# Dynamic Temporal Alignment of Speech to Lips

**Tavi Halperin*** **Ariel Ephrat*** **Shmuel Peleg**

***The Hebrew University of Jerusalem***

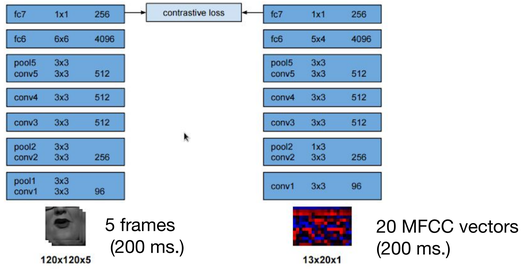
## Automatic Dialogue Replacement (ADR)

- In film post-production, when the quality of speech recorded on set is too low, the actor re-records the speech in the studio.
- Currently, the actor repeats the lines until the speech is perfectly synced to the lips movements in the video.
- If the quality of the on set audio is good the two tracks can be synced automatically using audio-to-audio methods [2].

⇒ **We propose an audio-to-video method for ADR.**

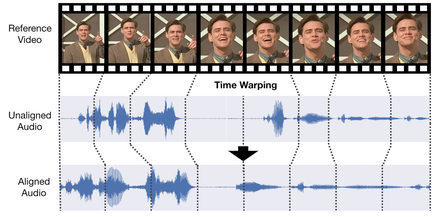
## SyncNet deep features

- Pretrained deep features from [1].
- Shared embedding space for short video clips and audio sequences containing visual speech.




## Our method

- **Dynamic time warp (DTW) of unaligned auditory features to reference visual features.**



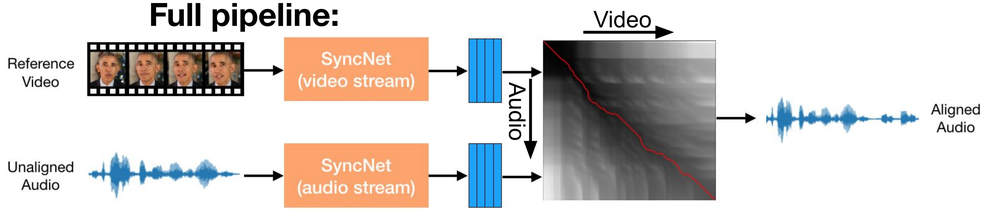

  - Alignment cost uses syncnet embedding
  - All available audio & video combinations are used
  - Cost = $\min\{D(A_i,V_j), D(V_i,A_j), D(A_i,A_j), D(V_i,V_j)\}$

Audio (studio) Video (set) Video (studio) Audio (set)

- Since an earlier voice (wrt the video frames) is much more disturbing than a delayed voice, we replace the cost function $C_{i,j}$ by:

$$\tfrac{1}{2}*C_{i,j} + \tfrac{1}{4}*C_{i-1,j} + \tfrac{1}{4}*C_{i-2,j}$$

**Full pipeline:**




## Results

### Experimental setup

- Data collection
  - Recorded every subject twice
  - Computed ground truth on the clean recordings using DTW on MFCC features
  - Manually verified the warped recordings were synchronized
- Experimental setup
  - Mix with noise, with varying SNR
- Evaluation metric
  - % of windows that are mapped outside of undetectable region

| | "Crowd" noise | | | "Wind" noise | | |
|---|---|---|---|---|---|---|
| | 0 dB | -5 dB | -10 dB | 0 dB | -5 dB | -10 dB |
| SyncNet [Chung and Zisserman 2016] | 88.49 | 88.49 | 88.49 | 88.49 | 88.49 | 88.49 |
| Adobe Audition [King et al. 2012] | 4.07 | 10.23 | 10.61 | 4.85 | 4.93 | 10.09 |
| Audio-to-Video | 7.26 | 7.26 | 7.26 | 7.26 | 7.26 | 7.26 |
| Audio-to-Video (with delay) | 4.12 | 4.12 | 4.12 | 4.12 | 4.12 | 4.12 |
| Audio to Video+Audio | 2.03 | 1.98 | **2.03** | 3.77 | 5.04 | 5.85 |
| Audio to Video+Audio (with delay) | **0.61** | **0.88** | 4.25 | **1.21** | **1.22** | **4.03** |

https://github.com/tavihalperin/AV-snap

## References

[1] Joon Son Chung and Andrew Zisserman. "*Out of time: automated lip sync in the wild.*" ACCV 2016
[2] Brian King, Paris Smaragdis, and Gautham J. Mysore: "*Noise-robust dynamic time warping using PLCA features*" ICASSP 2012