Audio-visual voice separation transformer

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

1. Introduction

Human voice is usually found together with other sounds. Think of people speaking in a cafeteria or in a social gathering, a journalist reporting on the scene, or an artist singing on a stage. In these situations we can find: multiple concurrent speeches, speech with background noise or a single or multiple singing voices with music accompaniment among others. Our brain is capable of understanding and concentrating on the voice of interest. This cognitive process does not only rely on the hearing. Some works have shown the sight helps to focus on the voice of interest [6] or to resolve ambiguities in a noisy environment [10]. In this paper we address the voice separation and enhancement problems from a multimodal perspective, leveraging the motion information extracted from the visual stream to guide the resolution of the problem.

The contributions of this work are several: i) We propose an audio-visual (AV) transformer-based model which produces state-of-the-art results in speech and singing voice separation. ii) We show how an enhancement stage based on a light network can boost the performance of AV models over larger complex models, reducing the computational cost and the required time for training. iii) We reveal that AV models trained in speech separation do not generalise good enough for the separation of singing voice because of the different voice characteristics in each case and that a dedicated training with singing voice examples clearly boosts the results. Finally, iv) our method is an end-to-end gpu-powered system which is capable of isolating a target voice in real time (including the pre-processing steps). Demos are available at http://ipcv.github.io/VoViT/.

2. Related work

In the last years there has been a fast evolution of deep-learning-based audio-visual works for speech separation and enhancement. We refer the reader to a recent review in [12]. Due to lack of space we review only the works most related to our proposal and the ones we compare with.

A two-step speech enhancement process was proposed in [1]. In the first step, a two-tower stream processed the audio-visual information to extract a binary mask that performed separation on the magnitude spectrogram. Afterwards, the phase of the spectrogram was predicted by passing the estimated magnitude spectrogram together with the noisy phase spectrogam through a 1D-CNN. Most AV source separation methods comprise of a two-tower stream architecture (one for each modality). These methods largely involve either of the two common variants: encoder-decoder CNNs (usually with a U-Net as backbone) (e.g. [5, 13]) or recurrent neural networks (RNNs), both conditioned on visual features (e.g. [4, 14]).

Transformers have been used in audio-only source separation [19]. Very recently, audio-visual transformers were investigated in [15] for main speaker localisation and separation of its corresponding audio. In [16] an audio-visual transformer was used for classification in order to guide an unsupervised source separation model. Finally, in [2] a transformer was used for audio-visual synchronisation.

Many AV speech separation methods rely on lips motion extracted from raw video frames to guide the task. To our knowledge, there are only two works that used face landmarks, instead of video frames, [14, 13].

Most recent algorithms use lips motion as well as appearance information, usually implementing cross-modal losses to pull together corresponding audio-visual features [5, 11].

3. Approach

Given an audio-visual recording with several speaking/singing faces, with or without any other accompaniment, our goal is to recover their isolated voices by guiding the voice separation with the visual information present in the video frames. More formally, given the audio signal of each speaker, $s_i(t)$ (where t denotes time), the mixture of sounds can be defined as $x(t) = \sum_i s_i(t) + n(t)$ where $n(t)$ denotes any other sound present in the mixture, i.e. background sounds. Therefore, the task of interest can be defined as the estimation of each individual isolated voice $\hat{s}_i(t)$. In our approach $\hat{s}_i(t) = F(x(t), v_i(t))$, where F is a function represented by a neural network that receives the visual information of the speaker of interest, $v_i(t)$.

Our solution comprises of a two-stage neural network that operates in the time-frequency domain. The first stage consists of an AV voice separation network which can isolate the target voice at a good quality. To alleviate the computational cost, we propose to use downsampled spectrograms in this stage. The second stage consists of a recursive lead voice enhancer network that works with full resolution spectrograms. The networks at both stages are trained independently. The whole model is illustrated in Fig.1.

Stage 1: Audio-Visual Voice Separation. For simplicity, let us denote by $s(t)$ the voice signal we want to isolate, and by $S(f,t)$ its corresponding spectrogram. The audio waveform of the mixture, $x(t)$, is transformed into a complex spectrogram $X(f,t)$ applying the Short-Time Fourier Transform (STFT). Once the waveform is mapped to the time-frequency domain, we can define a complex mask $M(f,t)$ that allows to recover the spectrogram of the esti-

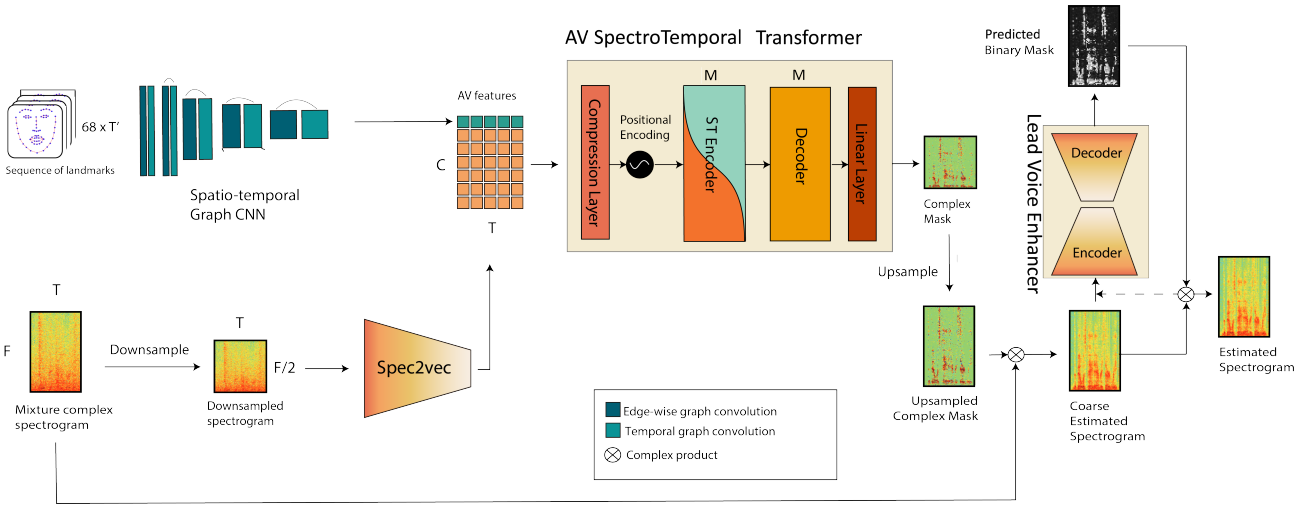

Figure 1. Proposed network. Audio and video features are concatenated in the channel dimension before being fed to the transformer.

mated source with a complex product, denoted as $*$, that is: $S(f,t) = X(f,t) * M(f,t)$ Then, the goal of the network in the first stage is to estimate the complex mask $\hat{M}(f,t)$. The optimal set of parameters of the network is found by minimising the following loss: $\mathcal{L} = \|G \odot (M_b - \hat{M}_b)\|^2$, where M_b and $\hat{M}_b$ are, respectively, the ground truth and estimated bounded complex masks, $\odot$ denotes the element-wise product, $\|\cdot\|$ is the $L2$-norm and G is a gradient penalty term which weights the time-frequency points of the mask according to the energy of the analogous point in the mixture spectrogram X. The audio waveform of the estimated source can be computed through the inverse STFT of the estimated spectrogram $\hat{S}(f,t) = X(f,t) * \hat{M}(f,t)$.

To solve the voice separation problem, we propose to leverage the face motion information present in the video frames of the target person whose voice we want to isolate. For that, we use a spatio-temporal graph neural network that processes the face landmarks to generate motion features. The audio features are generated by a CNN encoder, denoted as *Spec2vec*. Both audio and motion features preserve the temporal resolution and are concatenated in the channel dimension, then they are fed into a transformer. All the submodules have been carefully designed to achieve a high-performance low-latency neural network.

Spatio-temporal graph CNN: To reduce the computational cost of the visual stream, we propose to use face landmarks together with a spatio-temporal graph CNN [18]. This network, similar to that in [13], was redesigned to preserve the temporal resolution. It consists of a set of blocks which apply a graph convolution over the spatial dimension followed by a temporal convolution. This way we can considerably reduce the amount of data to process and to store, from $96 \times 96 \times 3 \approx 3 \cdot 10^4$ values per frame to $68 \times 2 \approx 10^2$. This supposes a substantial reduction in the storage necessities when working with large audio-visual datasets.

Spec2vec: Transformers need proper embeddings to achieve high performance. We use the audio encoder of [4] to generate embeddings without losing temporal resolution.

AV spectro-temporal transformer: Transformers appeared as an efficient solution to RNNs. They are trained with a masking system allowing to process all the timesteps of a sequence in parallel. However, these architectures operate sequentially at the time of inference, like the RNNs. To overcome this issue we use an encoder-decoder transformer, which can solve the source separation problem in a single forward pass.

We design our AV ST-transformer encoder upon the findings of [19]. The AV ST-transformer has 512 model features across 8 heads. We tried 256 features but it works worse. The compression layer is nothing but a fully connected layer followed by GELU [9] activation which maps the C incoming channels to the 512 channels required by the architecture. It is composed by M encoders and M decoders. The encoder is a set of two traditional encoders in parallel, which processes the signal from a temporal and a spectral point of view [19].

Stage 2: Lead voice enhancer. Although lips motion is correlated with the voice signal and may help in source separation, it is not always accessible or reliable. For example, the scenarios involving a side view of the speaker or a partial occlusion of the face or an out-of-sync audio-visual pair make it challenging to incorporate the lips motion information in a useful way; all such scenarios may appear in unconstrained video recordings. In [13], the authors show that audio-only models tend to predict the predominant voice in a mixture when there is no prior information about the target speaker. Based on this idea, we hypothesise that, if the first stage of the AV voice separation network outputs a reasonable estimation of the target voice, this voice will be predominant in the estimation. Upon this idea, we use an

audio-only network which identifies the predominant voice and enhances the estimation without relying on the motion, just on the pre-estimated audio. To do so, we simply use a small U-Net which takes as input the estimated magnitude spectrogram (at its original resolution) and returns a binary mask. We trained this network to optimise a weighted binary cross entropy loss.

There are different reasons to use binary masks. On the one hand, we found qualitatively, by inspecting the results, that the secondary speaker is often attenuated but not completely removed. In [7], the authors show that binary masks are particularly good at reducing interferences. On the other hand, complex masks appeared as an evolution of binary masks and ratio masks, as a way of estimating, not only the magnitude spectrogram, but the phase too. In our case, the phase has already been estimated by using complex masks in the previous stage and thus we can simplify the task in the second stage by predicting binary values.

Note that this refinement network can run recursively, although we empirically found that applying it once leads to the best results in terms of SDR and a considerable boost in SIR. Further iterations reduce the interferences (at a lesser extent) but at the cost of introducing more distortion.

Pursuing the real applicability of our model, we curated an end-to-end gpu-powered system which can pre-process (from raw audio and video) and isolate the target voice of 10s of recordings in less than 100ms using floating-point 32 precision, and in less than 50 ms using floating-point 16.

Face landmarks: In order to achieve real-time results, we estimate the 3D face landmarks using an optimised version of [8] and register the face landmarks to a frontal view omitting an image warping step. Thanks to the 3D information, we can recover lips motion from side views. Finally, we drop the depth and consider just the first two spatial coordinates in the nodes of the graph.

Audio: Waveforms are re-sampled to 16384 Hz. Then, we compute the STFT with a window size of 1022 and a hop length of 256. This leads to a $512 \times 64\,n$ complex spectrogram where n is the duration of the waveform in seconds. To reduce the computational cost of both training and inference we downsample the spectrogram in the frequency dimension by 2 in Stage 1.

4. Experiments

Experiments are carried out in two different datasets: *Voxceleb2* [3], a dataset of celebrities speaking; and *Acappella* [13], a dataset of solo-singing videos. Both datasets are a collection of YouTube videos. From the unseen-unheard test set of *Voxceleb2* we curated two different subsets. A *wild test set* of 1,000 samples randomly selected. And a *clean test set* of 1,000 samples, from which 500 of them have a high-quality content with the following characteristics: frontal (or almost) point of view, low background noise and good image quality.

The metrics used for comparing results are Source-to-Distortion Ratio (SDR) and Source-to-Interferences Ratio (SIR) [17]. We report both the mean and standard deviation.

Speech separation. We first consider the task of AV speech separation and work with *Voxceleb2* dataset. We use 2s audio excerpts which correspond to 50 video frames from which we extracted their face landmarks. We mix two voice samples from *Voxceleb2* which are normalised with respect to their absolute maximum. This normalisation aims to have two voices which are codominant in the mixture.

A quantitative comparison with respect to state-of-the-art methods is provided in Table 1. Our model outperforms all the previous AV speech separation models. Compared to Visual Voice [5], it achieves a much better SIR and slightly better SDR, both for the wild and clean test sets. In particular, for the clean test set, when the motion cues are more reliable, our model has a much lower standard deviation. Some aspects need to be taken into account: i) The face landmark extractor has been trained with higher quality videos than the ones in *Voxceleb2*. On the contrary, the Visual Voice video network has been trained specifically for *Voxceleb2*. ii) Our visual subnetwork, the graph CNN, has 10 times less parameters than its counterpart in Visual Voice. iii) Apart from motion cues, Visual Voice takes also into account speaker appearance features which are correlated with voice features, and which can be crucial in poor quality videos where lip motion is unreliable.

Singing voice separation. In this second experiment we consider the case of singing voice. We are interested in exploring the applicability of models trained for speech separation in the context of singing voice. Since speech models were trained with two voices and no extra sounds and in *Voxceleb2*, which contains mainly English, we restricted to similar types of mixtures in singing voice with *Acappella*. Table 2 compares the results of models trained directly with samples of singing voice (top block of results in Table 2) versus models trained with speech samples (bottom block). When training our model for singing voice we used just the first stage and a 4-block AV ST-transformer (instead of the 10-block one as in speech). We observe that dedicated models for singing voice perform largely better than models trained for speech. This may be explained to particular differences between a speaking and a singing voice. For example, vowels are much more sustained in singing voice, there is much less coarticulation of consonants with surrounding vowels and vibrato is not present in speech. Moreover, singing voice contains varying pitches covering a wider frequency range.

5. Conclusions

In this work we present a lightweight audio-visual source separation method which can process 10s of recordings in

	# parameters		Wild Test set		Clean Test set	
	Visual Net.	Whole Net.	SDR ↑	SIR ↑	SDR ↑	SIR ↑
Visual Voice Audio-only	–	46.14	7.7	13.6	–	–
The conversation [1]	–	–	8.89	14.8	–	–
Visual Voice Motion-only	9.14	55.28	9.94	17	–	–
Y-Net [13]	1.42	9.7	5.29 ± 5.06	8.45 ± 6.8	5.86 ± 4.78	9.25 ± 6.44
Visual Voice [5]	20.38	77.75	9.92 ± 3.56	16.11 ± 4.8	10.18 ± 3.36	16.49 ± 4.5
Ours	1.42	58.2	**10.03 ± 3.35**	**18.18 ± 4.72**	**10.25 ± 2.61**	**18.65 ±3.8**

Table 1. Evaluation on *Voxceleb2* unheard-unseen test sets. Number of parameters in millions. Results in the first block are taken from the original papers and that is why we only report certain columns. In the second block we report the mean and the standard deviation values.

less than 0.1s in an end-to-end GPU powered manner. Besides, the method shows competitive results to the state-of-the-art in reducing distortions while clearly outperforming in reducing interferences. We show that face landmarks are computationally cheaper alternatives to raw video and help to deal with large-scale datasets. For the first time, we evaluate AV speech separation systems in singing voice, showing empirically that the characteristics of the singing voice differ substantially from the ones of speech.

References

[1] T. Afouras, J. S. Chung, and A. Zisserman. The conversation: Deep audio-visual speech enhancement. In *Interspeech*, 2018. 1, 4

[2] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman. Audio-visual synchronisation in the wild. In *32nd British Machine Vision Conference, BMVC*, 2021. 1

[3] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech*, 2018. 3

[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. 1, 2

[5] R. Gao and K. Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, 2021. 1, 3, 4

[6] E. Z. Golumbic, G. B. Cogan, C. E. Schroeder, and D. Poeppel. Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *The Journal of Neuroscience*, 33:1417 – 1426, 2013. 1

Model	SDR ↑	SIR ↑
Y-Net [13]	11.08 ± 7.51	17.18 ± 9.68
Ours (stage 1, 4 blocks)	**14.85 ± 7.87**	**21.06 ± 9.69**
Ours (stage 1)	3.89 ± 9.28	5.89 ± 11.15
Ours	4.04 ± 10.30	**7.21 ± 13.26**
Visual Voice [5]	**4.52 ± 8.64**	7.03 ± 7.11

Table 2. Singing voice separation. Mixtures of two singers with no additional accompaniment from the test set unseen-unheard (only samples in English) of *Acappella*. Results in top block: models trained directly with samples of singing voice; bottom block: models trained with speech samples.

[7] E. M. Grais, G. Roma, A. J. Simpson, and M. Plumbley. Combining mask estimates for single channel audio source separation using deep neural networks. In *Interspeech*, 2016. 3

[8] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, 2020. 3

[9] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 2

[10] W. J. Ma, X. Zhou, L. A. Ross, J. J. Foxe, and L. C. Parra. Lip-reading aids word recognition most in moderate noise: a bayesian explanation using high-dimensional feature space. *PloS one*, 4(3):e4638, 2009. 1

[11] N. Makishima, M. Ihori, A. Takashima, T. Tanaka, S. Orihashi, and R. Masumura. Audio-visual speech separation using cross-modal correspondence loss. In *ICASSP*, 2021. 1

[12] 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. *IEEE/ACM TASLP*, 29:1368–1396, 2021. 1

[13] J. F. Montesinos, V. S. Kadandale, and G. Haro. A cappella: Audio-visual singing voice separation. In *BMVC*, 2021. 1, 2, 3, 4

[14] 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 *ICASSP*, 2019. 1

[15] T.-D. Truong, C. N. Duong, H. A. Pham, B. Raj, N. Le, K. Luu, et al. The right to talk: An audio-visual transformer approach. In *ICCV*, pages 1105–1114, 2021. 1

[16] E. Tzinis, S. Wisdom, T. Remez, and J. R. Hershey. Improving on-screen sound separation for open-domain videos with audio-visual self-attention. *arXiv preprint arXiv:2106.09669*, 2021. 1

[17] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE/ACM TASLP*, 14(4):1462–1469, 2006. 3

[18] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proc. of the AAAI conf. on artificial intelligence*, 2018. 2

[19] A. Zadeh, T. Ma, S. Poria, and L.-P. Morency. Wildmix dataset and spectro-temporal transformer model for monoaural audio source separation. *arXiv preprint arXiv:1911.09783*, 2019. 1, 2