre-trained model for face recognition into the speech separation, which guide the system in predicting a target speaker mask in the time-frequency domain. The experimental results show that a pre-enrolled face image is able to benefit separating expected speech signals. Additionally, face information is complementary to voice reference and we show that further improvement can be achieved when combining both face and voice embeddings[1].

Index Terms: target speech separation, multimodal speech separation, face reference, speech recognition

1. Introduction

Speech separation aims to recover a clean speech signal from a mixture signal produced by multiple speakers simultaneously, e.g. in a cocktail party environment. Despite the significant progress on speech separation technologies over the past few years [1–3], the permutation problem is still challenging for the speech signal processing community. The permutation problem arises from label ambiguity — the arbitrary order of multi-output — which leads to an inconsistent gradient update and makes a neural network hard to converge during training. According to whether additional information can be available, approaches for solving the problem can be mainly divided into two categories: blind speech separation and target speech separation. The blind speech separation task is to isolate a clean output for each individual source signal without any other information about the observed speech mixture, as shown in Figure 1 (a). To alleviate the permutation problem, deep clustering [1] and its variant, a deep attractor network [4], were proposed to disambiguate the label permutation. Permutation invariant training [5] was presented to predict the best label permutation, whereas the unknown number of sources and invalid outputs are still big challenges in this direction [6].

Different from blind speech separation, target speech separation only recovers the desired single signal guided by auxiliary information, e.g. source directions or target speaker identity, as shown in Figure 1 (b). By leveraging the target speaker reference, target speech separation avoids the permutation problem and is independent of the number of source speakers, since there is only one output per time in this case.

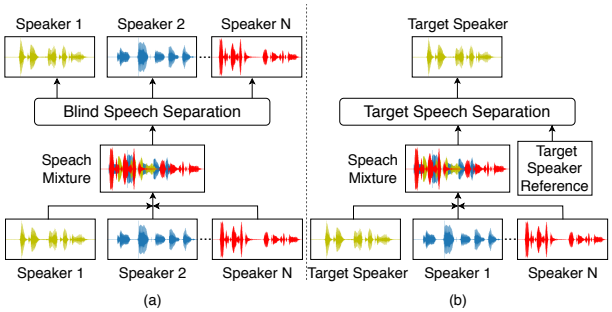

Figure 1: *Comparison of (a) blind speech separation and (b) target speech separation.*

More recently, multimodal audio-visual approaches have shown impressive results in target speech separation and attracted a lot of attention from the computer vision community, for instance, utilizing the lip movement sequences in videos to predict target time-frequency masks or directly generate the target waveform.

Inspired by VoiceFilter [7] which performed target speech separation with speaker voice embeddings and achieved good performance, in this paper, we extend the audio-only VoiceFilter to the audio-visual domain and explore to what extent the visual modality (face embedding) can benefit target speech separation. Additionally, previous audio-visual methods strictly require simultaneous visual streams and highly depend on the visual temporal information. This is hard to meet in most real-world cases, because the speaker's mouth may be concealed by microphone [8] or be undetectable sometimes. Therefore, it is difficult to generalize to devices without cameras. To solve this problem, we propose to integrate the speaker face information into the system, which can be enrolled beforehand and easily applied to more challenging scenarios, for example, if an assistance robot works in public spaces with unknown people addressing it for the first time, then their voice embedding is yet unavailable, while their face image is available.

2. Related Work

Target speech separation: Researchers working in this field try to inform models to only concentrate on the target output utilizing auxiliary information, such as source directions [9, 10], spatial features [11], speaker identity for multi-channel [12] and single-channel [13] setups, speaker profile for both

[1]Web demo: https://leyuanqu.github.io/INTERSPEECH2020/

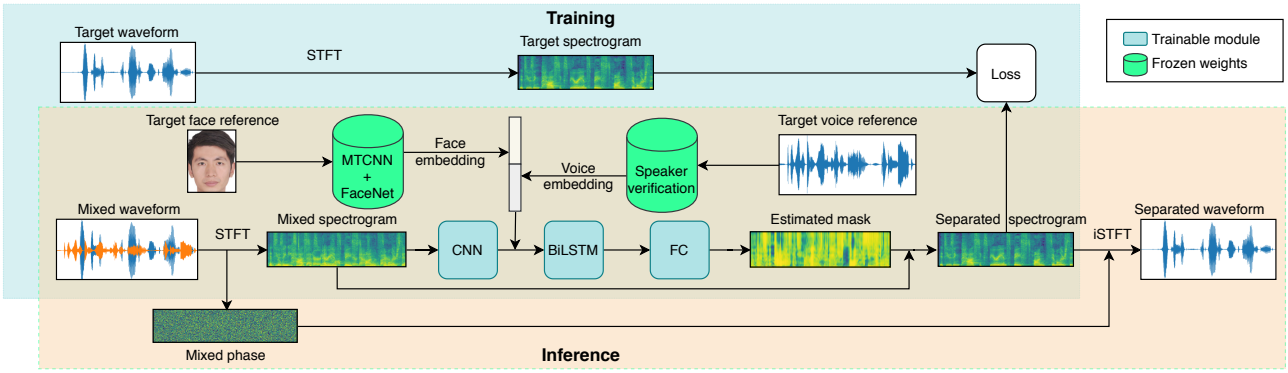

Figure 2: *Overview of the proposed target speech separation architecture. The model receives inputs, i.e. the mixed spectrogram, the face embedding and/or the voice embedding to predict a target speaker time-frequency mask which is used to estimate the target spectrogram.*

the target and competing speakers [14], and so on.

Recently, there has been a growing interest in using multimodal audio-visual methods in target speech separation. Rather than only refining a target spectrogram and reconstructing a waveform with the phase from noisy speech, Afouras *et al.* [2] use convolutional neural networks for both magnitude and phase estimation conditioning on lip regions in the corresponding video. Furthermore, considering the fact that the visual streams may be corrupted from realistic environments, for example, when the mouth region of the speaker is occluded by a microphone, Afouras *et al.* [8] combine lip movement and self-enrolled voice representation to improve the robustness of the proposed system and to prevent the domination of the visual modality. In a similar work, Ephrat *et al.* [3] validate the effectiveness of using the whole face embedding, instead of just the lip area [2, 8], to learn the target speaker magnitude mask based on a large-scale dataset in real-world scenarios. Different from previous works focusing on time-frequency masks, Wu *et al.* [15] directly estimate a raw waveform in the time domain by extending the audio-only (single-modal) TasNet [16] into the audio-visual (multimodal) domain. Gu *et al.* [17] explore the effectiveness of using more information, i.e. speaker spatial location, voice characteristics, and lip movements, in target speech separation. A factorized attention mechanism was introduced to dynamically weigh the three kinds of additional information at the embedding level. Different from previous audio-visual works using corresponding video streams as auxiliary information, the objective of this paper is to investigate the benefit of the pre-enrolled face image for target speech separation.

Learning associations between faces and voices: Inspired by the finding by neuroscientists [18, 19] and psychologists [20, 21] that there is a strong relationship between faces and voices and sometimes humans can even infer what one's voice sounds like by only seeing the face, or vice versa, researchers in computer science have conducted a large number of studies on learning face and voice association that can be mainly divided into two categories: crossmodal representation and joint/shared representation.

Work on the crossmodal representation has led to the possibility of generating one modality from another, e.g. reconstructing human faces by only conditioning on speech signals. Oh *et al.* [22] design neural networks to directly map speech spectrogram to face embeddings which were pretrained for face recognition, then decoded the predicted face representation to canonical face images with a separate reconstruction model. Wen *et al.* [23] utilize generative adversarial networks (GAN) to generate human faces from the output of a pretrained voice embedding network. Instead of using a pretrained network, Choi *et al.* [24] build speech and face encoders on a speech to face identity matching task, and train the encoders and a conditional generative adversarial network end to end to conduct face generation.

Researchers working on joint representation learning attempt to find a joint or sharing face-voice embedding space for tasks of crossmodal biometric retrieval or matching, e.g. searching a corresponding face image via a given speaker voice. Nagrani *et al.* [25] adopt a self-supervision training strategy to learn joint face and voice embeddings from videos without requiring any labelled data. Kim *et al.* [26] introduce triplet loss to learn overlapping information between faces and voices by using VGG16 [27] and SoundNet [28] for visual and auditory modality respectively. Wen *et al.* [29] propose DIMNet to leverage identity-sensitive factors, such as nationality and gender, as supervision signals to learn a shared representation for different modalities. Based on the strong association between faces and voices, we propose to utilize face embedding to guide models in tracking desirable auditory output.

3. Model Architecture

As shown in Figure 2, our proposed model contains three neural networks: a pretrained FaceNet for face embedding extraction, a pretrained speaker verification net for voice embedding extraction, and a mask estimation net (the trainable modules) for target speaker mask prediction.

3.1. Face embedding net

The face embedding net is based on a Multi-task CNN (MTCNN) [30] and FaceNet [31] used in a sequence. Before feeding the original face images into FaceNet, an MTCNN is used for face detection, since the MTCNN performs better in some hard conditions, such as partial occlusion and silhouettes. We crop only the face region and reshape all faces to 160x160 size for face embedding extraction. FaceNet directly learns a unified embedding for different tasks, for example face recognition and face verification, and achieves good results on different benchmarks. In this paper, we use FaceNet Inception-ResNet-v1 in Pytorch[2]. The model is pretrained on the VGGFace2 dataset and achieves 99.65% accuracy on the evaluation set.

[2]https://github.com/timesler/facenet-pytorch

3.2. Voice embedding net

The voice embedding net is based on the model proposed by Wan *et al.* [32] for speaker verification, which consists of 3 LSTM layers with 768 nodes in each layer and one linear layer with 256-dimensional outputs. A generalized end-to-end loss was performed to cluster the utterances from the same class closer while increasing the distance between utterances from different classes during training. The pretrained model[3] used in our paper is trained on the VoxCeleb2 [33] dataset with thousands of speakers. The input spectrogram is extracted using the Short Time Fourier Transform (STFT) with a 40ms hop length and a 80ms window size. The model achieves 7.4% equal error rate on the VoxCeleb1 test dataset (first 8 speakers).

3.3. Mask estimation net

The mask estimation network (the trainable modules in Figure 2) is to predict a target speaker mask in the time-frequency domain, which is heavily inspired by VoiceFilter [7] and the architecture proposed by Wilson *et al.* [34]. As shown in Table 1, the network begins with 7 Conv2D layers with different kernel sizes to capture the variations in time and frequency. Stacked dilated factors enable the network to have larger receptive fields. The output from the last CNN layer is concatenated with voice or/and face embeddings (repeated N times where N is the dimension of the spectrogram in time) as the input of the following bidirectional LSTMs layers. Two fully connected (FC) layers are used to map the high-dimensional outputs from LSTM to the dimension of spectrogram frequency. We use batch normalization and ReLU activation between each layer and a sigmoid function at the output layer. The separated spectrogram is obtained by multiplying the estimated mask and the mixed input. During inference, the separated waveform is reconstructed by the inverse STFT with the phase from the noisy mixture.

Table 1: *Configuration of mask estimation network. Kernel is the kernel size in time and frequency. Dilation is the dilation factor in time and frequency.*

Layer	Kernel		Dilation		Channels/Nodes
	Time	Freq	Time	Freq	
CNN1	1	7	1	1	128
CNN2	7	1	1	1	128
CNN3	5	5	2	1	128
CNN4	5	5	4	1	128
CNN5	5	5	8	1	128
CNN6	5	5	16	1	128
CNN7	1	1	1	1	8
BiLSTM1	-	-	-	-	400
BiLSTM2	-	-	-	-	400
FC1	-	-	-	-	601
FC2	-	-	-	-	601

4. Experimental Setup

4.1. Dataset

We generate the training and test sets based on the Lip Reading Sentences 3 (LRS3) [35] dataset which consists of thousands of speakers' videos from TED and TEDx. The dataset is transcribed on word-level which will be used in our speech recognition experiments.

[3]https://github.com/mindslab-ai/voicefilter

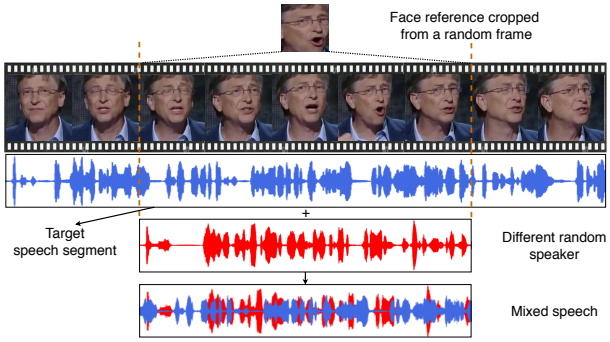

Figure 3: *Dataset building.*

As shown in Figure 3, for the training set, we crop 3s clips from each video where the audio part is treated as the target speech and the visual part is used to get the speaker face from a random frame. To augment the face variants, 10 random faces are extracted from each visual part. The 10 faces are completely out of order and only one face is visible at a time during training. The mixed speech is simulated by directly adding the same length speech from a different random speaker to the target speech. The voice embedding is extracted from a different utterance by the same speaker. Finally, we get 200k samples for around 2k speakers.

For the test set, we use the same process but keep the utterance length in the LRS3 test set and discard the speakers who have only one utterance or there the utterance length is less than 3s. Finally, we get 1171 utterances for 270 speakers. There is no speaker overlap between training and test sets.

4.2. Training

All experiments are conducted on a single NVIDIA Quadro RTX 6000 GPU with 24G memory. We used the Adam optimizer with an initial learning rate of 0.001 and anneal the learning rate with a value of 1.1 after every epoch.

Subsequently, we extract 601-dimension mel-spectrograms with a 25ms window size and a 10ms hop length from mixed speech as model input. Normalization is performed for each mel-frequency bin with the mean and variance.

4.3. Evaluation metrics

We evaluate the model performance with two metrics: Source to Distortion Ratio (SDR) [36] and Word Error Rate (WER). SDR[4] relates the estimated target signal to the noise terms and was found to negatively correlate with the amount of noise left in the separated audio signal [3]. We also evaluate the signal quality with WER by feeding the separated speech into a *Jasper* [37] speech recognition system which is trained on the 960h LibriSpeech dataset and achieves 3.61% WER on LibriSpeech dev-clean set. The evaluation is performed based on the OpenSeq2Seq[5] toolkit published by NVIDIA.

[4]http://craffel.github.io/mir_eval/

[5]https://nvidia.github.io/OpenSeq2Seq/html/speech-recognition.html

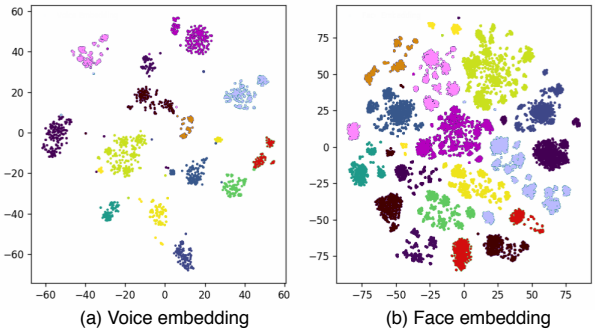

Figure 4: *The visualization of (a) voice and (b) face embeddings for 14 randomly chosen speakers in training set with t-SNE.*

5. Results and Discussion

5.1. Results of speech separation

We visualize the voice and face embeddings from 14 random speakers in the training set. The face embeddings are 10 times more than voices since we randomly crop 10 face images for each mixed speech. As shown in Figure 4, the voice embedding points belonging to the same speakers tend to gather together and significantly far away from other classes. However, the face embedding points from the same speaker are dispersed and close to other classes. We found this is caused by different face angles since all videos are in the wild and the speaker may turn the head from left to right profile while talking.

To investigate the effect of head poses on our experiments, we randomly extract 10 faces for each sample during inference. As shown in Table 2, the performance of using face embeddings fluctuate wildly according to different head poses (Std Dev: 0.32).

Compared to the result of only using voice embedding (10.32±0.11 dB), face information achieves competitive performance (9.23±0.32 dB). The quality of separated speech can be further improved by combining both face and voice references. After checking the output audios, we find that face and voice embeddings are complementary in some cases — in other words, when two voices sound similar, the corresponding faces may be distinguishable, for example, with different skin colors.

Table 2: *Source to distortion rate results for models using only-voice embedding, only-face embedding and both voice+face embeddings (higher is better).*

Reference	SDR (dB)
Voice	10.32±0.11
Face	9.23±0.32
Voice+Face	10.65±0.28

5.2. Results of speech recognition

We test the speech recognition results by Jasper in three settings, clean speech input, mixed speech input and speech separated by our proposed model. The Jasper system achieves 11.8% WER on the clean inputs, whereas the performance dramatically drops down to 71.2% WER when using mixed speech input.

We investigate the separated speech inputs for ASR in two conditions. One is the Separated Speech (Clean) in which we test the performance of our proposed model with clean speech input. A robust speech separation system should not only recover desirable output from mixture, but also have good performance on clean speech input. Table 3 lists the similar results for voice (13.46±0.08% WER), face (15.31±0.19% WER) and voice+face (13.36±0.12% WER) versus clean input (11.83% WER). The other is the Separated Speech (Mixed) in which the ASR receives the separated speech from mixed signals. We can see, in Table 3, the ASR performance can be significantly improved by feeding enhanced speech compared to the 71.22% WER when directly using noisy speech as input. The speech separation system using voice embedding is superior to the one using face embedding. Combining both voice and face references achieves the lowest WER, which is consistent with the evaluation on SDR.

Table 3: *Word error rate on Jasper speech recognition system.*

Input Speech	Model	WER(%)
Clean Speech	-	11.83
Mixed Speech	-	71.22
Separated Speech (Clean)	Voice	13.46±0.08
	Face	15.31±0.19
	Voice+Face	13.36±0.12
Separated Speech (Mixed)	Voice	25.60±0.11
	Face	29.94±0.25
	Voice+Face	23.32±0.12

6. Conclusion and Future Work

In this paper, we propose a novel approach of integrating pre-enrolled face information into the target speech separation task. Our model avoids the speaker permutation problem and the problem of unknown number of source speakers, which audio-only approaches suffer from. In addition, different from the conventional audio-visual speech separation methods which heavily rely on the temporal information from the visual sequences, our system can also be easily adapted to those devices without cameras or to scenarios where no simultaneous visual streams are available. The experimental results on speech separation and speech recognition reveal the effectiveness of face information and the complementarity to voice embeddings.

The face embedding used in our paper is extracted from a model mainly trained on frontal faces (VGGFace2) which is sensitive to the profile views of faces, as indicated in Figure 4. Future work will focus on adding faces from different angles to the face embedding net training. It is also possible to learn the face embeddings via crossmodal distillation [28] in which the voice embedding net transfers its knowledge to the face embedding net. This can be applied to scenarios where no voice embedding is available, for instance, a lecture or a colloquium where the clean speaker voice reference is usually not available, but the speaker face image is accessible on a poster or website.

7. Acknowledgements

The authors gratefully acknowledge partial support from the China Scholarship Council (CSC) and from the German Research Foundation DFG under project CML (TRR 169).

8. References

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

[2] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "The conversation: Deep audio-visual speech enhancement," in *Proceddings of INTERSPEECH*. IEEE, 2018, pp. 3244–3248.

[3] 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.

[4] 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.

[5] 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.

[6] Y. Luo and N. Mesgarani, "Separating varying numbers of sources with auxiliary autoencoding loss," *arXiv preprint arXiv:2003.12326*, 2020.

[7] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proceedings of INTERSPEECH*. IEEE, 2019, pp. 2728–2732.

[8] T. Afouras, J. S. Chung, and A. Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," *arXiv preprint arXiv:1907.04975*, 2019.

[9] J. Liu, D. Perez-Gonzalez, A. Rees, H. Erwin, and S. Wermter, "A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation," *Neurocomputing*, vol. 74, no. 1-3, pp. 129–139, 2010.

[10] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 36–40.

[11] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE Spoken Language Technology Workshop*. IEEE, 2018, pp. 558–565.

[12] 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 *Proceedings of INTERSPEECH*. IEEE, 2017, pp. 2655–2659.

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

[14] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. Dimitriadis, J. Droppo, and Y. Gong, "Single-channel speech extraction using speaker inventory and attention network," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 86–90.

[15] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," *arXiv preprint arXiv:1904.03760*, 2019.

[16] 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.

[17] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *arXiv preprint arXiv:2003.07032*, 2020.

[18] P. Belin, S. Fecteau, and C. Bedard, "Thinking the voice: neural correlates of voice perception," *Trends in Cognitive Sciences*, vol. 8, no. 3, pp. 129–135, 2004.

[19] L. W. Mavica and E. Barenholtz, "Matching voice and face identity from static images." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 2, p. 307, 2013.

[20] V. Bruce and A. Young, "Understanding face recognition," *British Journal of Psychology*, vol. 77, no. 3, pp. 305–327, 1986.

[21] S. R. Schweinberger, D. Robertson, and J. M. Kaufmann, "Hearing facial identities," *Quarterly Journal of Experimental Psychology*, vol. 60, no. 10, pp. 1446–1456, 2007.

[22] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2Face: Learning the face behind a voice," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7539–7548.

[23] Y. Wen, B. Raj, and R. Singh, "Face reconstruction from voice using generative adversarial networks," in *Advances in Neural Information Processing Systems*, 2019, pp. 5266–5275.

[24] H.-S. Choi, C. Park, and K. Lee, "From inference to generation: End-to-end fully self-supervised generation of human face from speech," *arXiv preprint arXiv:2004.05830*, 2020.

[25] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable PINs: Cross-modal embeddings for person identity," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 71–88.

[26] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," *arXiv preprint arXiv:1805.05553*, 2018.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016, pp. 892–900.

[29] Y. Wen, M. A. Ismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," *arXiv preprint arXiv:1807.04836*, 2018.

[30] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[31] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[32] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[33] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[34] K. Wilson, M. Chinen, J. Thorpe, B. Patton, J. Hershey, R. A. Saurous, J. Skoglund, and R. F. Lyon, "Exploring tradeoffs in models for low-latency speech enhancement," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 366–370.

[35] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[36] 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.

[37] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," *arXiv preprint arXiv:1904.03288*, 2019.