# VocaLiST: An Audio-Visual Synchronisation Model for Lips and Voices

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

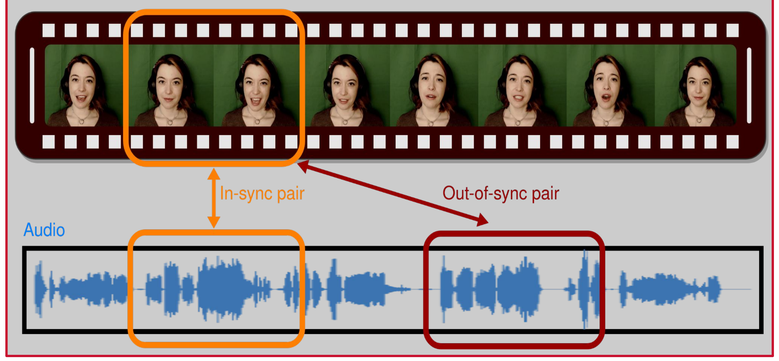
Universitat Pompeu Fabra Barcelona

INTERSPEECH 2022
Sep 18 - 22 • Incheon Korea

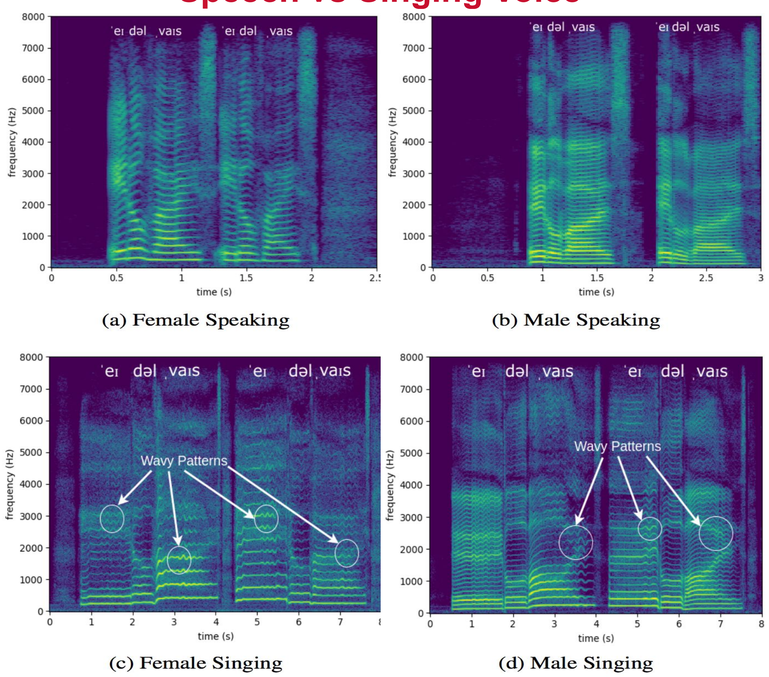
## Abstract

We address the problem of lip-voice synchronisation in videos containing human face and voice. Our approach is based on determining if the lips motion and the voice in a video are synchronised or not, depending on their audio-visual correspondence score. We propose an audio-visual cross-modal transformer-based model that outperforms several baseline models in the audio-visual synchronisation task on the standard lip-reading speech benchmark dataset LRS2. While the existing methods focus mainly on the lip synchronisation in speech videos, we also consider the special case of singing voice. Singing voice is a more challenging use case for synchronisation due to sustained vowel sounds. We also investigate the relevance of lip synchronisation models trained on speech datasets in the context of singing voice. Finally, we use the visual features extracted by the pre-trained visual encoder of the lip synchronisation model in the singing voice separation task to outperform a baseline audio-visual model which was trained end-to-end.

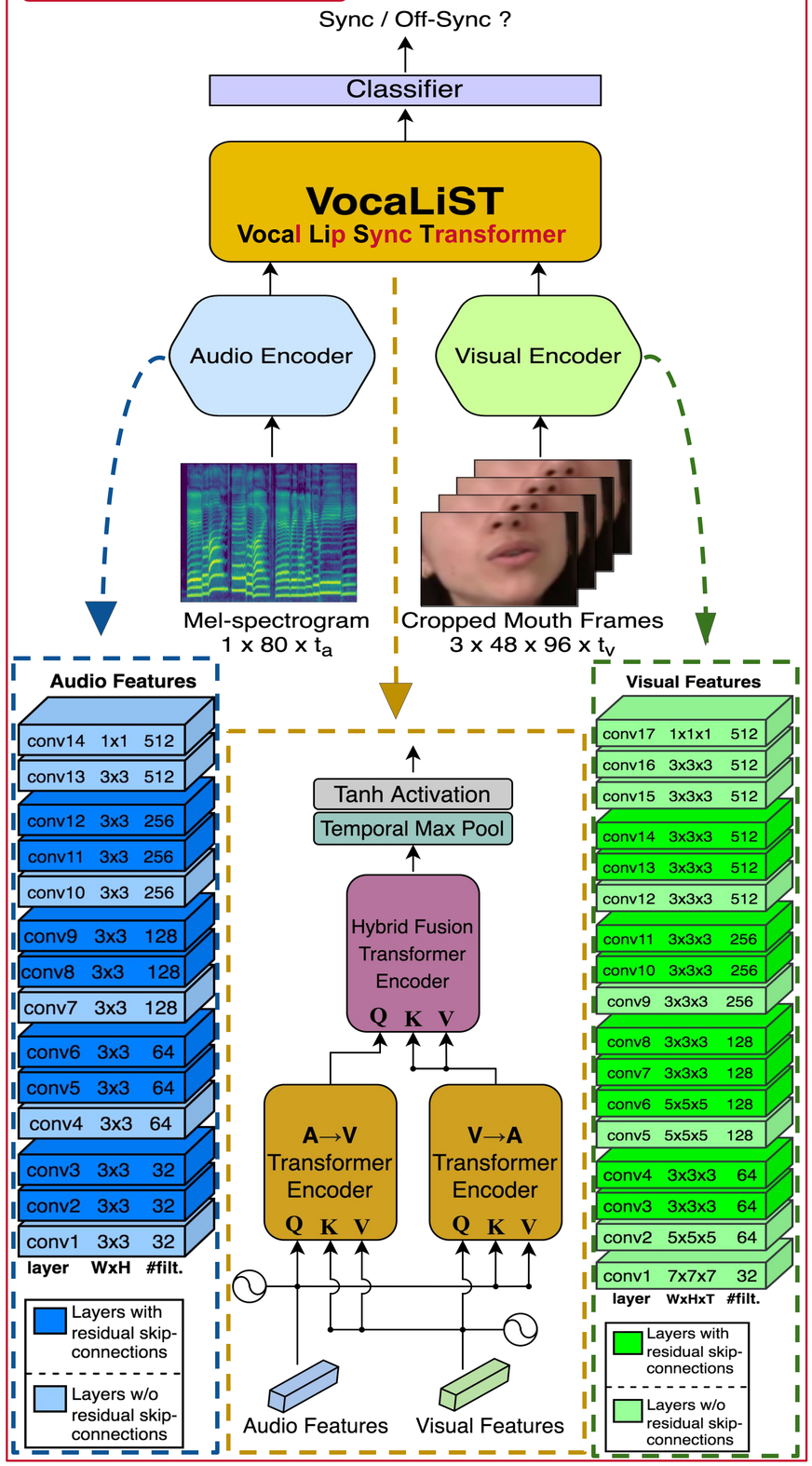
## Goal




## Speech vs Singing Voice




## Model Architecture




## Results

### Speech

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 | 13.6M | 75.8 | 82.3 | 87.6 | 91.8 | 94.5 | 96.1 |
| PM | 13.6M | 88.1 | 93.8 | 96.4 | 97.9 | 98.7 | 99.1 |
| AVST | 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** |

SyncNet: J.S. Chung, A. Zisserman. Out of time: automated lip sync in the wild. ACCV, 2016.
PM: S.W. Chung, J.S. Chung, H.-G. Kang. Perfect match: Improved cross-modal embeddings for audio-visual sync. ICASSP, 2019.
AVST: Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman. Audio-visual synchronisation in the wild. BMVC, 2021.
LRS2: T. Afouras, J.S. Chung, A. Senior, O. Vinyals, A. Zisserman. Deep audio-visual speech recognition. IEEE PAMI, 2018.

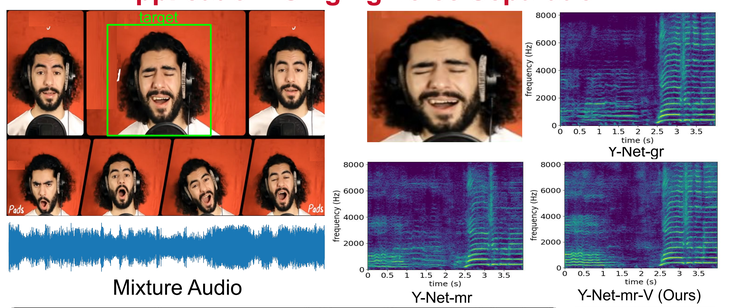
### Singing Voice

Accuracy of lip sync. 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** |

SyncNet*: K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, C. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. ACM Multimedia, 2020.
Acappella: J. F. Montesinos, V. S. Kadandale, G. Haro. A cappella: Audio-visual singing voice separation. BMVC, 2021.

### Application: Singing Voice Separation




| Architecture | Method | Source Separation Metrics | |
|---|---|---|---|
| | | SDR | SIR |
| Y-Net-mr | 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** |

Y-Net: J. F. Montesinos, V. S. Kadandale, and G. Haro. A cappella: Audio-visual singing voice separation. In BMVC, 2021.

SCAN ME

**Demos available!**

## Conclusion

- We propose **a new audio-visual transformer-based lip-voice synchronisation model** that detects synchronisation between the lips motion and the voice in a given voice video.
- The model produces **state-of-the-art results** both in **speech and singing voice**.
- Lip sync. in **singing voice is harder than speech** due to sustained vowels and it needs larger context windows.
- The model **learns powerful visual features** that are useful for singing voice separation.