ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

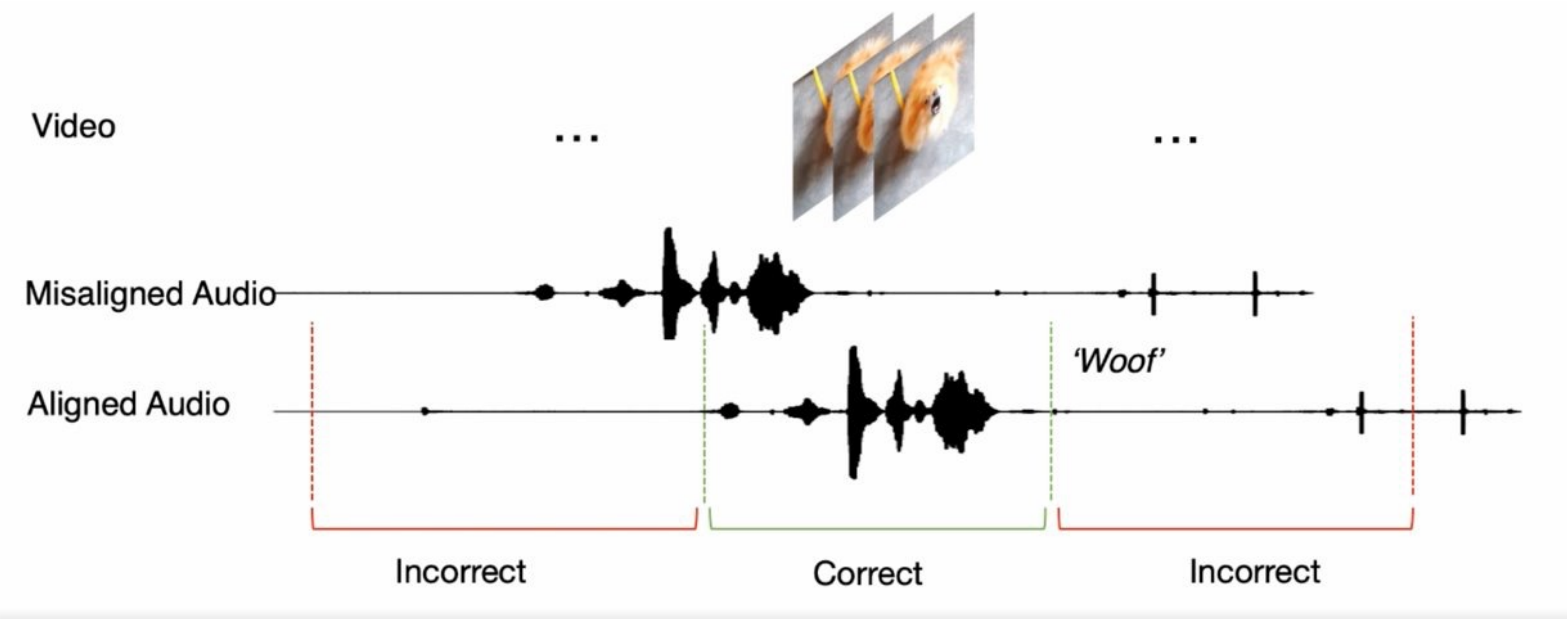
# ModEFormer: Modality-Preserving Embedding for Audio-Video Synchronization using Transformers

**Akash Gupta**
aksg@nyu.edu
New York University

**Rohun Tripathi**
rt443@cornell.edu
Amazon Studios

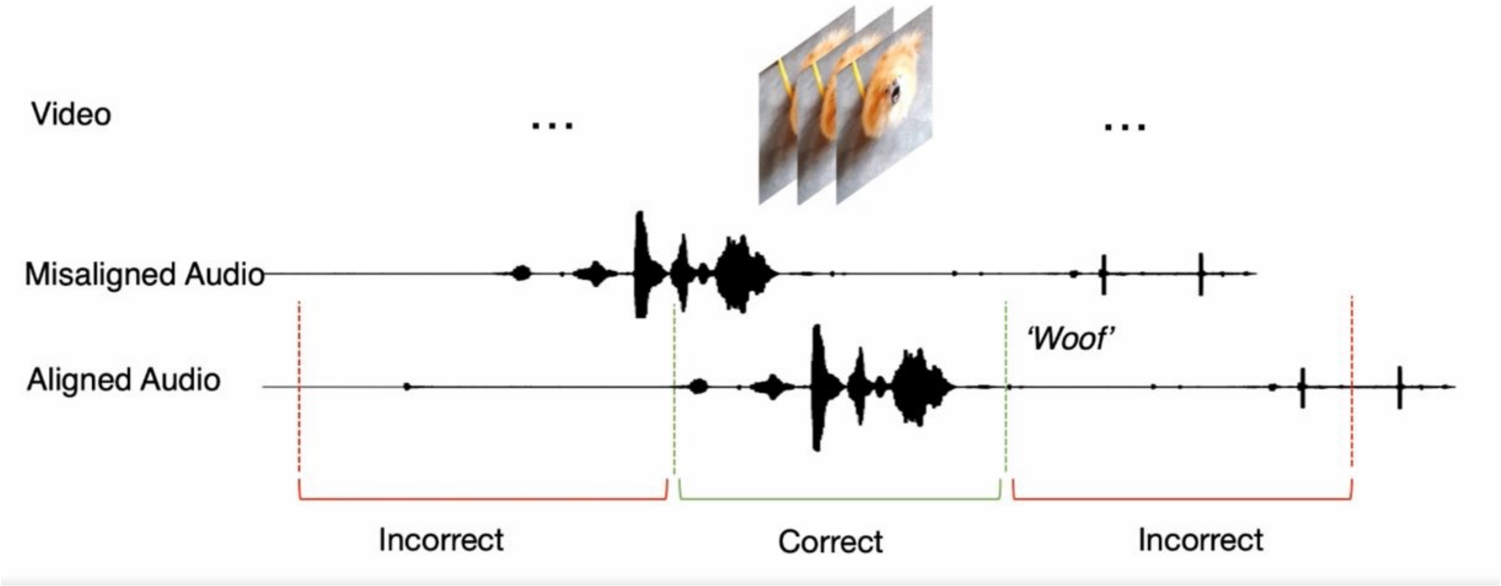
**Wondong Jang**
dotol1216@gmail.com
Amazon Studios

# Audio-video synchronization in videos



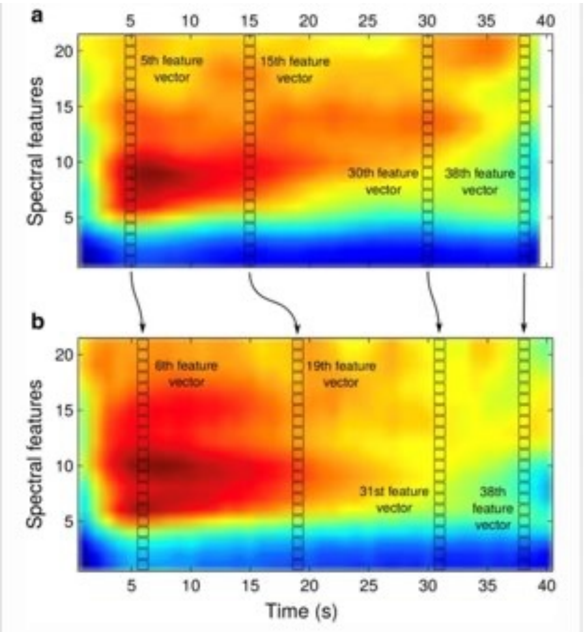

*Gupta et al. ModEFormer: Modality-Preserving Embedding for Audio-Video Synchronization using Transformers. ICASSP 2023*

# Audio-video synchronization in videos




*These errors occur when the audio and video components of a video are not synchronized properly, leading to a poor viewing experience.*

*__Time warping__*



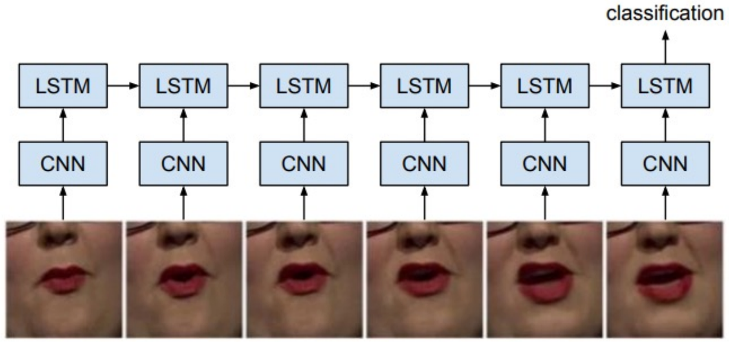

*Requires manual supervision to align audio with the video but it is time consuming and prone to human errors*

# Motivation

- An automated off-sync detector can help identify these errors and provide a more accurate synchronization between audio and video.
- Additionally, an off-sync detector can help video creators save time and resources by automating the process of detecting and correcting these errors.
- Some other practical applications -

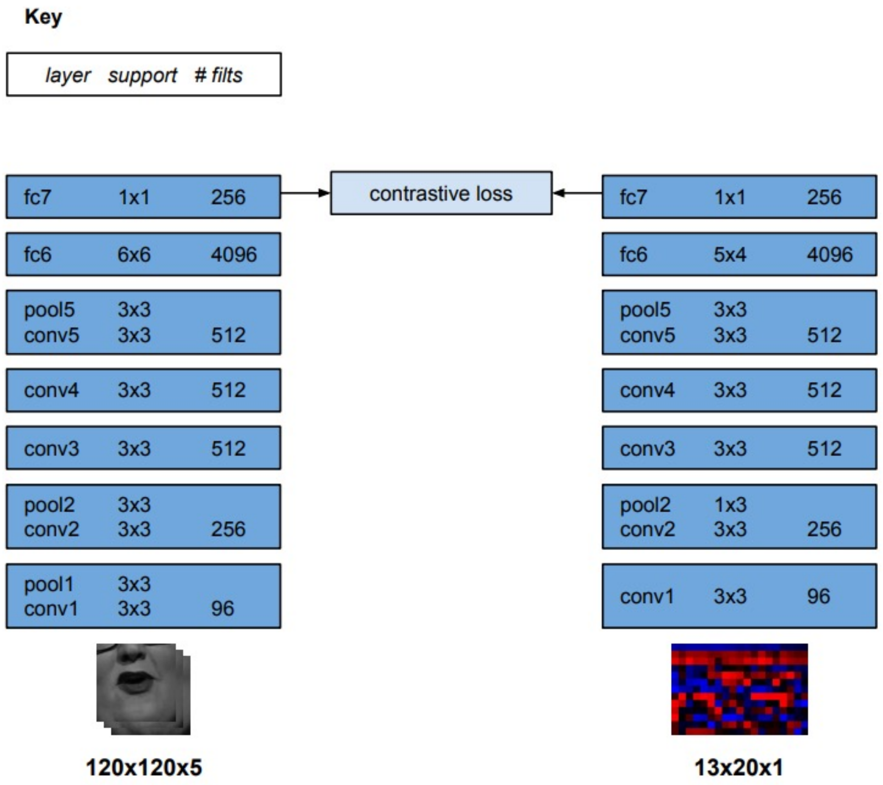
1. Active speaker detection

2. Lip reading

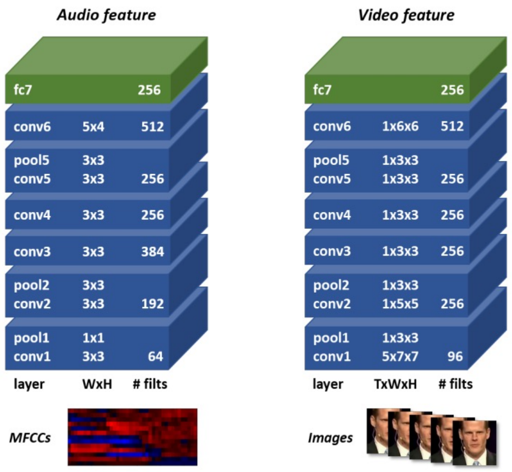



*Gupta et al. ModEFormer: Modality-Preserving Embedding for Audio-Video Synchronization using Transformers. ICASSP 2023*

# Previous approaches



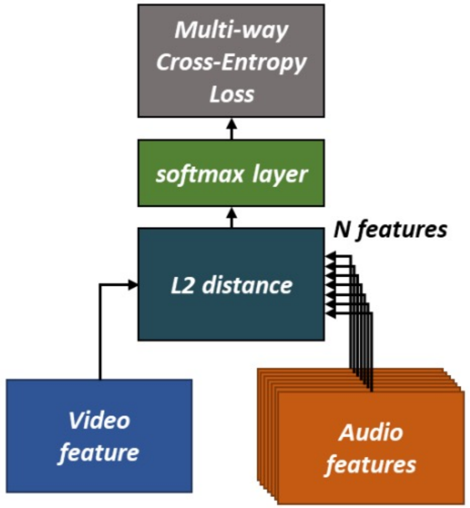

**SyncNet** [Chung et al. 2016] – *ConvNet Siamese style architecture trained with a Euclidean distance contrastive loss for off-sync detection*



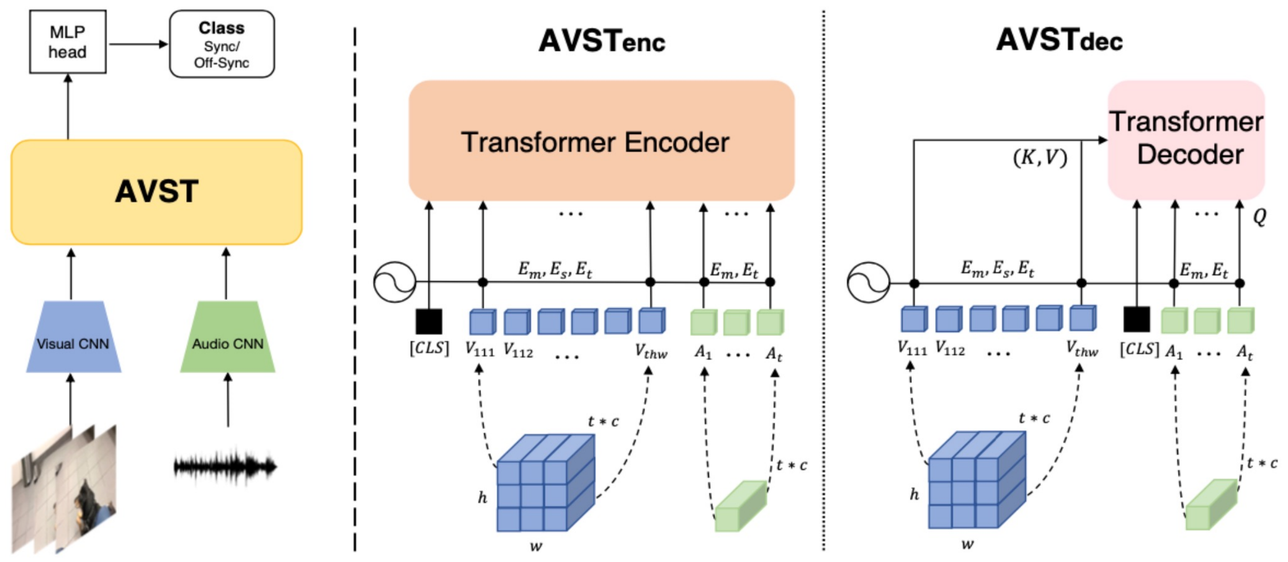

(a) Audio stream (b) Visual stream

**Perfect Match** [Chung et al. 2019] – *Introduces a 3D-Conv based image encoder to include RGB images from the video stream*

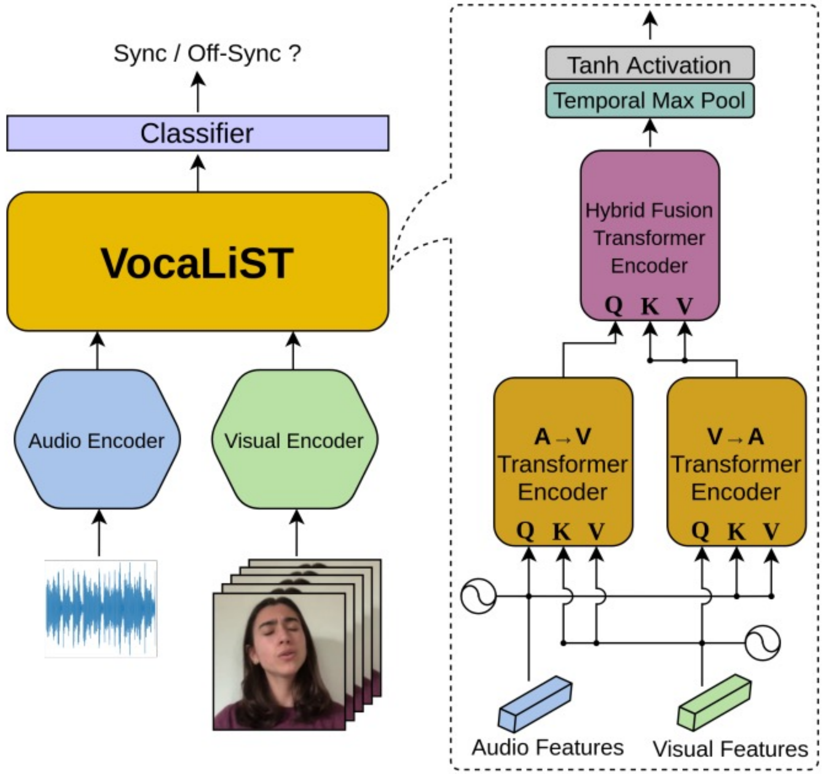
*A multi-way cross entropy loss is used to process a batch of 1 video feature, 1 positive audio feature and N-1 negative features and performs multi-class classification*



# Previous approaches



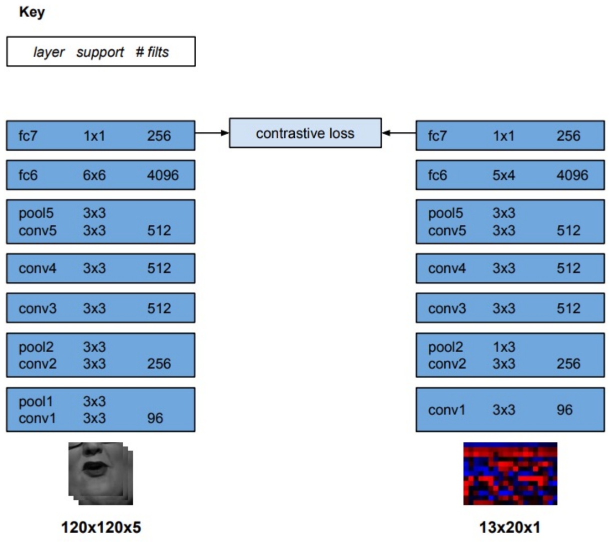

**AVST** [Chen et al. 2021] - *Introduced attention to learn correlation between longer audio and video sequences as informative portions can be localized in a short subsequence.*

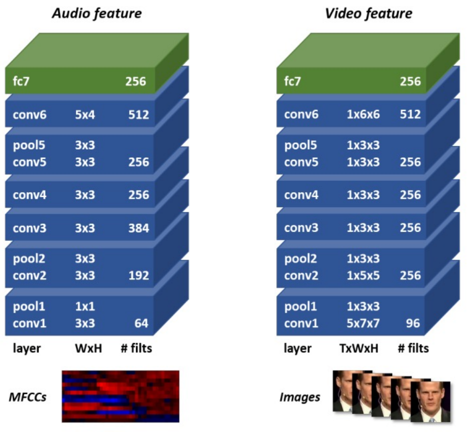



**VocaLiST** [Kadandale et al. 2022] – *Multiple cross modal transformers thereby learning audio-video, video-audio and hybrid correlations*

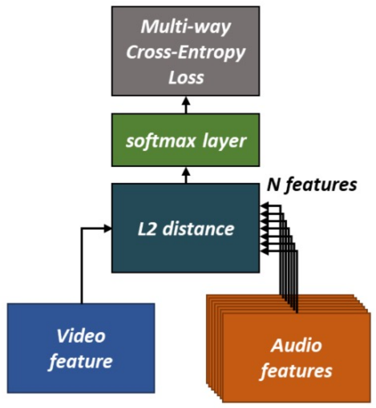
# Previous approaches

## CNN-based



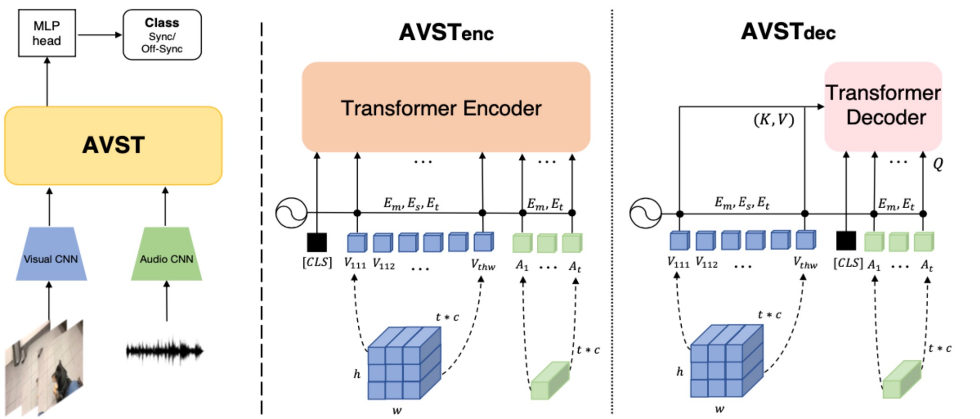



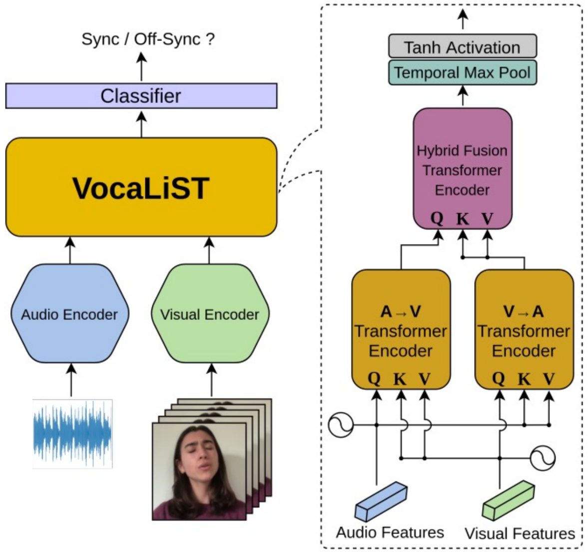

(a) Audio stream (b) Visual stream




**State-of-the-art contrastive learning techniques require large batch size with abundant negative samples for learning good modality-specific features**

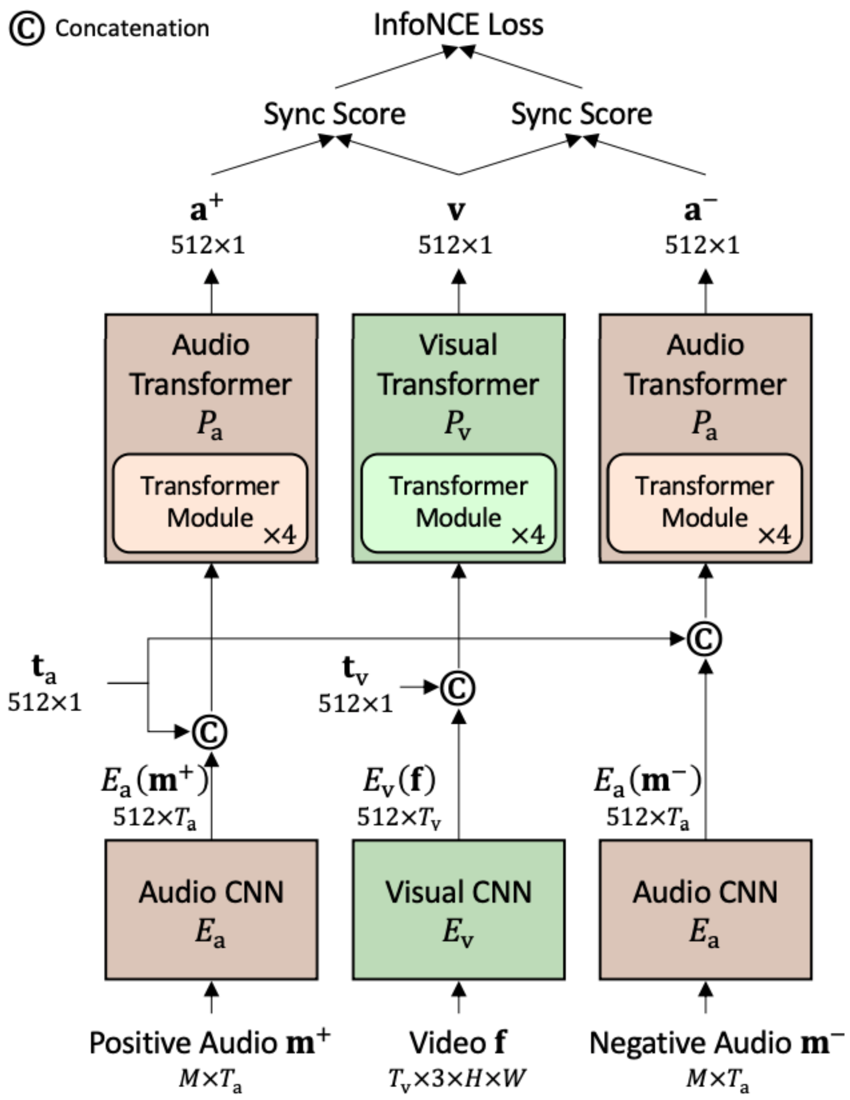
## Self and cross attention



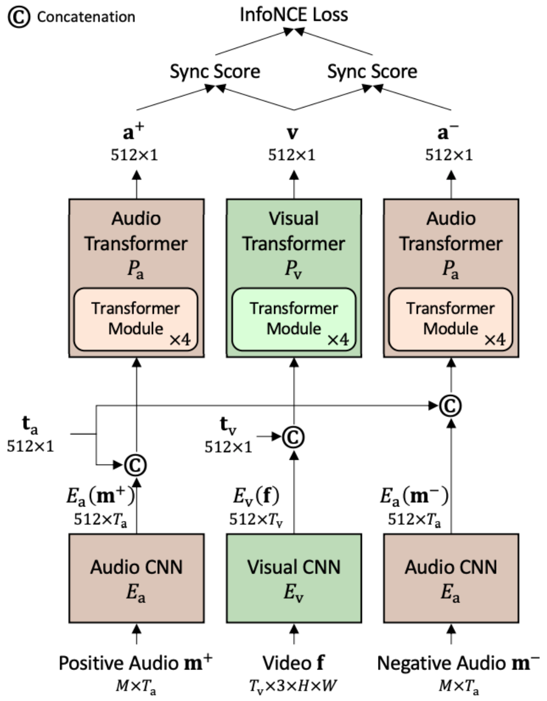

*Gupta et al. ModEFormer: Modality-Preserving Embedding for Audio-Video Synchronization using Transformers. ICASSP 2023*

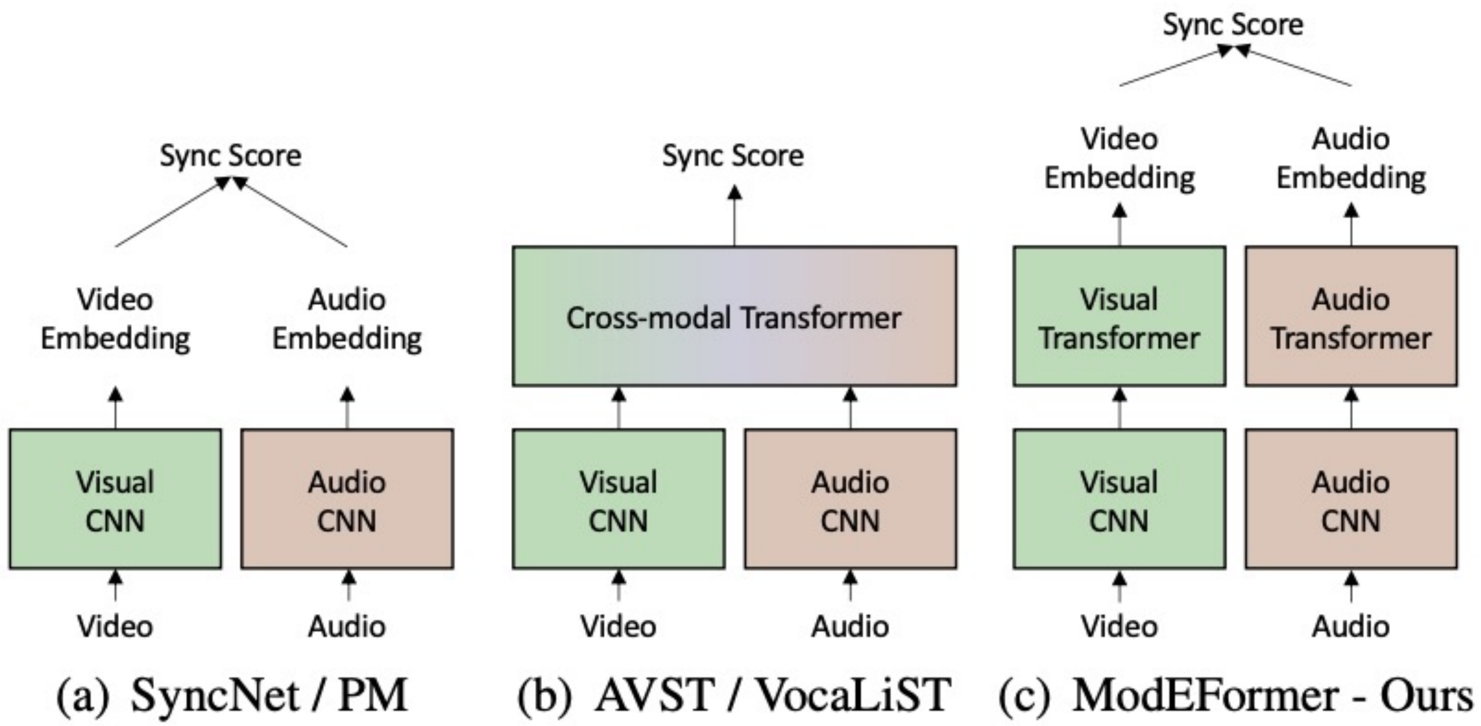
# ModEFormer: Modality-Preserving Embedding for Audio-Video Synchronization using Transformers




- **ModEFormer** has separate encoders for audio and video modalities and extracts the corresponding embeddings
- Video branch takes a sequence of RGB frames while the audio branch takes a fixed size crop from the mel-spectrogram.
- Each modality branch contains a CNN encoder to extract intermediate representations
- The above representations are concatenated with sinusoidal positional encodings and are passed to modality-specific transformers.

# ModEFormer: Modality-Preserving Embedding for Audio-Video Synchronization using Transformers

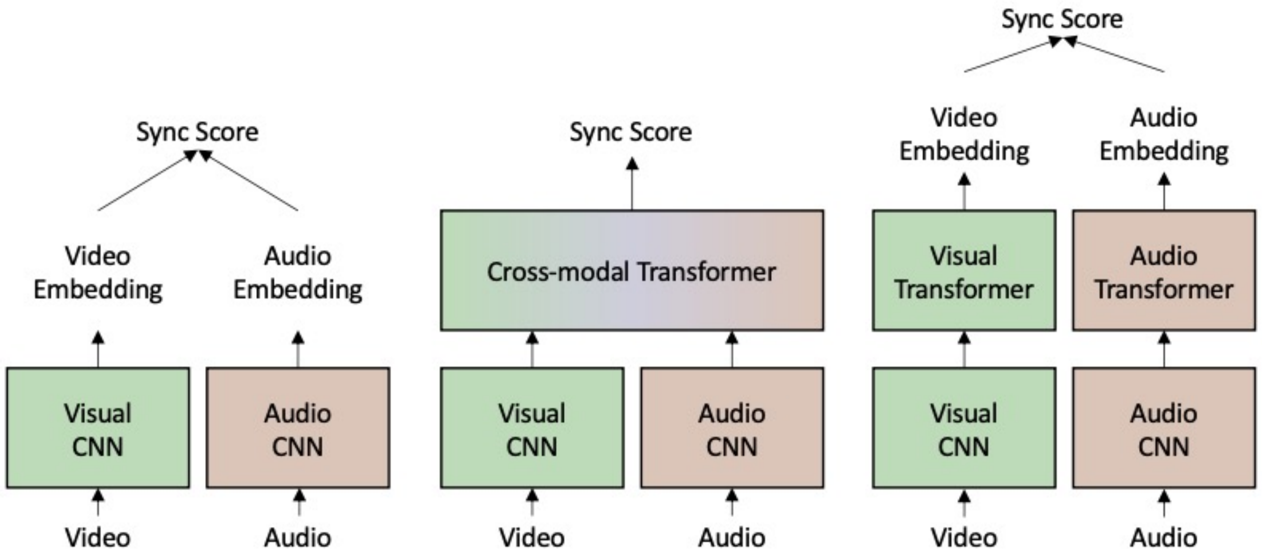



- Unlike previous approaches, we ensure no mixing between modalities at any step.
- We take the learned [CLS] token representation from the transformer encoder as the final embedding for each modality.
- To enable contrastive learning, each video modality is paired up with a bunch of audio samples illustrating positive and negative examples.

*Gupta et al. ModEFormer: Modality-Preserving Embedding for Audio-Video Synchronization using Transformers. ICASSP 2023*

# ModEFormer: Modality-Preserving Embedding for Audio-Video Synchronization using Transformers




(a) SyncNet / PM (b) AVST / VocaLiST (c) ModEFormer - Ours

$$L = -\frac{1}{B} \sum_{\mathbf{v},\mathbf{a}^+ \in \mathcal{P}} \log \frac{e^{(\phi(\mathbf{v},\mathbf{a}^+)/\tau)}}{\sum_{\mathbf{a} \in \mathcal{N}(\mathbf{v})} e^{(\phi(\mathbf{v},\mathbf{a})/\tau)}},$$

InfoNCE loss function

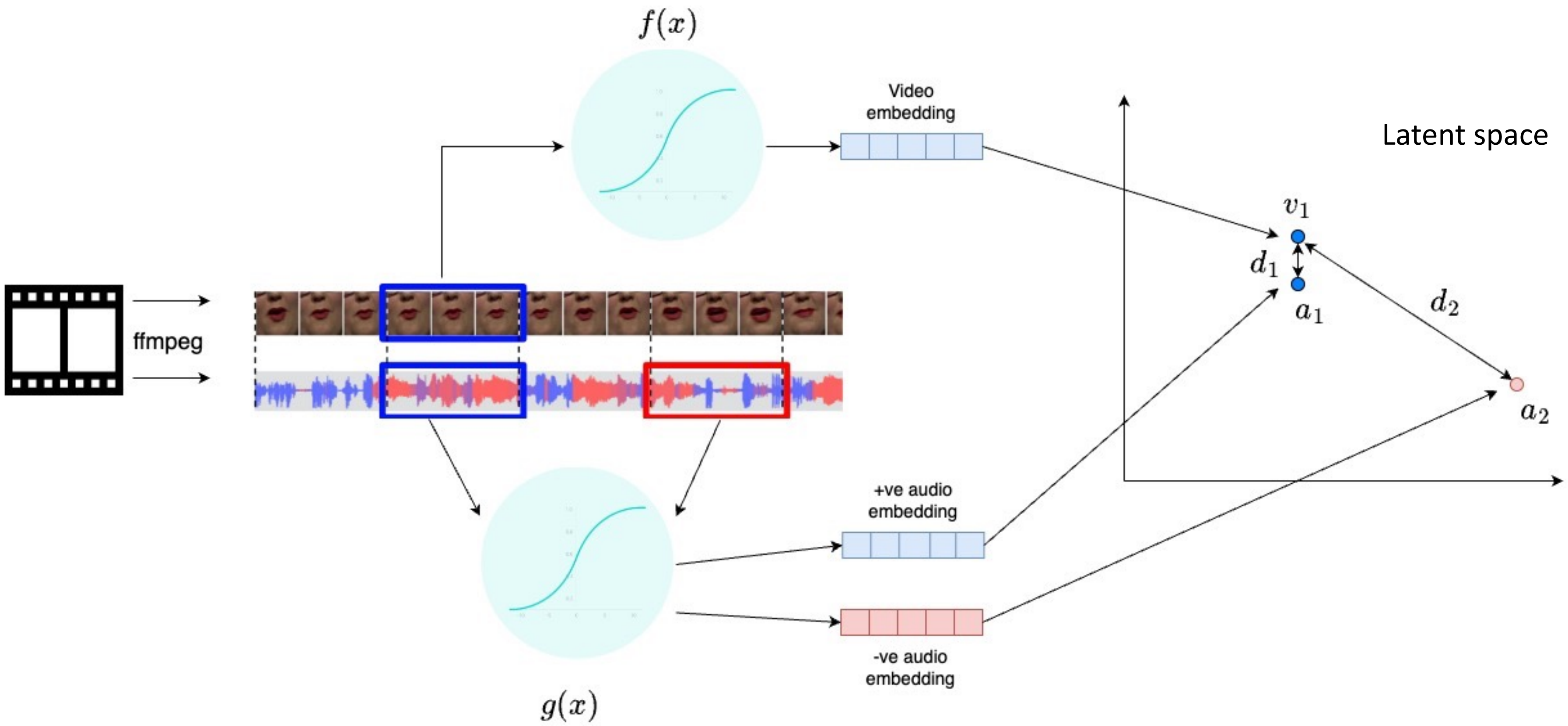
$$\phi(\mathbf{v}, \mathbf{a}) = \frac{\mathbf{v}}{|\mathbf{v}|} \cdot \frac{\mathbf{a}}{|\mathbf{a}|}.$$

Cosine similarity to calculate sync score

- Unlike previous approaches, we ensure no mixing between modalities at any step.
- We take the learned [CLS] token representation as the final embedding for each modality.
- To enable contrastive learning, each video modality is paired up with a bunch of audio samples illustrating positive and negative examples
- We calculate a sync score and use InfoNCE loss minimization which offers better generalization allowing to learn discriminative and noise-invariant features

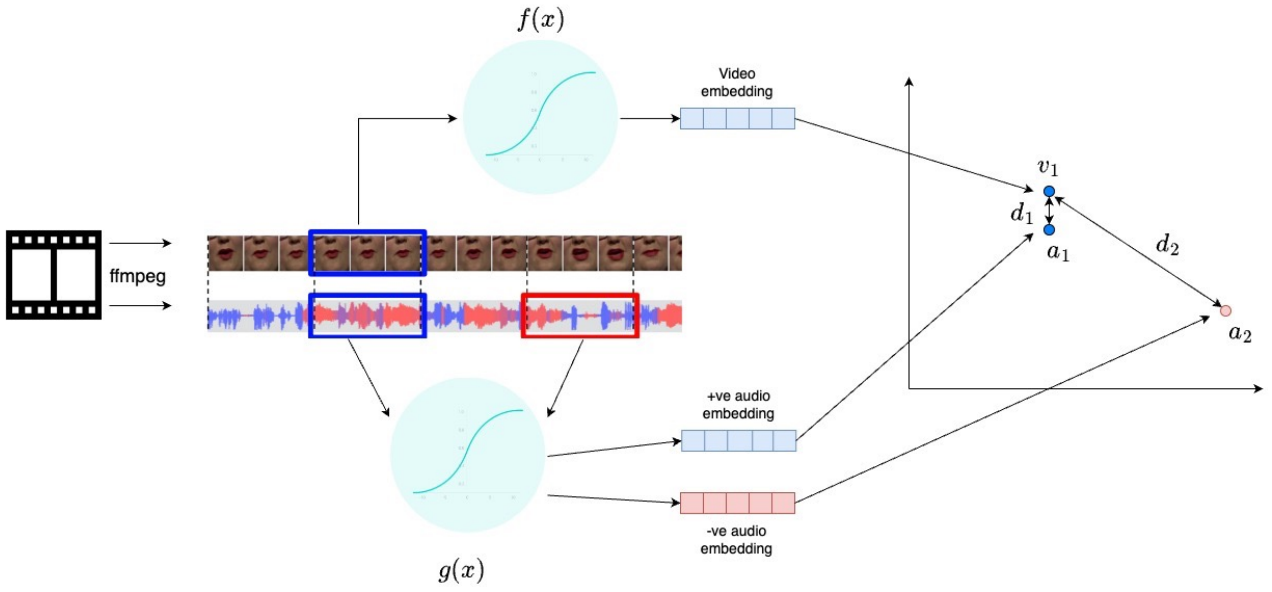
# Audio-Video Contrastive learning

- Push aligned audio-video latent representations closer to each other and misaligned latent representations far apart.

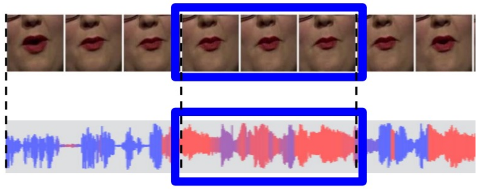


# Audio-Video Contrastive learning

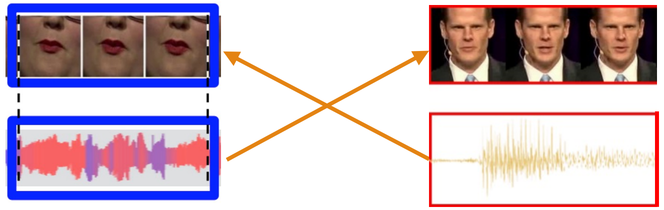
- Push similar (positive) latent representations closer to each other and dissimilar (negative) latent representations far apart.




- Sampling strategy -
  - Positives – Audio and video are temporally aligned coming from the same clip.



  - Easy negatives – Audio and video coming from a different clip.



  - Hard negatives – Audio and video from the same clip but temporally shifted

# Experimental setup

- **ModEFormer training** – Carried out in two stages

  - Stage 1 – Here we take a large batch size of 2000 and where each batch entry is from a unique clip and has two corresponding hard negative audio samples
  - Stage 2 – Here we increase the number of hard negatives and also start incorporating easy negatives in the batch.
  - We develop such paradigm to obtain benefits of large batch size from contrastive learning (stage 1) and also efficiently incorporate diversity in training samples for better generalization (stage 2)

- **Datasets used** – We used Lip reading sentences (LRS) datasets

  - LRS2 - Contains thousands of spoken sentences from BBC television with a length of upto 100 characters.

| Set | Dates | # utterances | # word instances | Vocab |
|---|---|---|---|---|
| Pre-train | 11/2010-06/2016 | 96,318 | 2,064,118 | 41,427 |
| Train | 11/2010-06/2016 | 45,839 | 329,180 | 17,660 |
| Validation | 06/2016-09/2016 | 1,082 | 7,866 | 1,984 |
| Test | 09/2016-03/2017 | 1,243 | 6,663 | 1,698 |

  - LRS3 - Contains thousands of spoken sentences from TED and TEDx videos. We created the val set by randomly slicing the 40% of the “Trainval” partition.

| Set | # videos | # utterances | # word instances | Vocab |
|---|---|---|---|---|
| Pre-train | 5,090 | 118,516 | 3.9M | 51k |
| Trainval | 4,004 | 31,982 | 358k | 17k |
| Test | 412 | 1,321 | 10k | 2k |

# Results

- We use lip-synchronization accuracy as defined by previous approaches on different input video clip lengths to compare the performance of ModEFormer on the LRS test datasets.

| Dataset | Model | Var | Clip Length in Frames (Seconds) | | | | | | # of params |
|---|---|---|---|---|---|---|---|---|---|
| | | | 5 (0.2s) | 7 (0.28s) | 9 (0.36s) | 11 (0.44s) | 13 (0.52s) | 15 (0.6s) | (M=Millions) |
| LRS2 | AVST[3] | ✓ | 91.9 | 97.0 | 98.8 | 99.6 | 99.8 | 99.9 | 42.4M |
| | SyncNet[1] | | 75.8 | 82.3 | 87.6 | 91.8 | 94.5 | 96.1 | 13.6M |
| | PM[2] | | 88.1 | 93.8 | 96.4 | 97.9 | 98.7 | 99.1 | 13.6M |
| | VocaLiST[4] | | 92.8 | 96.7 | 98.4 | **99.3** | 99.6 | **99.8** | 80.1M |
| | ModEFormer - Ours | | **94.5** | **97.1** | **98.5** | **99.3** | **99.7** | **99.8** | 59.0M |
| LRS3 | AVST[3] | ✓ | 77.3 | 88.0 | 93.3 | 96.4 | 97.8 | 98.6 | 42.4M |
| | ModEFormer - Ours | | **90.9** | **93.1** | **96.0** | **97.7** | **98.7** | **99.2** | 59.0M |

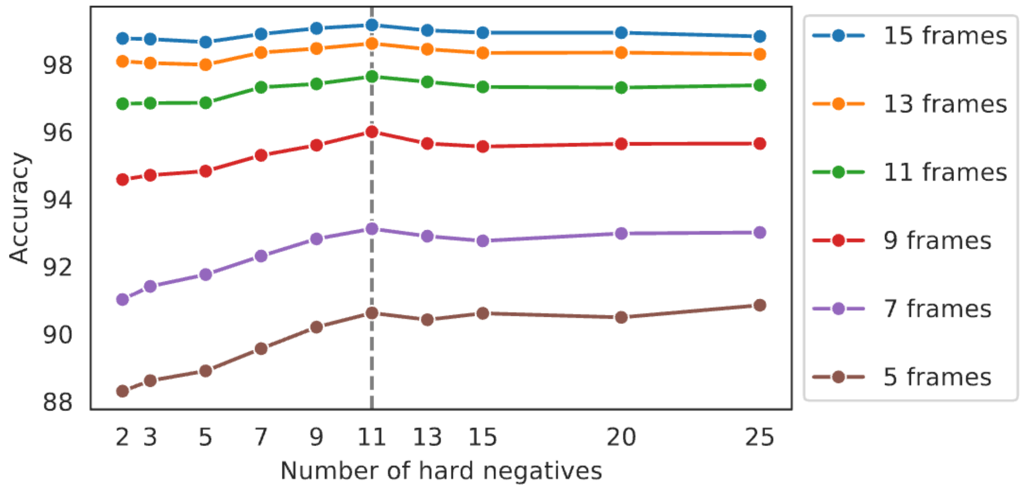
- ModEFormer outperforms all the previous approaches using a fixed number of input frames.
- The significant increase in performance is due to the modality-preserving architecture and the novel sampling strategy including multiple hard negatives during training.
- Since AVST has seen clips of variable length input during training, it cannot be compared with other approaches

# Ablation Study

- Architectural ablation – We study the effect of using transformers in addition to the CNN encoders for each modality branch
- We build a 3D-SyncNet architecture by removing the transformer encoders in each branch and train with the same InfoNCE loss and sampling strategy
- On the LRS3 test dataset we see a remarkable increase in the accuracy of 8.1%

**Table 2.** Results of 3D-SyncNet and ModEFormer on LRS3 test set

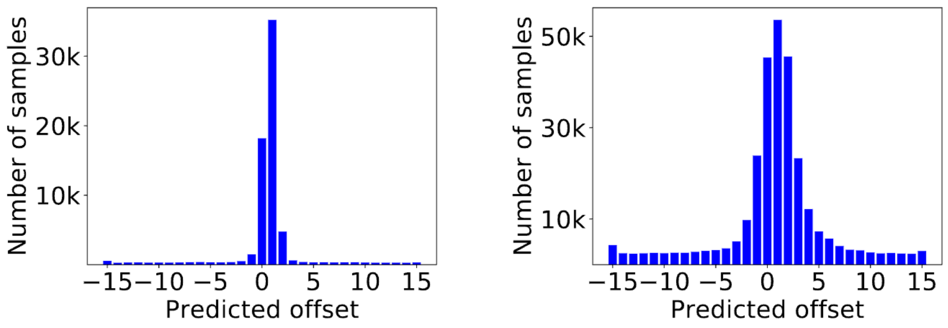
| | 3D-SyncNet | ModEFormer (1st stage) | ModEFormer (2nd stage) |
|---|---|---|---|
| Accuracy | 80.2% | 88.3% | 90.9% |




- Negative sampling strategy – We also experiment to find the optimal number of hard negatives between 2 to 25 to be used during training.
- The overall lip-sync accuracy peaks when the number of hard negatives is 11.
- We see a further increase of 2.6% in second stage training that validates the benefit of our negative sampling strategy.

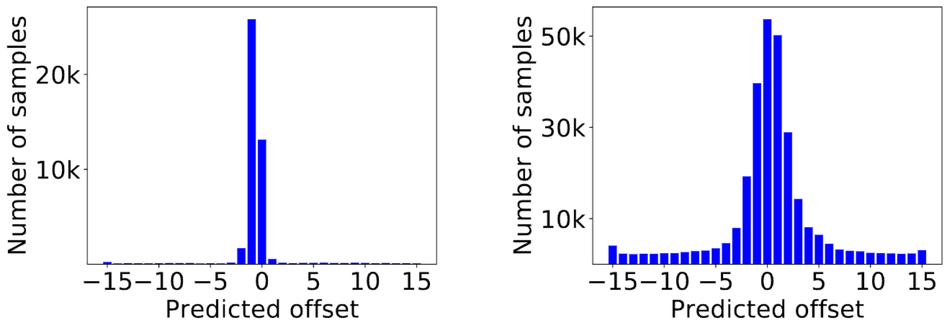
# Applications

- ➢ <u>Offset detection</u> – We apply a trained ModEFormer to detect any audio-video lag in a given test clip
- ➢ For a given clip, we compute cosine similarities at every video frame for audio windows in its neighborhood
- ➢ We identify the predicted offset as the audio window with highest cosine similarity and generate the histogram.
- ➢ Using this analysis, we found that LRS2 and LRS3 are out-of-sync by one frame using a third out-of-distribution dataset, VoxCeleb2.




(a) (b)

Trained on LRS2 and tested on (a) LRS3, (b) VoxCeleb2




(c) (d)

Trained on LRS3 and tested on (c) LRS2, (d) VoxCeleb2

ICASSP 2023
4 - 10 JUNE,
RHODES ISLAND, GREECE

# Thank you for your attention!