FaceFilter: Audio-visual speech separation using still images

Soo-Whan Chung[1], Soyeon Choe[2], Joon Son Chung[2] and Hong-Goo Kang[1]

[1]Department of Electrical & Electronic Engineering, Yonsei University, South Korea
[2]Naver Corporation, South Korea

jsh6293@dsp.yonsei.ac.kr

Abstract

The objective of this paper is to separate a target speaker's speech from a mixture of two speakers using a deep audio-visual speech separation network. Unlike previous works that used lip movement on video clips or pre-enrolled speaker information as an auxiliary conditional feature, we use a single face image of the target speaker. In this task, the conditional feature is obtained from facial appearance in cross-modal biometric task, where audio and visual identity representations are shared in latent space. Learnt identities from facial images enforce the network to isolate matched speakers and extract the voices from mixed speech. It solves the permutation problem caused by swapped channel outputs, frequently occurred in speech separation tasks. The proposed method is far more practical than video-based speech separation since user profile images are readily available on many platforms. Also, unlike speaker-aware separation methods, it is applicable on separation with unseen speakers who have never been enrolled before. We show strong qualitative and quantitative results on challenging real-world examples.

Index Terms: audio-visual, speech separation, speaker isolation, cross-modal biometrics.

1. Introduction

In recent years, there has been great progress in the field of automatic speech recognition, achieving human-level transcription accuracy in quiet environment [1, 2, 3]. However, it still remains a challenge to accurately recognise speech in noisy environments or multi-talker background. Humans have a remarkable ability in separating one person's voice from others [4]. In a *cocktail party* environment, for example, humans can focus on one speaker's voice while filtering out the rest.

The goal of this work is to extract a target speaker's speech from a mixed signal. While there have been researches on separating simultaneous speakers based only on the audio [5, 6], the permutation problem remains unsolved. Although a permutation invariant training method [7, 8] relieves the problem, the permutation problem still exists in the inference stage since there is no explicit constraint for the channel assignment. Furthermore [9, 10] shows that an embedding of the target speaker's voice can be used to separate simultaneously speaking speech as auxiliary information. However, this requires a pre-enrollment of speaker embedding in a clean environment, which also limits its applications.

In recent works of [11, 12, 13, 14, 15], it has been shown that the use of video helps in solving the cocktail party problem. The methods are able to separate speech of a particular speaker by conditioning on his/her lip movements in the corresponding face video. These audio-visual models have demonstrated

---

Video examples: https://youtu.be/ku9xoLh62E4

impressive results, but the dependence on the lip movements means that their usefulness is limited to scenarios where high frame rate video is available.

But what if a single face image can substitute the full video stream? This can be used in a wide range of scenarios since users' profile images are already available in many mobile devices, social networks and company groupware. Once considered a seemingly impossible task, recent works have shown that voice representations can be obtained from face images and vice-versa [16, 17, 18, 19, 20].

We propose a novel audio-visual speech separation (AVSS) model, *FaceFilter*, that is conditioned on a still face image of a target speaker. With our approach, the face image provides speaker identity information to the network, thus it resolves the ambiguity of assigning the separated voice to the speaker. We make the following contributions: (i) we propose a novel method for audio-visual speech separation using only a single face image of the target speaker. The method exploits cross-modal biometric representation obtained from face appearance to solve the permutation problem in speech separation; (ii) the model trained on a large-scale dataset is evaluated on unseen and unheard speakers, on which we demonstrate strong qualitative and quantitative performance; (iii) we conduct further experiments with additional methods for improving separation performance such as temporal attention [21] and speaker representation loss [22].

2. Related works

Research in speech separation has seen significant progress due to the application of neural network models and the availability of new large-scale datasets [23, 24]. The majority of works on speech separation use the *audio* modality only, and they focus on how to effectively extract the portion of a target speaker by analysing spectral characteristic. Spectral embeddings [5, 25] and time-domain speech analysis [6, 26] have shown great success on multiple speaker separation tasks by distinguishing one speaker's voice from mixed input speech. However, they emit separated signals into random channels on the inference stage, where it requires additional clustering step that is vulnerable to permutation. Some works have experimented on the use of speaker embeddings as conditional vectors to regulate the speech assignment. In [9, 10, 27], they obtained speaker embeddings of target speaker from a pre-trained speaker representation model such as x-vector or i-vector [28, 29], and then used the representation as a conditional vector to extract the target speaker's voice from a noisy input speech signal. The performance of these methods far exceeds the previous methods not conditioned on speaker-specific information.

There have been some works that use auxiliary audio-visual information, namely lip movements on video [11, 12, 30, 31]. In these methods, visual embedding learns linguistic informa-
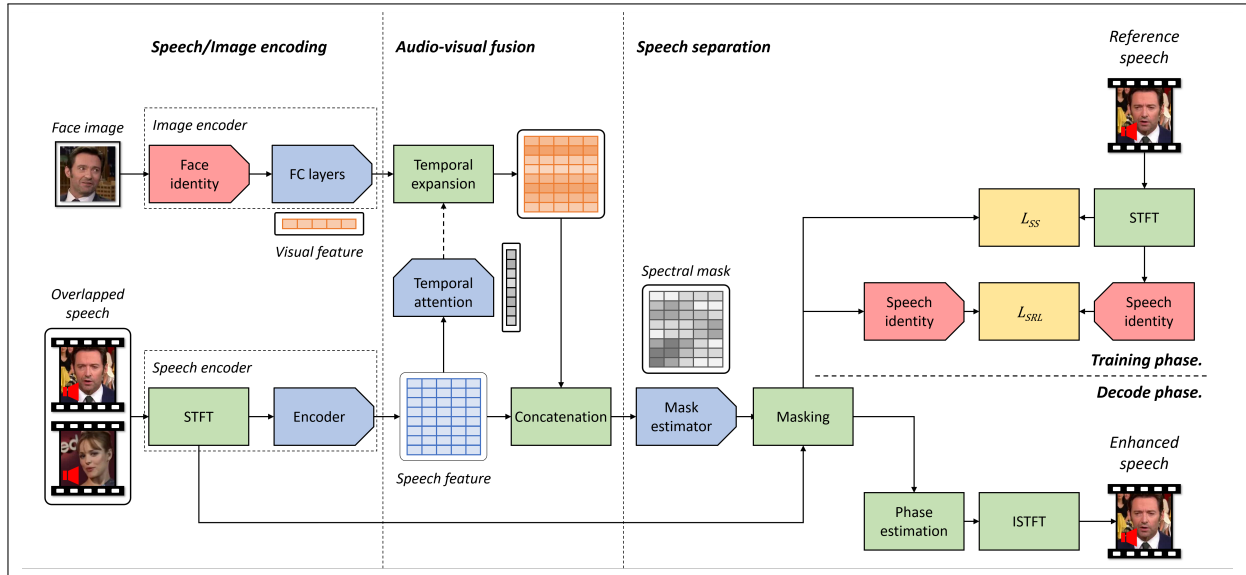
Figure 1: *Description of the proposed audio-visual speech separation network. Blue blocks are trainable neural networks whereas the red blocks are networks pre-trained in cross-modal biometric task.*

tion in a synchronisation task which matches natural correspondence between speech and visual streams. They effectively extract target voice, even with a large number of simultaneous speech; however, these methods are very sensitive to the quality of video.

3. Cross-modal identity representation

Face and voice analysis are popular tools to verify user identity due to their non-invasive nature and good performance. Although the two types of signals share common information (*i.e.* person identity), researches in each field have evolved independently [28, 32] due to the difficulty in building a unified statistical model using heterogeneous features from the different modalities. Speaker recognition analyses consistent spectral characteristics over utterances such as vocal tract shape, pitch and prosody variation, whereas face recognition looks into the shape of facial landmarks. However, there is common higher-order information that are captured by both tasks, including but not limited to age, nationality and gender.

Recent works have demonstrated that joint embedding of face and voice characteristics can be learnt using deep neural networks trained on large-scale datasets. In [16, 18, 33], the problem is set up as a cross-modal matching task using a two stream network. Representations of speech segments and facial appearances are mapped using the networks onto a joint embedding space, and they are trained to predict whether or not the pair comes from same identity. The embeddings are learnt in a self-supervised manner; if audio segment and face image are taken from same video clip, the learning criterion minimises the distance between embeddings whereas it is maximised if they are from different clips.

We exploit these cross-modal identity embeddings to retrieve speaker identity from facial appearance. In particular, we use the implementation of [18] in this paper, where the embeddings are trained as a one-of-many matching task. We sample images and audio segments from talking face videos and train the model with a multi-way matching objective [18], where the model has to select one of the 200 audio segments that comes

from the same video as the face image. Both audio and video streams are based on the VGG-M network [34] – detailed network architecture is given in [18].

4. Audio-visual speech separation

In this section, we describe the proposed architecture and training strategies for the AVSS network.

The most relevant work to the proposed method is the speaker-conditioned speech separation tasks. In this paper, we replace the speaker identity vector with cross-modal identity embeddings extracted from face images instead of that from pre-enrolled speakers' voices. Without having a speaker pre-enrollment step, we can retrieve speaker identity from a profile image in the inference stage, even for *unseen* speaker's voice. The network structure that we use is similar to the one proposed in [11], but with a few changes. It consists of three sub-blocks such as speech/image encoders, audio-visual fusion, and speech separation. The overall structure is illustrated in Figure 1.

Speech/Image encoder. In this stage, speech and face appearances are extracted and encoded onto a latent space, which prepares for the audio-visual fusion. The image encoder block extracts face representations to condition mask estimator. The face identity extractor is pre-trained with the cross-modal identity matching task (Section 3), so that the learnt representation encapsulates joint information between face and voice. The output of the identity extractor is ingested by two additional fully-connected layers that are jointly trainable with the speech separation network. In the speech encoder block, input speech signal is first transformed into a magnitude spectrum using short-time Fourier transform (STFT), then the encoder network transforms the magnitude spectrum into a latent domain representation so that it is easy to combine with a visual representation extracted from the image encoder.

Audio-visual fusion. In the audio-visual fusion stage, we generate audio-visual joint features to represent speaker information as well as spectral information to be separated. Joint audio-visual features are obtained by concatenating speech embeddings and visual embeddings along the channel axis.

We implement two strategies to assign visual identity on speech embeddings. The first one uses equal speaker information on each frame of speech embeddings as done in [9]. The visual identity is concatenated on every speech embedding frame, after verifying its effectiveness in regulating the separation task by consistent guidance. The other strategy is based on self-attention mechanism [21] to provide differently weighted constraints for the separation. Since each speech frame contains different states such as silence, target speech only, interfering speech only and overlapped speech, the degree of separation needs to be varied depending on the characteristic of each frame. We extract weights referred to speech embeddings, and they are multiplied with the identity vector. The weighted visual embedding vector is concatenated with its corresponding speech embedding frame.

Speech separation. The mask estimator predicts a time-frequency mask, which only leaves the desired voice from the observed signal. In particular, it notices the target speaker using audio-visual features and generates a soft mask that considers the power ratio between target and interfering signals. The magnitude spectrum is filtered with the estimated mask to remove undesired components including other speaker's voice. Once the spectrogram is separated, the speech signal is synthesized using phase reconstruction and inverse STFT. We use the Griffin-Lim algorithm [35] to reconstruct phase components because it is better than using the phase of input noisy signal, especially when the input signal is severely distorted by interfering speech.

Learning strategy. Our training strategy is described into two different criteria for speech separation L_{SS} and speaker isolation L_{SRL} respectively.

$$L_{TOT} = L_{SS} + L_{SRL} \qquad (1)$$

The training criterion for the separation is based on minimising mean-squared-error (MSE) between masked spectrogram and magnitude spectrum of reference speech in the logarithm scale (Equation 2).

$$L_{SS} = \left\| log\left(\frac{X}{f(\mathbf{A} \mathbin{+\mkern-8mu+} \mathbf{V}) \odot S} \right) \right\|^2 \qquad (2)$$

where X is target magnitude spectrum, S is mixed input speech spectrum and $f(\cdot), \mathbf{A}, \mathbf{V}$ are the mask estimator, audio embeddings and weighted visual embeddings respectively, where $\mathbin{+\mkern-8mu+}$ is concatenation mark. All network parameters are jointly trained except the face identity extractor in the image encoder.

In addition, we exploit another learning criterion, speaker representation loss [22], where it maximises the cosine similarity of latent distributions between the separated speech and the reference speech using the speech identity extractor pre-trained on cross-modal biometrics task (Equation 3).

$$L_{SRL} = 1 - \cos\big(g(X), g(f(\mathbf{A} \mathbin{+\mkern-8mu+} \mathbf{V}) \odot S)\big) \qquad (3)$$

where $g(\cdot)$ is the speech identity extractor. It helps the model to isolate the target speaker from multiple speakers since the embeddings from the speech identity extractor represent speaker identity, hence only allowing the target speaker's identity to remain in separated signal.

5. Experiments

5.1. Experimental setup

Both the cross-modal biometrics model and the AVSS model are trained on the VoxCeleb2 dataset [36] which contains

Table 1: *Architecture of audio-visual speech separation network. The configurations of FCBlock, ConvBlock, and ResBlock are described in Figure 3.*

Model	Layer	# Channels
FC layers (Image encoder)	FCBlock	1,536
	FCBlock	512
Temporal attention	FC + Sigmoid	1
Encoder (Speech encoder)	ConvBlock	1,536
	ResBlock ($\times 4$)	1,536
	FCBlock	512
Mask estimator	ConvBlock ($\times 15$)	1,536
	FC + Sigmoid	257

celebrity voices from YouTube. The dataset contains 5,994 speakers with a total of 1,092,009 clips in the training set and 118 speakers with 36,237 clips in the test set. Video clips in VoxCeleb2 include various environmental factors such as ambient noise, reverberation, and channel effects.

We prepare the training data for speech separation by mixing two clips from different speakers, while the protocol for generating the training data for the biometric task is identical to that in [16]. There are 2,000,000 samples for training and 100,000 samples for validation, both of which are from the training set speakers. For evaluation, we reserve two types of data, which are *seen-heard* and *unseen-unheard* speakers, and each type has another 4 gender pair subsets, *i.e.,* male-male (*M-M*), male-female (*M-F*), female-male (*F-M*), and female-female (*F-F*). Each evaluation set has 1,000 samples. There is no overlapped evaluation data for the separation task with training set for the biometric task.

Audio representation. Audio stream of the cross-modal biometric network ingests 40-dimensional mel-spectrogram in logarithm scale, extracted at every 10ms with 25ms window length. Audio input to the speech separation module has same settings for slicing and shifting of frames, and input signal is transformed into 257 dimension using short-time Fourier transform. The length of input audio segment is 2 seconds, and the it spectral feature dimension is 40×200 for the biometric network and 257×200 for the separation network.

Image representation. The input to the visual network is a single image containing the person's face of which size is 224×224. For every video in VoxCeleb2, we sample 10 image frames, from which one image is randomly sampled during training. The images are taken from the same video clip as the audio since the cross-modal training does not make use of the identity labels, and this also helps to prevent mismatch in person's physiological characteristics such as age. For the evaluation, the images are randomly taken from other clips with same speaker, which is in line with the practical scenario where a profile image would be used.

Network parameters. Each sub-network of the proposed AVSS model mostly consists of convolutional layers. The detailed parameter settings for the architecture are given in Table 1, and the structures of *FCBlock*, *ConvBlock* and *ResBlock* are described in Figure 3.

5.2. Evaluation protocol

The method is evaluated on two different tasks – speech separation performance and speaker isolation. The first measures

Male-Male | Male-Female
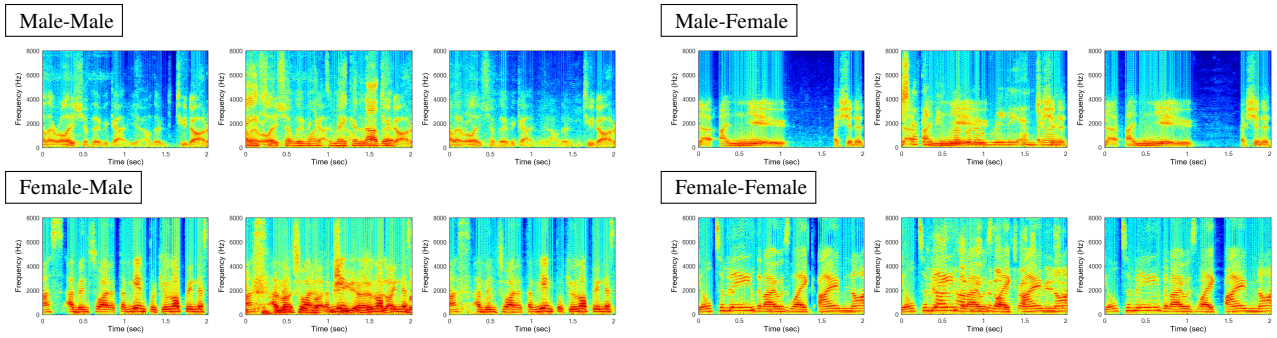Female-Male | Female-Female

Figure 2: *Spectrogram for experimental results. There are 4 samples which are combinations of genders, where genders on the left are the targets and the right are set as interference. Each set demonstrates clean, mixed and separated speech from left to right respectively.*
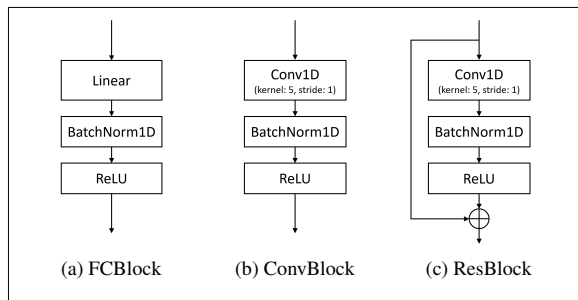

(a) FCBlock   (b) ConvBlock   (c) ResBlock

Figure 3: *Configuration of neural network blocks*

Table 2: *Audio-visual speech separation and speaker isolation results on VoxCeleb2*

Method	Criterion	SDRi	Accuracy
Audio-only	MSE w. PIT	0.690 dB	49.94 %
Audio-visual	MSE	2.348 dB	80.56 %
+attention	MSE	2.413 dB	80.89 %
+attention	MSE w. SRL	2.528 dB	80.20 %

the separation performance, and signal-to-distortion improvement (SDRi) is used to measure the quality of separated signals. The secondary evaluation metric is speaker isolation accuracy which shows whether the output speech is from the target speaker given in the face image. The accuracy indicates the effectiveness of visual information to indicate correct target speaker. Audio-only separation method (baseline) shows an accuracy of 49.94% (*i.e.* random selection).

5.3. Experiment results

Table 2 reports the results of our proposed methods using the metrics described in Section 5.2. Our baseline is a separation network trained using the permutation invariant training (PIT) loss [7], a popular method for speech separation. We compare three variants of the proposed method – the basic AVSS model, the model with attention-based fusion, and the model trained with both L_{SS} and L_{SRL} losses. The basic AVSS model does not include temporal attention and visual vector is padded uniformly along the time direction with the same values.

Note that the SDRi values are lower than those typically reported for source separation on other datasets, since clips in VoxCeleb2 contain environmental noise which degrades the SDR numbers when the network correctly removes the noise from the target.

Table 3: *Speech separation results for gender combinations*

Mixture	SDRi	Accuracy
Seen-Heard Speaker		
Male-Male	2.084 dB	74.8 %
Male-Female	3.039 dB	94.4 %
Female-Male	3.821 dB	93.8 %
Female-Female	1.930 dB	71.2 %
Unseen-Unheard Speaker		
Male-Male	1.256 dB	61.2 %
Male-Female	3.066 dB	92.1 %
Female-Male	3.830 dB	94.5 %
Female-Female	1.198 dB	59.6 %

We report accuracy for different gender pairs, which significantly affects the performance as shown in Table 3. Mixtures with same gender, *i.e., M-M* and *F-F* sets, are more difficult to distinguish by given the images, and the SDR improvement is relatively small. The results for different gender sets, *i.e., F-M* and *M-F* sets, show significantly higher SDR gain and identification accuracy since it is much easier to select the target speech using the given image. Unseen-unheard test sets consist of samples that are not seen during the training of both cross-modal biometrics and speech separation tasks. Although the discrimination performance of unseen people is slightly lower than that of seen pairs, the performance is still far above the baseline.

6. Conclusion

We proposed a novel audio-visual speech separation method that can isolate a specific speaker from multi-talker simultaneous speech using a conditional embedding represented by face image. By using self-supervision, speaker representation can be retrieved from a face image in latent space, which is then used to condition the speech separation network. This approach overcomes the permutation problem that is unavoidable in audio-only source separation, and consistently reconstructs speech from the target identity. The experimental results confirm its effectiveness on the speech separation task.

7. Acknowledgements

8. References

[1] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*, 2016.

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. ICML*, 2016.

[3] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al., "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. ICASSP*, 2018.

[4] Nima Mesgarani and Edward F Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[5] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*. IEEE, 2016, pp. 31–35.

[6] Yi Luo and Nima Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *Proc. ICASSP*. IEEE, 2018, pp. 696–700.

[7] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*. IEEE, 2017, pp. 241–245.

[8] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[9] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John R. Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *INTERSPEECH*, 2019, pp. 2728–2732.

[10] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li, "Time-domain speaker extraction network," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 327–334.

[11] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep audio-visual speech enhancement," in *INTERSPEECH*, 2018.

[12] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 112:1–112:11, 2018.

[13] Rui Lu, Zhiyao Duan, and Changshui Zhang, "Audio–visual deep clustering for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1697–1712, 2019.

[14] Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani, "Multimodal speakerbeam: Single channel target speech extraction with audio-visual speaker clues," *Proc. Interspeech 2019*, pp. 2718–2722, 2019.

[15] Andrew Owens and Alexei A Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. CVPR*, 2018.

[16] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proc. CVPR*, 2018, pp. 8427–8436.

[17] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proc. ECCV*, 2018.

[18] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang, "Perfect match: Self-supervised embeddings for cross-modal retrieval," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2020.

[19] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik, "Speech2face: Learning the face behind a voice," in *Proc. CVPR*, 2019, pp. 7539–7548.

[20] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik, "On learning associations of faces and voices," in *Proc. ACCV*, 2018.

[21] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio, "A structured self-attentive sentence embedding," in *Proc. ICLR*, 2017.

[22] Seongkyu Mun, Soyeon Choe, Jaesung Huh, and Joon Son Chung, "The sound of my voice: Speaker representation loss for target voice separation," in *Proc. ICASSP*, 2020.

[23] Joon Son Chung and Andrew Zisserman, "Learning to lip read words by watching videos," *Computer Vision and Image Understanding*, 2018.

[24] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*. IEEE, 2015, pp. 5206–5210.

[25] Yi Luo, Zhuo Chen, and Nima 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.

[26] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug 2019.

[27] Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Tsubasa Ochiai, Tomohiro Nakatani, Lukáš Burget, and Jan Černocký, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[28] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5329–5333.

[29] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[30] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg, "Visual speech enhancement," in *INTERSPEECH*, 2018.

[31] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," in *INTERSPEECH*, 2019.

[32] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015, pp. 815–823.

[33] Arsha Nagrani, Joon Son Chung, Samuel Albanie, and Andrew Zisserman, "Disentangled speech embeddings using cross-modal self-supervision," in *Proc. ICASSP*, 2020.

[34] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. BMVC.*, 2014.

[35] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[36] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.