A cappella: Audio-visual Singing Voice Separation

Juan F. Montesinos, Venkatesh S. Kadandale*, Gloria Haro*

Universitat Pompeu Fabra

`juanfelipe.montesinos, venkatesh.kadandale, gloria.haro @ upf.edu`

Abstract

Music source separation can be interpreted as the estimation of the constituent music sources that a music clip is composed of. In this work, we explore the single-channel singing voice separation problem from a multimodal perspective, by jointly learning from audio and visual modalities. To do so, we present *Acappella*, a dataset spanning around 46 hours of *a cappella* solo singing videos sourced from YouTube. We propose Y-Net, an audio-visual convolutional neural network which achieves state-of-the-art singing voice separation results on the *Acappella* dataset and compare it against its audio-only counterpart, U-Net, and a state-of-the-art audio-visual speech separation model. Singing voice separation can be particularly challenging when the audio mixture also comprises of other accompaniment voices and background sounds along with the target voice of interest. We demonstrate that our model can outperform the baseline models in the singing voice separation task in such challenging scenarios. The code, the pretrained models and the dataset are publicly available at `https://ipcv.github.io/Acappella/`

Index Terms: singing voice separation, audio-visual processing

1. Introduction

Voices form an integral part of our daily lives. In the form of speech, human voice serves as an effective means of communication. The same voice, when vocalised in sustained tonality and/or rhythm, turns into something musical: the singing voice. The singing voice has become a vital element in the music industry today. Apart from its usage as lead singing voice in songs, it is also found in other diverse forms like rap music, opera singing, solfège, scatting, humming and beatboxing to name a few. *A cappella* refers to a musical arrangement with single or multiple singing voices without any instrumental accompaniment. We are interested in isolating the target voices of interest from such complex musical arrangements.

The particular case of singing voice separation has been largely explored in the context of separating voice from the instrumental accompaniment. The timbral characteristics of singing voice is clearly different from that of the accompanying musical instruments. The audio-only models developed for separating the singing voice from the instrumental accompaniment (*e.g.* [1, 2, 3]) largely benefit from this difference. However, such models do not perform well in the case of separating a particular voice from a mixture of voices or when the volume of the desired target voice is low. In fact, a very similar problem appears in speech separation when there are overlapping speech segments from different sources in a speech mixture. The audio-visual speech separation methods that leverage the visual information to isolate the desired target speech have been shown to outperform their audio-only counterparts [4, 5, 6, 7]

---

* Authors contributed equally.

(the reader is referred to [8] for an extensive review of audio-visual speech separation works). Likewise, we are interested in improving upon the audio-only singing voice separation method by incorporating the visual information. We show that using the visual features is particularly advantageous in the singing voice separation task, especially when the volume of the desired target voice is lower than the background sounds in the audio mixture.

In the audio-visual speech separation works, there are multiple ways in which the visual features are extracted, depending on the front-end representation of the visual information. Many of such works [6, 7, 9, 10, 11] operate directly on the mouth region of the video input to extract the lip motion features. In [12], the motion vectors of face landmarks are used as input to the network that learns the visual features. On the other hand, [4] makes use of face embeddings [13] extracted on the input video frames containing the whole face. These face embeddings are invariant to illumination, pose, and facial expression. The authors show that, apart from the region around the mouth, the facial parts like eyes and cheeks also contribute to the speech separation performance. A very recent work [5] leverages not only the lip motion features but also the facial appearance of the speaker since it is related to certain speech attributes. Their network is trained in a multi-task fashion that jointly learns audio-visual speech separation and cross-modal face-voice embeddings that assist in establishing face-voice mappings. In [14], a single face image of the target speaker is used to condition their audio-visual source separation model on facial appearance. The correlation of voice traits and facial attributes has also proven useful in speaker identification [15] and image generation [16] tasks. Further, [17] points out that facial expressions are helpful in the visual speech recognition task.

While there are different audio-visual benchmark datasets for speech separation (reviewed in [8]), to the best of our knowledge, to date there is no public dataset available for audio-visual singing voice. One of the contributions of the paper is a new dataset with videos of solo performances of people singing *a cappella*, *i.e.* with no musical accompaniment. This dataset can be used to train audio-visual networks for singing voice separation. In a concurrent work, Li [18] created a similar dataset, which is significantly smaller than ours and it has not been made public yet.

We also propose a new audio-visual network for singing voice separation. It is based on a U-Net that processes a complex spectogram and it is conditioned by the motion features extracted by a subnetwork that receives a video sequence of faces cropped around the lips region. The U-Net architecture has been extensively used both in audio-only source separation – both on the spectral [19, 20, 21] and time [22] domains – as well as in its audio-visual counterpart [23, 24, 25, 26, 27, 28]. We can also find works on source separation that condition the U-Net on prior information such as the presence of certain types of musical instruments [29], phoneme activation for singing voice separation [30] or the fundamental frequency contour of each

type of voice sources in choir ensembles [31].

In summary, our contributions are two-fold: i) a new dataset of singers performing with no accompaniment, and ii) a new audio-visual deep neural network for singing voice separation. Both are, to the extent of our knowledge, the first ones presented in the literature with publicly available code and data for reproducibility.

2. The Dataset

In order to exploit the visual information in the singing voice separation problem, we gathered a new dataset of people singing *a cappella*.

The dataset, named *Acappella*, comprises around 46 hours of *a cappella* solo singing videos sourced from YouTube, sampled across different singers and languages. We considered four language categories: English, Spanish, Hindi and others.

The samples in our dataset are defined based on the timestamps corresponding to the segments of interest in each of the videos. We provide these timestamps as a part of the dataset. They have been manually selected to exclude parts of the videos that do not satisfy any of the following characteristics: single frontal face view without occlusions, minimal background noise, no beatboxing, no snapping fingers, songs with lyrics (*e.g.* we avoid humming and yodelling).

Along with the dataset, we provide the splits for training set, validation set and test set. The training set makes up around 80% of the total dataset. Around 7% of the dataset forms the validation set which is used during the training to save the best checkpoint. The test set is divided into the following subsets: seen and unseen. The former consists of samples from known singers, *i.e.* singers seen in the training set but singing different songs. The latter contains singers who are not a part of the training set. The unseen test subset also contains samples from languages not seen in the training set. It presents an approximately uniform distribution of samples across language categories and gender. Extended statistics are shown in Figure 1.
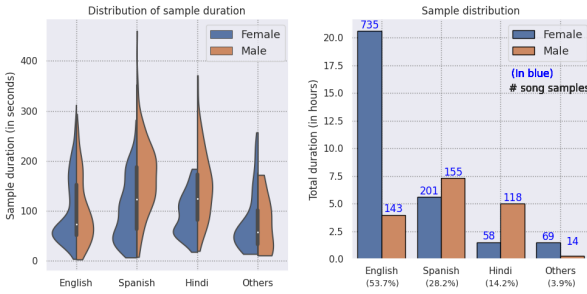

Figure 1: *Acappella dataset statistics.*

In a concurrent work, Li [18] created a similar dataset that has not been made public yet. It comprises 491 YouTube videos and only spans 8h. Thus, to our knowledge, the dataset presented in this paper is the biggest dataset for audio-visual singing voice and, at present, the only one which is public.

We also wanted to test our models to separate voices in multi-voice *a cappella* videos where multiple singing faces are put together in a single view. Since such videos do not provide us with the individual voices for each face, it is not possible to quantitatively evaluate our models on them. Hence, we assembled a multi-voice video ourselves. The mixture contains six voices song by the same person. The lead voice sings in En-

glish, there is a voice emulating a flute and the rest sing in Zulu. The latter singing at unison in pairs most of the time. Background accompaniment music is also added to the mixture.

3. Singing voice separation model

The model is a multimodal convolutional neural network which takes as input a video and its corresponding audio waveform and returns a complex mask. The video consists of a sequence of RGB frames cropped either around the mouth or the face (in case we use visual embeddings) of the target singer. The waveform is mapped into the time-frequency domain by a short-time Fourier transform (STFT). The estimated mask allows to recover the separated voice of the target singer by computing the complex product between the mask and the spectrogram.

Our network is designed for a single singer mainly for two reasons: i) it allows to reduce and bound the memory required for training and ii) it broadens the applicability of the model since the video just needs to visualise the face of the singer with no extra visual information of the additional sources. This way, the model can address mixtures of singing voice with accompaniments of different nature: musical instruments, backing vocals, beatboxing, snapping fingers, ambient sounds or different types of noise.
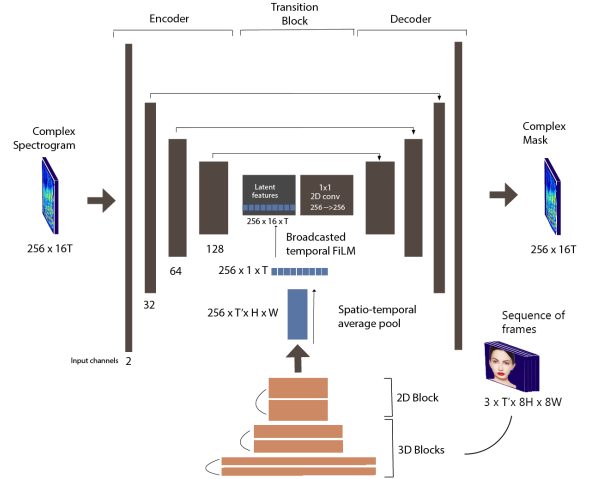

Figure 2: *Y-Net model scheme. The system works with chunks of $4n$ seconds, where $n \in \mathbb{N}$. The audio network takes as input a $256 \times 16Tn$ complex spectrogram and returns a complex mask. The visual network takes as input a set of $100n$ frames cropped around the mouth of the target singer in case of Y-Net-m and Y-Net-r, or face embeddings in case of the Y-Net-e version. The visual features are fused with the audio network's latent space through a FiLM layer. The FiLM broadcasts the $256 \times 1 \times T$ visual features into the $256 \times 16 \times T$ audio ones.*

The architecture is a two-stream convolutional neural network for processing video and audio, it is denoted as Y-Net and illustrated in Fig. 2. The audio network consists of a 4-block U-Net which predicts a two-channel tensor. The U-Net [32] is an encoder-decoder architecture with skip connections in between which allows to preserve the spatial structure while increasing the receptive field through blocks. Each block consists of a sequence of a 5×5 kernel 2D convolution, batch normalization, leaky relu and max pooling.

For the video network we experiment with two different options:

1) **Y-Net-m:** We use a 3-block 3D-ResNet-like network were the first two blocks are traditional 3D convolutional blocks and the last block is a 2D convolutional one. The 3D convolutional blocks process motion information. This design turns into a network with 3M parameters (M stands for million). In contrast, a traditional 3D-ResNet18 has 33.4M and the 2D-ResNet18 has 11.4M. This way, the visual network keeps the capacity to model spatio-temporal information, as suggested in Tran *et al.* [33], while having a contained amount of parameters not to overfit. This network is fed directly with the video frames cropped around the lips region.

2) **Y-Net-e:** In this case, we consider the visual network used in Ephrat *et al.* [4]. The input to this visual network are the face embeddings extracted from the video frames cropped around the face, rather than the video frames themselves, just like in [4]. The face embeddings are obtained for each of the video frames using FaceNet [34], a face recognition model pretrained on the VGGFace2 [35] dataset, a large-scale dataset for face recognition which sums up to 3.31M images. The visual network comprises six 1D dilated convolutional blocks which add up to around 2.56M trainable parameters.

The visual features are fused together with the audio networks's latent features via FiLM conditioning [36]. Note that both, audio and visual features are processed with convolutions, thus the time-frequency and spatio-temporal structures are kept. This allows to fuse both by aligning them in the temporal dimension. We apply a spatio-temporal average pooling to the video features to match the audio ones. At inference time the model can work with chunks larger than 4s, limited by the memory available. This enables a fast processing without artifacts arisen from concatenating masks.

3.1. Pre-processing

Video processing. Videos are resampled to 25 fps to unify the sampling rate. We pre-processed the video stream using a face detector[1] to extract face keypoints, cropping around the face and aligning the face along all the frames in the video. The resulting sequence is resized to 160×160 in case of Y-Net-e. In case of Y-Net-m it is cropped around the mouth region and then resized to 96×96. We feed the visual network with a sequence of 100 RGB frames, corresponding to 4s of video. We normalise w.r.t the mean and the standard deviation. These frames correspond to the face of the target singer.

Audio processing. The audio signal is resampled to 16384 Hz. We consider a 4s-audio excerpt and compute its STFT using a Hanning window of size 1022 and a hop length of 256 (as in [25, 23]) which leads to a 512×256 spectrogram. This specific shape is useful to achieve a perfect alignment between the downconvolutional and the upconvolutional blocks of the U-Net, which are connected through the skip connections. For computational efficiency, we downsample the spectrogram in the frequency dimension to 256×256. Finally, we feed the network with a 256×256 complex spectrogram.

3.2. Training strategy, training target and loss

We train the networks in a self-supervised way generating mixtures artificially. Given a set of N waveforms, $s_1, ..., s_N$, we generate an artificial mixture by taking the average, *i.e.* $s_m = \frac{1}{N} \sum s_i$. This way we can ensure the resulting mixture is bounded between -1 and 1. The network is trained to optimise an L_2 loss on bounded complex ratio masks [37].

---

[1] https://github.com/DinoMan/face-processor

Let $S_i(f, t)$ be the STFT of a generic waveform s_i. Note that $S_i(f, t)$ is a complex matrix. We define the ideal complex ratio mask as follows:

$$M(f, t) = \frac{S_i(f, t)}{\sum S_i(f, t)} \qquad (1)$$

Since the mask M in (1) is not bounded, we apply a hyperbolic tangent on the real and imaginary parts, M^r and M^i, respectively, to obtain a bounded complex mask:

$$M_b(f, t) = \tanh M^r(f, t) + \tanh M^i(f, t)\, i \qquad (2)$$

We also apply a gradient penalty (3) on the loss term so that the points of the spectrogram with higher energy contribute more to the loss.

$$G(f, t) = \max(\min(\log(1 + |S_m(f, t)|), 10), 10^{-3}) \quad (3)$$

Let $\hat{M}_b$ be the bounded mask estimated by the network. The loss function is defined as follows:

$$\mathcal{L} = \left\| G^{\frac{1}{2}} \odot (\hat{M}_b^r - M_b^r) \right\|_2^2 + \left\| G^{\frac{1}{2}} \odot (\hat{M}_b^i - M_b^i) \right\|_2^2$$

where $\odot$ denotes the element-wise product.

4. Experiments

We conduct a set of experiments comparing the Y-Net against its audio-only counterpart, the U-Net (*i.e.* our Y-Net without the visual subnetwork), and a state-of-the-art model for speech separation, the model of Ephrat *et al.* [4], that we denote as LLCP. We consider different variants of our model: Y-Net-m, Y-Net-e (both defined in Sect. 3), and Y-Net-r. Y-Net-r is the same network as Y-Net-m but it has been trained with mixtures in which 50% of the time the mixture includes two lead voices (from *Acappella* dataset) instead of just one. The purpose is to have a model that better separates a singing voice in this kind of mixtures.

A very common problem with multimodal networks is how to force the model to pay attention to one modality when the task is easy to solve from the other modality alone. When the patterns of each sound source are clearly different the source separation is easier from the audio modality. Thus, we artificially create mixtures with different types of accompaniments, including human voices. Since we only need the face of the target singer, we mix samples from *Acappella* together with samples from AudioSet [38]. Audioset is an in-the-wild large-scale dataset of audio events across more than 600 categories. We gathered the categories related to the human voice and some typical accompaniments. These categories are: acappella, background music, beatboxing, choir, drum, lullaby, rapping, theremin, whistling and yodelling. In addition, we use MUSDB18 dataset [39] to have pop and rock examples as accompaniment. In order to generate an artificial mixture we ensure that all the samples from *Acappella* are used in each epoch. Those are mixed with a random sample from AudioSet or MUSDB18. We adjust the sampling distribution so that all the categories are sampled uniformly. Including Audioset in the training strategy increases the robustness of the model and addresses over-fitting.

All the models have been trained using stochastic gradient descent, with a momentum of 0.8 and a weight decay of 10^{-5}. The learning rate is 0.01. In case of Y-Net-m, we use pretrained weights from Kinetics [40] and its statistics to normalise the input frames.

We are interested in analysing the role of different types of visual information in different kind of mixtures. For that, we evaluate the models in two different scenarios: mixing a single singing voice with accompaniment (SV+A) and mixing two singing voices with accompaniment (2SV+A). Besides, we use different volume levels in the singing voice, so that experiments range from predominant singing voice to non-dominant one. To do so, each source s_i in the mixture is normalised by its energy and then the singing voice is further multiplied by a factor α, where $\alpha \in \{0.25, 0.5, 1, 1.25\}$. Lastly, we rescale the sources to ensure they are bounded between -1 and 1 while respecting the relative preset volumes. Results for these experiments in terms of Signal-to-Distortion Ratio (SDR) and Signal-to-Interference Ratio (SIR) are shown in Fig. 3.
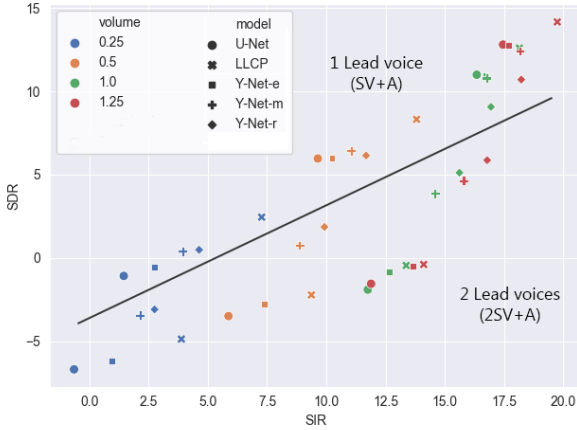

Figure 3: *SDR and SIR values for different models and experiments on the test seen set. Different symbols are assigned to the different models and different colours to the different volume levels of the lead voice. Results of experiments where two lead voices are mixed are shown below the line and the ones with one lead voice above it (they have bigger SDR and SIR values since it is a simpler task).*

For the SV+A setup, we can observe that U-Net performs really well for higher volume levels. We hypothesise that the system is capable of learning what the predominant voice is and separating that from the accompaniment even if it consists of backing vocals. On the contrary, when the volume of the singing voice is low, visual information helps to better recover the target voice. Both the SDR and SIR values of the audio-only method are more degraded with respect to the audio-visual ones as the volume level of the lead voices decreases.

Some interesting results arise from the 2SV+A setup as well. As there is no predominant voice, U-Net fails to recover the isolated signal. On the other hand, Y-Net-m incorporates motion information from the lips and outperforms LLCP in such a challenging situation despite having three times less parameters. We hypothesise that visual embeddings do not sufficiently encode motion information. This follows the observations of [13], which explains that visual embeddings ignore factors of variation related to the instant such as lightning, pose and expression (this latter more related to lips position). Nevertheless, we also tested Y-Net-e in order to discard that the improvement is due to the use of U-Net as backbone. Furthermore, it can be observed that if we train Y-net-m with two lead voice samples, *i.e.* the Y-Net-r model, results get even better.

Finally, we conduct experiments for unseen singers and unseen languages to check how the network generalises. Results are in Table 1. The first four methods have been trained with the same kind of mixtures (only one lead voice) and the best results are in bold font. We found the performance is similar for all the languages, even for unseen ones. Regarding genders, we found the performance is better for female, as the dataset is unbalanced and contains more female samples (see Fig. 1). The U-Net is biased to predict female voices over male ones while audio-visual models can better predict male voices thanks to the visual information. Table 1 also provides the quantitative results of different models in the separation of the lead vocals in our multi-voice video. This singer was not seen in the training set.

Model	English		Unseen languages		Multi-voice	
	SDR	SIR	SDR	SIR	SDR	SIR
U-Net	-2.10	11.66	-1.98	10.64	2.79	6.67
LLCP	-1.07	13.75	-1.25	11.67	4.72	8.43
Y-Net-e	-1.69	12.21	-1.46	11.33	2.46	7.32
Y-Net-m	**2.39**	**13.76**	**1.81**	**12.25**	**6.95**	**10.36**
Y-Net-r	3.16	13.71	2.13	12.55	6.11	10.81

Table 1: *SDR and SIR values for different models in three different experiments. Two of them mix two lead voices of the test unseen set: singers singing in English (left values) and singers singing in unseen languages in the training set (center values). The third experiment (right values) corresponds to the multi-voice video we assembled. The first four methods have been trained with the same kind of mixtures (only one lead voice), bold font indicates best results for those methods.*

Some examples of different kind of mixtures (including the multi-voice one) and the estimated sources in each case are provided in the website of the project: `https://ipcv.github.io/Acappella/`.

5. Conclusions

This paper explores the singing voice separation problem from a new perspective, by exploiting both the audio and visual information. For that, we introduce a new dataset of video recordings of *a cappella* solo performances. We also propose a new audio-visual singing voice separation model, based on a U-Net conditioned on the lips motion of the target singer. The experiments show how audio-visual methods improve upon audio-only ones in challenging scenarios where there are different voices or the target voice has a relative low volume. The presented method is compared to a state-of-the-art audio-visual speech separation method trained in the new dataset. Our method better exploits the lips motion and thus outperforms when separating two singing voices.

6. Acknowledgements

7. References

[1] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 106–110.

[2] D. Samuel, A. Ganeshan, and J. Naradowsky, "Meta-learning extractors for music source separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2020, pp. 816–820.

[3] T. Li, J. Chen, H. Hou, and M. Li, "Sams-net: A sliced attention-based neural network for music source separation," in *Int. Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021.

[4] 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," in *SIGGRAPH*, 2018.

[5] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," *arXiv preprint arXiv:2101.03149*, 2021.

[6] V.-N. Nguyen, M. Sadeghi, E. Ricci, and X. Alameda-Pineda, "Deep variational generative models for audio-visual speech separation," *arXiv preprint arXiv:2008.07191*, 2020.

[7] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[8] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *arXiv preprint arXiv:2008.09586*, 2020.

[9] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 3244–3248.

[10] C. Li and Y. Qian, "Deep audio-visual speech separation with attention mechanism," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2020, pp. 7314–7318.

[11] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Interspeech*. ISCA, 2018, pp. 1170–1174.

[12] G. Morrone, S. Bergamaschi, L. Pasa, L. Fadiga, V. Tikhanoff, and L. Badino, "Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2019, pp. 6900–6904.

[13] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, "Synthesizing normalized faces from facial identity features," in *Proceedings of the IEEE Con. on Computer Vision and Pattern Recognition*, 2017, pp. 3703–3712.

[14] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, "FaceFilter: Audio-Visual Speech Separation Using Still Images," in *Proc. Interspeech*, 2020, pp. 3481–3485.

[15] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 276–292.

[16] 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 *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7539–7548.

[17] A. Fernandez-Lopez, O. Martinez, and F. M. Sukno, "Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database," in *IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2017)*, 2017, pp. 208–215.

[18] B. Li, *Multi-Modal Analysis for Music Performances*. PhD thesis, University of Rochester, 2020.

[19] G. Meseguer-Brocal and G. Peeters, "Conditioned-u-net: Introducing a control mechanism in the u-net for multiple source separations," in *Int. Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[20] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 23–27.

[21] V. S. Kadandale, J. F. Montesinos, G. Haro, and E. Gómez, "Multi-channel u-net for music source separation," in *IEEE Int. Workshop on Multimedia Signal Processing (MMSP)*, 2020.

[22] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2018.

[23] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2019.

[24] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[25] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proc. of the European Conf. on Computer Vision (ECCV)*, 2018, pp. 570–586.

[26] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, "The sound of motions," in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2019, pp. 1735–1744.

[27] X. Xu, B. Dai, and D. Lin, "Recursive visual sound separation using minus-plus net," in *Proc. of the Int. Conf. on Computer Vision*, 2019, pp. 882–891.

[28] L. Zhu and E. Rahtu, "Visually guided sound source separation using cascaded opponent filter network," in *Proc. of the Asian Conference on Computer Vision*, 2020.

[29] O. Slizovskaia, L. Kim, G. Haro, and E. Gomez, "End-to-end sound source separation conditioned on instrument labels," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2019, pp. 306–310.

[30] G. Meseguer-Brocal and G. Peeters, "Content based singing voice source separation via strong conditioning using aligned phonemes," in *Int. Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[31] D. Petermann, P. Chandna, H. Cuesta, J. Bonada, and E. Gomez, "Deep learning based source separation applied to choir ensembles," *arXiv preprint arXiv:2008.07645*, 2020.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[33] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. of the IEEE conf. on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[34] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. of the IEEE conf. on Computer Vision and Pattern Recognition*, 2015.

[35] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *IEEE int. conf. on automatic face & gesture recognition*, 2018.

[36] V. Dumoulin, E. Perez, N. Schucher, F. Strub, H. d. Vries, A. Courville, and Y. Bengio, "Feature-wise transformations," *Distill*, vol. 3, no. 7, p. e11, 2018.

[37] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, 2015.

[38] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2017.

[39] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[40] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.