A Deep Dive Into Neural Synchrony Evaluation for Audio-visual Translation

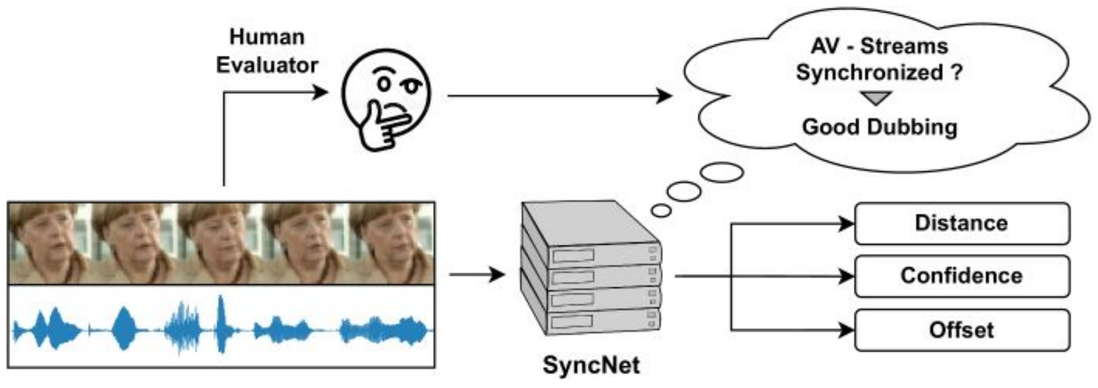
Shravan Nayak, Christian Schuler, Debjoy Saha, Timo Baumann

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

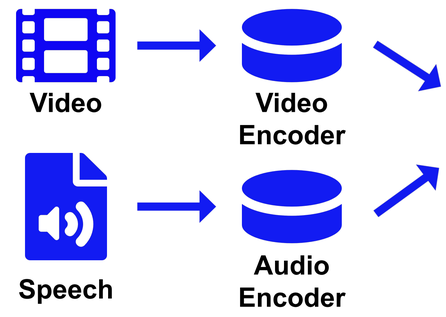
OSTBAYERISCHE TECHNISCHE HOCHSCHULE REGENSBURG

Motivation

Automated dubbing consists of automated translation of the source language, synchronous re-speaking in the target language and potentially adaptations to lip-movement visible on screen. Any neural evaluation tool should measure the synchrony between its audio and video streams (AV-Sync) and rate accordingly. However, for the AV-Sync scores to be reliable, they must align as best as possible with human perception.



SyncNet Fundamentals



Training SyncNet is trained using a contrastive learning framework; by minimizing (maximising) the L2 distance between the synced (unsynced) audio-video pairs.

Inference We test these metrics, extracted from SyncNet output, on their ability to characterize AV-Sync -

- Distance (LSE-D)
- Confidence (LSE-C)
- Offset

Dubbing proposes a unique challenge, where due to local asynchronies (jitter), audio and video streams can never be perfectly synchronized. SyncNet is not trained for, nor is it benchmarked for video data containing these issues.

The question is - Do SyncNet scores reliably quantify AV synchrony?

Our Findings

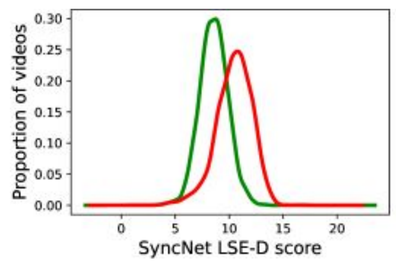
1. Empirical Analysis of Syncnet

Dataset	LRW	LRS2	LRS3	GRID	VoxCeleleb2	LipWav	Merkel	Heroes
LSE-D	7.01	6.74	6.96	6.87	7.51	6.93	7.81	8.60
LSE-C	6.93	7.84	7.59	7.68	7.00	7.71	6.29	3.60

We observe that SyncNet scores vary across datasets, usually worsening with noisy in-the-wild videos and different languages. However, scores are largely invariant to duration or speaker face angle, but shorter videos show a larger variance in scores. We observed a large negative correlation of -0.78 between metrics LSE-C and LSE-D.

2. SyncNet on dubbed data

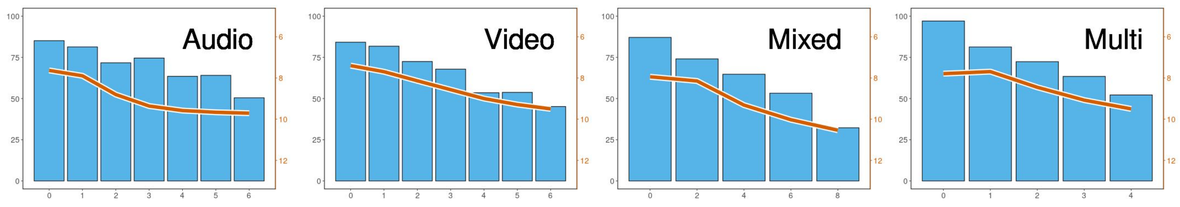
Type	Language	LSE-D	LSE-C
On-screen	English	8.6 (1.3)	3.6 (2.0)
	German (Dub)	10.5 (1.6)	1.9 (0.9)
Off-screen	English	9.7 (2.3)	1.5 (0.9)
	German (Dub)	9.8 (2.3)	1.5 (0.9)



SyncNet is able to distinguish between Original and Dubbed speech. The videos in the overlapping region of the density plots are of small durations, where SyncNet scores are unreliable. However, human dubbed speech is rated worse than off-screen speech.

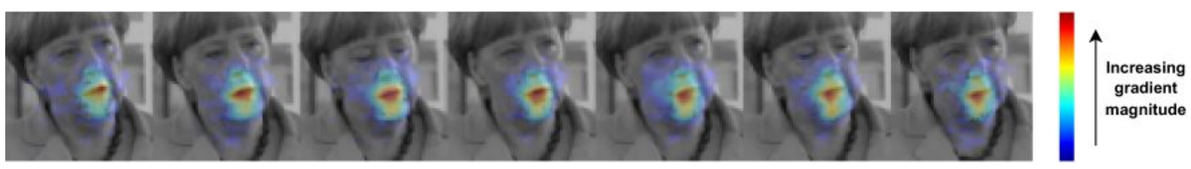
3. Syncnet scores relative to human judgement

We perform a study with human participants (83) who rate stimuli that have been manipulated (audio, video and combinations thereof) in order to yield asynchronies, assess the gravity of asynchronies in different types of stimuli, and check how well SyncNet matches human preferences. The condition *Mixed* combines audio and video asynchronies in opposite directions while *Multi* introduces synchronous manipulations by shifting both audio and video in the same direction.

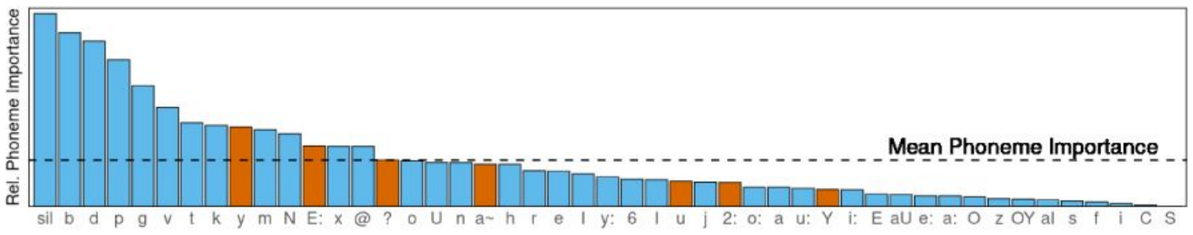


The general tendency of humans while assessing lip-synchrony in video material, corresponds to the LSE-D scores of SyncNet. Notable here is the stark contrast in regard of synchrony impairments introduced by editing the audio part of the video. SyncNet scores fall off quicker even for small impairments other than humans, who show a higher tolerance in this case.

4. SyncNet focus in Space and Time



By using explainability algorithm Integrated Gradients, we find that SyncNet consistently focuses near the mouth region of the speaker videos to capture lip movements.



From the gradients, we obtain the importances of each phoneme. We find that SyncNet pays very close attention to silences. Furthermore, bilabial and labio-dental phonemes (particularly plosives) are of highest importance, reaffirming the importance of lip closures for AV-synchrony.

Conclusion

We found that SyncNet scores by-and-large match human judgements but also that humans are more forgiving towards small audio shifts. We have also found that SyncNet bases decisions on similar criteria as are found to be important in dubbing. However, SyncNet is still lacking with regards to differentiating “reasonably” dubbed speech from badly or not at all dubbed speech which may hinder its applicability in assessing (and improving) dubbing applications.

Acknowledgments

We thank DAAD for granting internship scholarships to the first and third authors; we also want to thank Universität Hamburg's Language Technology group for the continued support after the internship.