Synchronisation of Lips and Voices

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

1. Introduction

Online media is brimming with user generated videos involving human voice activity in the form of speech and singing voice. A large number of these videos suffer from misalignment between the audio and visual streams arising due to the lack of appropriate care during video editing. As a result of this misalignment, the video viewers notice that the lip motion is not perfectly synchronised with the voice in the audio. Lip-voice synchronisation in such videos could be corrected by compensating the offset between the audio and visual modalities, which is rendering the video out-of-sync. A commonly used method to determine this offset is by training a model to tell lip-synchronised audio-visual pairs from the non-synchronised pairs and then choose the alignment between the audio and visual signals in the non-synchronised pairs that maximises the synchronisation score. We train our model to discriminate between synchronised and unsynchronised audio-visual pairs by learning their mutual audio-visual correspondence in a self-supervised fashion.

We make several contributions through this paper. We propose a novel audio-visual transformer-based lip-voice synchronisation model that estimates the extent of synchronisation between the lips motion and the voice in a given voice video. Our lip synchronisation model outperforms state-of-the-art models on the speech benchmark dataset Lip Reading Sentences 2 (LRS2) [1]. Besides, we also train and test our lip synchronisation models on the Acappella [8] dataset which contains videos of solo singing performances. We are the first to explore the relevance of lip sync models trained in speech videos to lip sync in singing voice. We use the learned visual features in the lip synchronisation task to outperform a singing voice separation baseline to showcase the practical utility value of our work. Demos are available at https://ipcv.github.io/VocaLiST.

2. Related Work

With the advent of deep learning techniques, different models have addressed the synchronisation of audio-visual signals in a self-supervised way, by automatically creating positive and negative audio-visual sync/out-of-sync pairs. While some works focus on synchronisation in general sounds [11, 2], others are specialised in speech signals [4, 3, 5]. The common trend is to extract features from each modality with an audio and a visual stream and then measure similarity/distance between the two embeddings using a sliding window to infer the offset of synchronisation. The first deep-learning-based model for audio-visual synchronisation in speech [3] uses a contrastive loss with a positive and a negative pair. The use of N negative pairs in a multi-way cross-entropy loss with softmax function [4] further improved the performance of the same model. Unlike these works, [5] directly train the model to determine the offset in the audio-visual pairs.

Transformers have emerged as powerful deep learning architectures capable of capturing long range dependencies in time series. Lately, transformers have been explored for several audio-visual tasks such as source separation [15, 9], source localisation [6] and speech recognition [7], including synchronisation [2]. Our work in this paper is closest to Audio-Visual Synchronisation with Transformers (AVST) [2]. At high-level, both models share the same outline as shown in the left part of Fig. 1. However, the overall architecture of our model is different with regard to the choice of audio and visual encoders and the design of the synchronisation block. The details of our model architecture are clearly outlined in Section 3.1. Besides, [2] use the InfoNCE [10] loss for optimising their model, we use binary cross entropy loss. The main focus in [2] is audio-visual synchronisation in general audio classes, while we focus on lip-synchronisation in speech and singing voice videos only. Finally, unlike in [2], we demonstrate a real-world application of the learned features of our synchronisation model.

3. Method

Architecture. The architecture of our model is shown in Fig. 1. We use a transformer-based classification model which ingests audio and visual features estimated by the audio and visual encoders.

We design a powerful cross-modal audio-visual transformer that can use the audio-visual representations learned in its cross-modal attention modules to deduce the inherent audio-visual correspondence in a synchronised voice and lips motion pair. We refer to our transformer model as VocaLiST, the **Vocal Lip Sync Transformer**. Its design is inspired by the cross-modal transformer from [16]. The cross-modal attention blocks track correlations between signals across modalities. The A→V unit takes in audio features as the query and the visual features as the key and values. The roles of these audio and visual features are swapped in the V→A unit. The output of the A→V unit forms the query to the hybrid fusion transformer unit, while its key and values are sourced from the output of the V→A unit.

Training Setup. We train our model for estimating the audio-visual correspondence score for a given audio-visual pair in an end-to-end manner. The positive examples correspond to the synchronised pairs in which the audio cor-

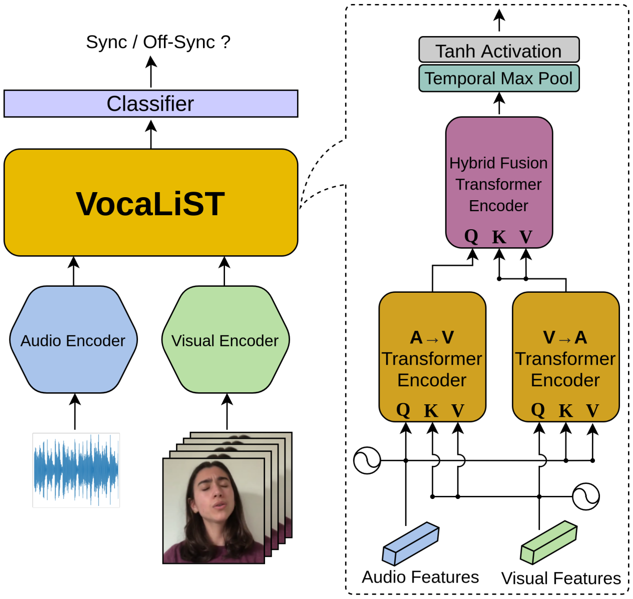

Figure 1. Architecture of our lip synchronisation model.

responds to the visual. The negative examples are obtained by introducing random temporal misalignment between the in-sync audio-visual pairs. This allows us to follow a self-supervised training pipeline. During the training, the positive and the negative examples are sampled equally. The audio-visual correspondence score needs to be maximised for the positive examples and minimised for the negative examples. The binary cross-entropy loss criterion is used to optimise the model parameters during the training. Unless stated otherwise, we train our model with audio-visual input corresponding to a sequence length of 5 visual frames (0.2s) sampled at 25 fps.

4. Experiments

In this section, we investigate the lip synchronisation in two different settings: speech and singing voice. Further, we show the usefulness of the lip synchronisation models for a practical audio-visual application.

4.1. Lip Synchronisation in Speech

Dataset. For the task of lip synchronisation in speech videos, we consider the LRS2 [1] dataset as in the baseline methods SyncNet [3], Perfect Match (PM) [4] and AVST [2]. This allows us to directly compare our method against these baselines. As in the baseline methods, we train our model on the 'pretrain' subset of LRS2 and evaluate on the 'test' subset.

Evaluation protocol. For a fair comparison, we mimic the evaluation protocol followed by the baseline methods. Given a sequence of cropped mouth frames and the mel-spectrograms, our model is tasked with predicting the correct synchronisation in the given pair of inputs. Human observers cannot tell unsynchronised speech videos from the synchronised ones when the temporal offset is in the ±1 frame range. Hence, for a synchronised pair of speech audio and lips motion inputs, the synchronisation is estimated to be correct if the predicted offset between the pair is within ± 1 frame range. As in the baseline models, we estimate this offset by finding the index of maximum audio-visual correspondence between the set of 5-frame visual features and all the audio feature sets (each temporally matching the length of 5 visual frames) lying in ±15 frame range.

Though we mainly train our model on audio-visual input pairs corresponding to the length of 5 visual frames, the model can be tested on inputs of larger lengths. This is particularly useful when a set of 5 visual frames is non-informative (e.g. silence in speech). To evaluate on inputs of larger context window, the audio-visual correspondence score is obtained for each possible 5-frame set within the evaluation window (with a temporal stride of 1 frame) and then averaged before determining the offset. This averaging has been shown to improve the accuracy in the baseline models. On the other hand, [2] showed that training models for specific context window sizes (other than 5) leads to better performance than the previous averaging strategy in the same context window.

Results and discussion. Table 1 shows a direct comparison of the lip synchronisation accuracy of our model against the baseline models on the LRS2 test set. The accuracy is computed for multiple context windows ranging from sizes of 5 to 15. The accuracy improves across all the models as we increase the context window size while the rate of improvement reduces.

For each of the context window sizes, our model outperforms all the listed baseline models on the LRS2 test set. This improved performance can be attributed to the powerful cross-modal transformer blocks in our model in combination with our 3D-CNN based visual encoder. We tried reducing the complexity of our model by eliminating up to 2 cross-modal transformer blocks and also by replacing our visual encoder with the 18-layered Mixed Convolution network [14]. These changes resulted in poorer performance.

4.2. Lip Synchronisation in Singing Voice

Dataset. We use Acappella [8], the only publicly available audio-visual singing voice dataset. It consists of around 46-hours of solo-singing videos spanning four language categories. We test on the unseen-unheard test subset of the dataset which contains 83 song performances evenly distributed across each of the language categories and the genders. All the videos in the Acappella dataset have been manually curated from YouTube with care taken to ensure that the samples appear synchronised to human visual observation. On the other hand, in LRS2, an automatic method was used to better synchronise the audio-visual signal pairs.

Models. We consider two models for the singing voice lip synchronisation: a baseline model and our VocaLiST. Instead of selecting SyncNet [3] as the baseline model, we choose the Lip Sync Expert Discriminator [12]. We refer to this new baseline model as SyncNet*. SyncNet* improves upon its predecessor SyncNet in the following ways: Unlike in SyncNet, i) SyncNet* operates on the RGB images, ii) the model is significantly deeper than the former with residual skip connections, and iii) the model is optimised using a cosine-similarity distance metric in combination with binary cross-entropy loss. We train both VocaLiST and SyncNet* on Acappella training set in an end-to-end manner.

Evaluation protocol. Unlike in speech videos, it is difficult for humans to notice out-of-sync in videos belonging to the general sound categories [2] when the offset between the modalities is less than 5. Singing voice also falls in such general sound category. The presence of sustained vowel sounds in singing voice, makes it difficult to notice synchronisation errors for the offsets less than 5. Thus, we consider synchronisation to be correctly estimated if the model predicts the maximum audio-visual correspondence within ± 5 frame range with respect to the ground truth. Unlike in [2], we do not decode the videos in lower frame rate for evaluation in singing voice.

Results and discussion. We show the results in Table 2. Firstly, we test the models trained on the speech dataset LRS2 directly on the singing voice samples. Among the baseline models discussed in Table 1, only the SyncNet model is publicly available. The Var column indicates if the results on larger context windows are obtained by the model trained with the exact length of the context window as the ones used for testing. Several observations arise from the results. First, the synchronisation accuracy in singing voice is lower than in the case of speech, even with a larger tolerance, showing that singing voice synchronisation is a harder task. Also, the increase of the context window size supposes a larger improvement, compared to the speech case. The best results are achieved with a dedicated network trained for 1s-length window. It can be noticed how our model trained for speech synchronisation generalises quite well for the singing voice with large enough context window sizes. We hypothesise that with larger contexts it is more probable to find portions of singing voice excerpt with non-sustained vowels, or consonants, producing audio-visual cues more similar to the ones found in speech. The LRS2 'pretrain' subset that we use for training spans around 195 hours, while the training set of Acappella totals up to less than 37 hours of videos. Hence, compared to LRS2, Acappella is a small dataset. Therefore, the models trained on LRS2 tend to perform better in Acappella than the other way round.

4.3. Singing Voice Separation

We would like to demonstrate the effectiveness of the features learned by the visual encoder of our synchronisation network by using them in the singing voice separation task. We use the Acappella dataset [8].

Training. We use the same training pipeline that is used for training the model Y-Net-mr in [8]. Y-Net-mr is a U-Net [13] conditioned by visual features extracted from cropped mouth frames using an 18-layer mixed convolution network [14]. Y-Net-mr estimates the complex masks corresponding to a voice present in an audio mixture when given with a spectrogram of an audio mixture and the temporal sequence of the cropped mouth frames corresponding to the target voice as input. During training, half of the input audio mixtures contain one voice mixed with a musical accompaniment, while the other half contains an additional voice besides the accompaniment.

Models. The main baseline models here are Y-Net-gr and Y-Net-mr from [8] which are trained end-to-end for singing voice separation. Y-Net-mr is a state-of-the-art audio-visual singing voice separation model among the models that operate directly on the cropped mouth frames (Y-Net-gr instead works with face landmarks). To investigate the contribution of the features learned by our lip synchronisation model, we propose Y-Net-mr-V. Y-Net-mr-V is nothing but a Y-Net-mr with its visual encoder replaced with that of our lip synchronisation model. In one setting, we train the Y-Net-mr-V for singing voice separation by loading the pretrained weights of the visual encoder which was trained as a part of VocaLiST for singing voice lip synchronisation and keep them frozen throughout the training. In another setting, we train the entire model Y-Net-mr-V in an end-to-end manner without using the pretrained weights learned from the lip synchronisation task. Finally, we also consider Y-Net-mr-S*, which is a Y-Net-mr with its visual encoder replaced by that of the SyncNet* [12]. For Y-Net-mr-S*, we only consider the setting where the pretrained visual encoder weights corresponding to the singing voice lip synchronisation task are loaded from the SyncNet* and frozen as we train the rest of Y-Net-mr-S* for singing voice separation.

Evaluation. We evaluate the singing voice separation performance by evaluating the source separation metrics [17] Source-to-Distortion Ratio (SDR) and Source-to-Interference Ratio (SIR) on the estimated target voices. The higher these metrics are, the better the performance.

Results and Discussion. Table 3 shows the performance of the models in singing voice separation. Our model Y-Net-mr-V outperforms Y-Net-mr when we use the pretrained weights from the lip synchronisation task for the visual encoder. It is challenging to train a source separation network that operates directly with video frames when the dataset is small, which is the case of Acappella. There is a tendency of over-fitting in such small datasets. Under such cir-

Table 1. Accuracy of lip synchronisation models in LRS2

Models	# params	Clip Length in frames (seconds)					
		5 (0.2s)	7 (0.28s)	9 (0.36s)	11 (0.44s)	13 (0.52s)	15 (0.6s)
SyncNet [3]	13.6M	75.8	82.3	87.6	91.8	94.5	96.1
PM [4]	13.6M	88.1	93.8	96.4	97.9	98.7	99.1
AVST [2]	42.4M	92.0	95.5	97.7	98.8	99.3	99.6
VocaLiST	80.1M	**92.8**	**96.7**	**98.4**	**99.3**	**99.6**	**99.8**

(M = million)

cumstances, knowledge transfer from the audio-visual synchronisation could guide the source separation despite the dataset size limitations. To further highlight the importance of this knowledge transfer, we also train the Y-Net-mr-V end to end without using the pretrained synhronisation visual features. In this case, the model does not generalise well to the test-unseen subset despite having 38M trainable parameters. Note that even Y-Net-mr-S* outperforms Y-Net-mr here. But since SyncNet* didn't perform as good as the VocaLisT in lip synchronisation task (see Table 2), as we expected, Y-Net-mr-S* did not outperform Y-Net-mr-V trained with knowledge transfer. Finally, Y-Net-mr-V achieves comparable results to the state-of-the-art model, Y-Net-gr: The metrics of Y-Net-mr-V are not statistically significant w.r.t. the results of Y-Net-gr ($p > 0.05$).

Table 2. Accuracy of lip synchronisation in Acappella dataset

Models	Var	Trained on	Clip Length in frames (seconds)				
			5 (0.2s)	10 (0.4s)	15 (0.6s)	20 (0.8s)	25 (1s)
SyncNet*	N	Acappella	57.7	63.9	69.9	75.1	78.7
SyncNet*	Y	Acappella	57.7	65.9	—	—	73.6
VocaLiST	N	LRS2	56.7	65.1	72.2	77.2	81.2
VocaLiST	N	Acappella	58.8	65.4	71.6	76.5	80.5
VocaLiST	Y	Acappella	**58.8**	**66.4**	—	—	**85.2**

5. Conclusions

This paper presents VocaLiST, a transformer-based model for voice-lip synchronisation. The model has been analysed both in speech and singing voice, producing state-of-the-art results. We have shown that it learns powerful visual features that are useful for solving the problem of singing voice separation in a mixture with more than one voice. It could perform even better for larger input contexts if we have dedicated models trained for each specific length of the input, as shown in Table 2 and also in [2].

Table 3. Performance metrics for Singing Voice Separation. Only the results that are not statistically significant w.r.t. the results of Y-Net-gr ($p > 0.05$) are dotted.

Architecture	Method	Source Separation Metrics	
		SDR	SIR
Y-Net-gr [8]	E2E	**6.41**	**17.38**
Y-Net-mr [8]	E2E	5.03	15.80
Y-Net-mr-V	E2E	1.14	11.72
Y-Net-mr-S*	PT - SyncNet*	5.44	16.17
Y-Net-mr-V	PT - VocaLiST	**6.32**	**17.08**

References

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. *IEEE trans. on Pattern Analysis and Machine Intelligence*, 2018. 1, 2

[2] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman. Audio-visual synchronisation in the wild. In *BMVC*, 2021. 1, 2, 3, 4

[3] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 1, 2, 3, 4

[4] S.-W. Chung, J. S. Chung, and H.-G. Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. 1, 2, 4

[5] Y. J. Kim, H. S. Heo, S.-W. Chung, and B.-J. Lee. End-to-end lip synchronisation based on pattern classification. In *IEEE Spoken Language Technology Workshop*, 2021. 1

[6] Y.-B. Lin and Y.-C. F. Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Proc. of the Asian Conf. on Computer Vision*, 2020. 1

[7] P. Ma, S. Petridis, and M. Pantic. End-to-end audio-visual speech recognition with conformers. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. 1

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

[9] J. F. Montesinos, V. S. Kadandale, and G. Haro. Vovit: Low latency graph-based audio-visual voice separation transformer. *arXiv preprint arXiv:2203.04099*, 2022. 1

[10] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1

[11] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1

[12] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proc. of the 28th ACM Int. Conf. on Multimedia*, pages 484–492, 2020. 3

[13] 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*, pages 234–241. Springer, 2015. 3

[14] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 2, 3

[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*, page 1105–1114, 2021. 1

[16] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Assoc. Computational Linguistics Meet.*, volume 2019, page 6558–6569, 2019. 1

[17] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech, and Language Process.*, 14(4), 2006. 3