Estimating Individual A Cappella Voices in Music Videos with Singing Faces

Venkatesh S. Kadandale [*] , Juan F. Montesinos [*] , Gloria Haro
Universitat Pompeu Fabra
{venkatesh.kadandale, juanfelipe.montesinos, gloria.haro}@upf.edu

1. Introduction

A cappella refers to a musical arrangement composed of single or multiple singing voices without any instrumental accompaniment. We are interested in isolating the target voices of interest in multi-voice *a cappella* music videos. 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.* [12, 19]) 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 [5, 7, 16]. 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 when the volume of the desired target voice is lower than the background sounds in the audio mixture and when there are overlapping singing voices.

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 [1, 11, 16] operate directly on the mouth region of the video input to extract the lip motion features. In [15], the motion vectors of face landmarks are used as input to the network that learns the visual features. On the other hand, [5] makes use of face embeddings [3] 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 [7] 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 [2], a single face image of the target speaker is used to condition their audio-visual source separation model on facial appearance.

In a concurrent work, Li [10] explored the specific task of audio-visual singing voice separation. Li's audio-visual singing voice separation method particularly outperformed the audio-only baseline methods when the input sample contained backing vocals in addition to the target voice. Our work is also along the similar lines but we focus on the effect of volume of the target voice in the source separation quality. Further, our approach also differs from Li's work in terms of the choice of baseline models, the proposed model architecture, the experimental setup and the dataset.

While there are different audio-visual benchmark datasets for speech separation (reviewed in [14]), 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.

The U-Net architecture has been extensively used both in audio-only source separation methods [9, 13, 18] as well as in its audio-visual counterpart [6, 22, 23]. In this paper, we propose a new audio-visual network based on U-Net. It is conditioned by the motion features extracted using a visual network that operates on a sequence of aligned faces cropped around the lips region.

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. The code, the pretrained models and the dataset are publicly available at https://ipcv.github.io/Acappella/

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 of around 46 hours of *a cappella* solo singing videos sourced from YouTube, sampled across different singers and languages. It encompasses 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. The segments have been manually selected to exclude parts of the videos that do not satisfy any of the following characteristics: single frontal face view

[*] Authors contributed equally.

without occlusions, minimal background noise, no beatboxing, no snapping fingers, songs with lyrics. The timestamps defining these segments are published as a part of the dataset.

The dataset also comes with the pre-defined splits for training, validation and testing. 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. Extended statistics are shown in Figure 1.

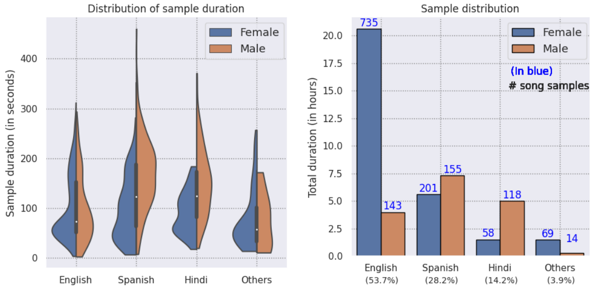

Figure 1. *Acappella* dataset statistics.

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 sourced from YouTube 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. It contains six voices sung by a female singer.

Li [10] created two datasets of audio-visual solo singing voices. One comprises 491 videos curated from YouTube and the other 65 recorded videos. Together, they sum up to 12 hours of solo singing videos. On the other hand, our dataset, *Acappella*, spans around 46 hours and to our knowledge, it is the biggest dataset of audio-visual solo singing voice and, at present, the only one which is public.

3. Singing voice separation model

The model is a multimodal CNN which takes in the video frames and the complex spectrogram of the audio mixture as the input and returns a complex mask for the target voice. The video consists of a sequence of RGB frames of the target singer cropped either around the mouth or the face (in case we use visual embeddings). 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 sound 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.

The architecture is a two-stream convolutional neural network for processing video and audio, denoted as Y-Net. The audio network consists of a 4-block U-Net which predicts a two-channel tensor corresponding to the real and imaginary parts of the complex mask. For the video network, we experiment with two options:

1) **Y-Net-m:** We use a 3-block 3D-ResNet-like network where the first block is a 3D convolutional block and the last blocks are 2D convolutional ones. The 3D convolutional block processes motion information. This design turns into a network with 3M parameters (M 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.* [20], 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.* [5]. The input to this visual network are the face embeddings extracted from the video frames cropped around the aligned face, rather than the video frames themselves, as in [5]. The visual network comprises of six 1D dilated convolutional blocks which add up to around 2.56M trainable parameters.

The visual features and the audio network's latent feature are aligned temporally and fused together through FiLM conditioning [4].

3.1. Pre-processing

Video processing. We process the video stream using a face detector[1] to crop and align the face along all the frames in the video. The resulting sequence is resized to 160 × 160 and visual embeddings are computed for each of the frames as in [5]. We store the frames after resizing them to 96 × 96. We feed the visual network with 100 RGB frames (4s, 25fps) containing 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 obtaining a 512×256 spectrogram. For computational efficiency, we downsample the spectrogram in the frequency dimension to 256×256. Finally, we feed the network with the complex spectrogram.

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

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, we generate an artificial mixture by taking their average. The network is trained to optimise an $\mathcal{L}_2$ loss on bounded complex ratio masks [21].

Let M be the ideal complex ratio mask for the target source. Since the mask M is not bounded, we apply a hyperbolic tangent on the real and imaginary parts of M, at every time-frequency coordinate, to obtain a bounded complex mask. The loss function to optimise is an $\mathcal{L}_2$ loss with a gradient penalty, that weights the mask error in every time-frequency coordinate proportionally to the magnitude of the spectrogram of the mixed audio.

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 network), and a state-of-the-art model for speech separation, the model of Ephrat *et al.* [5], that we denote as LLCP. 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 rather than one.

When the sound sources have clearly distinct texture, using the audio modality alone could result in a good level of source separation. But, such a model might not work well when it comes to separating a target voice from a mixture containing multiple voices, which is the case of multi-voice *a cappella*. To separate a target voice in such multi-voice *a cappella*, the lip motion information tracked in the visual modality could be useful. To make the model to pay attention to the lip motion during training, we also include human voices as accompaniments to the samples from *Acappella*. The accompaniment samples are sourced from MUSDB18 [17] and the following categories of AudioSet [8]: acappella, background music, beatboxing, choir, drum, lullaby, rapping, theremin, whistling and yodelling.

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 target 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 RMS value and then the singing voice is further multiplied by a factor α, where $\alpha \in \{0.25, 0.5, 1, 1.25\}$. Results for these experiments in terms of Signal-to-Distortion Ratio (SDR) and Signal-to-Interference Ratio (SIR) are shown in Fig. 2.

For the SV+A setup, we can observe that U-Net, an audio-only model, performs really well for higher volume levels of the target voice. 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. To the contrary, when the volume of the target voice is low, visual information helps to better recover the target voice. The audio-only method performs poorly in comparison with the audio-visual methods as the volume level of the target voice decreases.

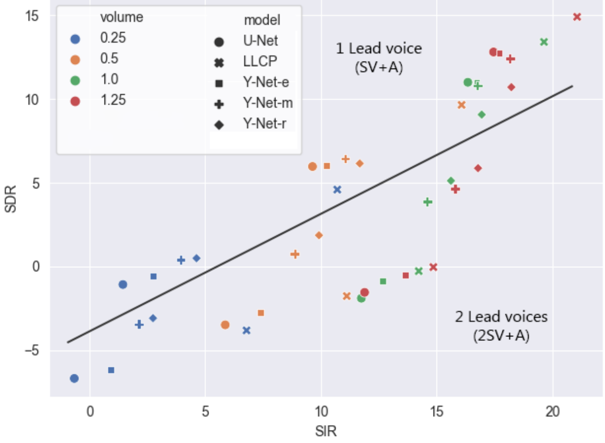

Figure 2. SDR and SIR values on the test seen set.

Some interesting results arise from the 2SV+A setup as well. When the target voice is not the predominant voice, the U-Net fails to recover the target voice. On the other hand, note that the 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 of LLCP do not sufficiently encode motion information. This follows the observations of [3], which explains that visual embeddings ignore factors of variation related to the instant such as lighting, pose and expression (which are related to the lips position). Nevertheless, Y-Net-e that makes use of visual embeddings, outperforms the U-Net. Furthermore, note that the Y-Net-r model performs better than the Y-net-m in 2SV+A setup.

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	-1.19	**14.22**	-1.20	**12.49**	5.63	9.55
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 on the test unseen sets (left and center) and our multi-voice video (right). The test unseen sets are evaluated in the 2SV+A setup.

Finally, we evaluate on the unseen singers, unseen languages and our multi-voice video to check how well the models generalise. Results are in Table 1. The first four models have been trained with mixtures in which 50% of the time the mixture includes two lead voices rather than one. Among them, Y-Net-m performs the best in terms of SDR. The test unseen sets in Table 1 contain both male and female singers. Evaluating on gender specific test subsets showed that the performance is better for female target voice. It could be because the dataset is unbalanced and contains more female samples (see Fig. 1). We notice that the U-Net is biased as it tends to predict female voices over the 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. Again, this singer is not a part of the training set.

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 videos of *a cappella* solo performances. We also propose a new audio-visual singing voice separation model, based on a U-Net conditioned on the lip motion of the target singer. Our experiments show that the audio-visual methods improve upon the audio-only method in challenging scenarios. The presented method is compared to a state-of-the-art audio-visual speech separation method trained on the new dataset. Our method better exploits the lip motion information and thus largely outperforms our baseline models in terms of SDR in separating a target voice mixed with another singing voice and an accompaniment.

References

[1] T. Afouras, J. S. Chung, and A. Zisserman. The conversation: Deep audio-visual speech enhancement. In *Proc. Interspeech*, pages 3244–3248, 2018. 1

[2] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang. FaceFilter: Audio-Visual Speech Separation Using Still Images. In *Proc. Interspeech*, pages 3481–3485, 2020. 1

[3] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Synthesizing normalized faces from facial identity features. In *Proc. IEEE Con. on Computer Vision and Pattern Recognition*, pages 3703–3712, 2017. 1, 3

[4] V. Dumoulin, E. Perez, N. Schucher, F. Strub, H. d. Vries, A. Courville, and Y. Bengio. Feature-wise transformations. *Distill*, 3(7):e11, 2018. 2

[5] 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. 1, 2, 3

[6] R. Gao and K. Grauman. Co-separating sounds of visual objects. In *Proc. of the IEEE Int. Conf. on Computer Vision*, pages 3879–3888, 2019. 1

[7] R. Gao and K. Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. *arXiv preprint arXiv:2101.03149*, 2021. 1

[8] 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. 3

[9] V. S. Kadandale, J. F. Montesinos, G. Haro, and E. Gómez. Multi-channel u-net for music source separation. In *IEEE Int. Work. on Multimedia Signal Processing*, 2020. 1

[10] B. Li. *Multi-Modal Analysis for Music Performances*. PhD thesis, University of Rochester, 2020. 1, 2

[11] C. Li and Y. Qian. Deep audio-visual speech separation with attention mechanism. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 7314–7318, 2020. 1

[12] 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. 1

[13] 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. 1

[14] 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. 1

[15] 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*, pages 6900–6904, 2019. 1

[16] 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. 1

[17] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner. The MUSDB18 corpus for music separation, 2017. 3

[18] 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. 1

[19] 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)*, pages 106–110, 2018. 1

[20] 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*, pages 6450–6459, 2018. 2

[21] D. S. Williamson, Y. Wang, and D. Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM trans. on audio, speech, and language processing*, 24(3), 2015. 3

[22] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *European Conf. on Comp. Vision (ECCV)*, pages 570–586, 2018. 1

[23] 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. 1