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

SHRAVAN NAYAK and CHRISTIAN SCHULER, Universität Hamburg, Germany
DEBJOY SAHA, IIT Kharagpur, India and Universität Hamburg, Germany
TIMO BAUMANN, OTH Regensburg, Germany and Universität Hamburg, Germany

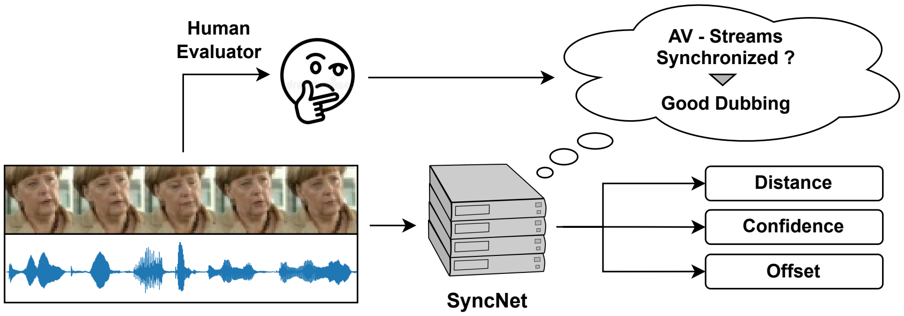

Fig. 1. Workflow of SyncNet when assessing audio-visual synchrony and its relation to human perception of dubbing.

We present a comprehensive analysis of the neural audio-visual synchrony evaluation tool SyncNet. We assess the agreement of SyncNet scores vis-a-vis human perception and whether we can use these as a reliable metric for evaluating audio-visual lip-synchrony in generation tasks with no ground truth reference audio-video pair. We further look into the underlying elements in audio and video which vitally affect synchrony using interpretable explanations from SyncNet predictions and analyse its susceptibility by introducing adversarial noise. SyncNet has been used in numerous papers on visually-grounded text-to-speech for scenarios such as dubbing. We focus on this scenario which features many local asynchronies (something that SyncNet isn't made for).

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **Applied computing** → **Language translation**.

Additional Key Words and Phrases: audio-visual synchrony, speech-lip synchrony, dubbing

ACM Reference Format:
Shravan Nayak, Christian Schuler, Debjoy Saha, and Timo Baumann. 2022. A Deep Dive Into Neural Synchrony Evaluation for Audio-visual Translation. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3536221.3556621

1 INTRODUCTION

Audio-visual synchrony (AV-sync), especially for visibly speaking faces has long been of interest in the signal processing community. Operations such as mixing, transmission, reception and processing of signals lead to unequal delays for each modality, which, if left uncorrected, may cause an unpleasant viewing experience. Television companies tackle this by embedding time stamps and tags in the audio and video streams facilitated by standardised frameworks [6]. Recently, multiple papers have proposed deep learning to rectify AV-sync errors using only the underlying audio-visual data [12, 19, 21, 36]. The first such model was SyncNet [12] which was proposed to correct time shifts between audio and video. Recent research has successfully used SyncNet for tasks that require to assess jitter (local discrepancies between audio and video), rather than time shift, like lip generation [31, 37, 42], visually-grounded speech synthesis [17, 18, 23], lip-reading [12, 39] and speaker diarization [2, 14], both as an AV-sync metric and for generating meaningful representations. SyncNet's popularity can partially be attributed to the open-sourced code repository[1] and model weights provided by the authors. In this paper, we look into SyncNet's effectiveness as an AV-sync metric for dubbing-related applications which also features a high amount of AV jitter.

Automated dubbing consists of automated translation of the source language [33], synchronous re-speaking in the target language and potentially adaptations to lip-movement visible on screen. Processes like speaker-video grounded speech synthesis or modifying the lip movements in the video according to the generated audio can help improve synchrony and boost dubbing quality and naturalness. A key aspect of automated dubbing tasks is that perfect synchronization is virtually impossible, although it is broadly characterized and evaluated using AV-sync. In these tasks, for the scores to be reliable, they must align as best as possible with human perception. While SyncNet has already been (successfully) used as a substitute for human evaluations [17, 18, 23], there have been no scientific studies yet regarding its agreement with human evaluations. So, in this paper, we attack the question of how reliably SyncNet mimics human judgements of audio-visual synchrony and can hence be expected to be useful as a target function in model training or as a substitute to human evaluations. Specifically, our contribution is threefold:

- We evaluate the SyncNet's agreement with human evaluations across a wide range of varying input properties like language, recording conditions and noise, and try to identify any underlying biases that may be due to a data distribution shift.
- We analyze the interpretability of the SyncNet model and identify the components in the audio-visual input data that are affecting the predicted AV-sync error. We relate these findings to expert heuristics when manually preparing and performing dubbed speech and find interesting similarities and differences that will be useful to further refine future models.
- We assess SyncNet performance on dubbed material and draw conclusions on its applicability in this domain.

2 AUDIO-VISUAL TRANSLATION: "DUBBING"

Dubbing, or audio-visual translation, is the process of replacing the audio track in a video clip with another, containing the dialogue translated into a different language. In many countries such as India, Spain, Italy and Germany, dubbed content is the preferred mode of foreign AV media consumption (as compared to others where subtitling is more prevalent, such as in Scandinavia). The key objective for dubbing is to create the illusion that the dubbed audio is spoken first-hand by the speaker on-screen. Maximising the AV-sync between the speech embedded in the audio and the video track can help ensure this.

The task of translation in itself provides potential for controversy [7] and one linguistic definition can be: "translation is a type of language mediation, socially serving to approximate a mediated bilingual communication to a common monolingual communication". [38, p.4] The process of translation is thus a complex cognitive process that consists of decoding the meaning of a source language text and re-encoding this meaning into a target language text, and requiring knowledge of not only the grammar, semantics, syntax, etc., but also of the culture corresponding to the respective communities that use both these languages. [3]

[1] https://github.com/joonson/syncnet_python

Dataset	LRW [11]	LRS2 [10]	LRS3 [1]	GRID [13]	VoxCeleb2 [9]	Lip2Wav [32]	Merkel [34]	Heroes [27]
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

Table 1. Mean LSE-D and LSE-C scores for several audio-visual datasets. Scores aggregated from [17, 18, 30, 34]. The language in all datasets is English, except for Merkel which is German.

With the addition of "audio-visual", the goal of translation extends to multi-modal alignment, requiring the source and target speech to align in terms of lip movements (and to a lesser degree to accentuations, facial mimicry and gesturing of the actor). The intended multi-modal perception of a film consists of "watching" and "listening" and these are inseparably intertwined and performed simultaneously. McGurk [24] found understanding difficulties based on active perception given conflicting auditory and visual cues. Similarly, Buchan *et al.* found that participants in an eye-tracking study paid more (visual) attention to lips under noisy conditions [5], making high lip-synchrony of dubbing key for understandability in noisy speech. Many experiments have found the overall quality perception of AV material to depend on each modality's quality as well as their matching [29] implying that the quality of dubbing will be ascribed as the quality of a given film.

Dubbing has been studied particularly in the arts and humanities and the manual dubbing process is largely driven by heuristics about what constitutes 'good' dubbing: *quantitative* similarity is concerned with the coordination of time of speech and lip movements and is meant to avoid visual or auditory phantom effects. *Qualitative* similarity is important once quantity is established, and is concerned with matching visemic characteristics (i.e., what speech sounds look like when pronounced) of source and target speech, such as opening angle of the jaw for vowels and lip closure for consonants. In this regard, there shall be a special focus on open vowels (such as /a:/ and particularly in stressed syllables), bilabial consonants (/p b m/) and labio-dental consonants /t d n/). [8]

3 AN EMPIRICAL BACKGROUND ON SYNCNET

SyncNet [12] is a multimodal model that predicts metrics indicating the synchrony of the input audio-video pair. Video is input to the SyncNet in the form of face crops and audio is passed to the model in the form of MFCC features [25]. Windows of 0.2 s are used to extract segments of video and audio inputs which are subsequently passed through separate video and audio encoders (comprised of convolutional layers) to generate video and audio embeddings respectively. The model is trained to minimize (maximise) the L2-distance between these set of embeddings for the sync'ed (unsync'ed) pairs using a max-margin loss and negative sampling. The model has been used for synchrony evaluation based on the two metrics LSE-C (Confidence) and LSE-D (Distance) [30]. For calculating the LSE-D metric, the video input is shifted (wrt. audio) by a range of offset values. For each shift value, the windowed L2-distance between the audio and video features, averaged over the entire duration of the snippet was used to quantify the synchrony score for that offset. The resulting LSE-D was taken as the minimum of these synchrony scores. A lower LSE-D implies better audio-visual synchronization. A higher LSE-C value denotes that the minimum distance offset obtained has a much lower LSE-D compared to adjacent offsets and therefore the model has higher confidence in it being the correct offset.

Table 1 shows LSE-D and LSE-C scores aggregated for various datasets. We observe that SyncNet scores vary across datasets. Particularly, VoxCeleb2 and Heroes dataset show a larger deviation from other English datasets. Further, scores for Merkel dataset (in German) are worse, calling their language independence [12] into question. While one has to be careful while comparing SyncNet scores from data coming from different sources, we find that they are reasonably consistent across varying conditions.

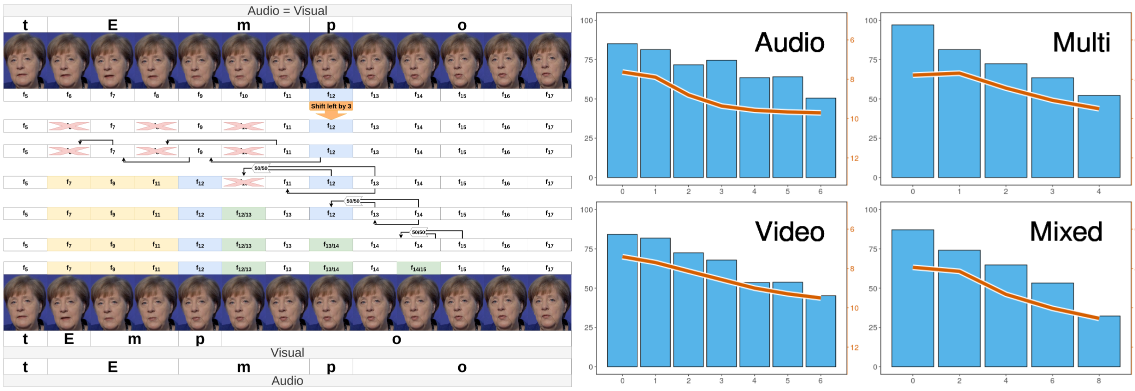

Fig. 2. **Left**: De-synchronisation process wrt. the target phoneme /p/ in "tempo" in the *video* condition: frames are skipped before/duplicated after the target, yielding a de-synchronised impression of the lip closure to the plosive. Similar de-synchronisation is performed by shortening/lengthening surrounding speech for the *audio* condition, both in opposing directions are combined for *mixed* and in the same direction for the *multi* condition. **Right**: Comparison of synchony judgements by humans (gray bars, higher is better) and SyncNet LSE-D (red line) grouped by the four applied types of modifications and over strength of introduced asynchrony.

We do a quantitative analysis of SyncNet performance with different varying input attributes, using a subset of the LRS3 and Merkel datasets. On videos of different durations (1–30 seconds) we observe that scores are invariant to duration on average, but shorter videos show a larger variance in scores. We also analyze wrt. speaker-face direction (roll, pitch and yaw) and find scores to be largely independent as well. Finally, we analyze the inter-dependencies between the SyncNet metrics LSE-C and LSE-D, and observe a Pearson correlation coefficient of -0.74 (on LRS3, -0.78 on Merkel) indicating a strong negative correlation.

4 SYNCNET SCORES RELATIVE TO HUMAN JUDGEMENTS

As outlined above, the perceptual gravity of asynchronies is thought to depend on what speech sounds and corresponding visemes are desynchronized and whether asynchronies stem from the auditive or visual channels. We perform a study with human participants who rate stimuli that have been manipulated in order to yield asynchronies, assess the gravitiy of asynchronies in different types of stimuli, and check how well SyncNet matches human preferences.

Among the wide range of corpora for audio-visual speech, we select the Merkel Corpus [34] as the base of our study which features web-streamed podcasts of former German chancellor Angela Merkel as the main speaker. It provides clear views of the speaker's face and lips at only moderate angles of the face and often as a close-up shot with little movement wrt. the camera or scene. Recordings are of studio quality with high frame rate, resolution and audio quality. The corpus comes with time-aligned transcripts which were further extended with phoneme-level timings from forced alignment [35] for this study. We hand-selected samples from the corpus which can be characterized as American Medium and Medium-Long Shots [4], an example of which can be seen in Figure 3. Each sample contains a target phoneme combination under consideration for manipulation (/p/ as a bilabial plosive and /a/ as an open vowel as well as inter-word pauses) particularly clearly and with varied surrounding speech material.

From each source sample, we derive variations in which the *audio* or *video* track (or both, termed the *mixed* condition) are locally de-synchronized around the target phoneme by shortening/lengthening the material before the target (and doing the opposite afterwards). Audio duration manipulation is performed using sox[2] while video manipulation is

[2] SoX is a cross-platform command line utility for performing audio manipulations (http://sox.sourceforge.net/sox.html).

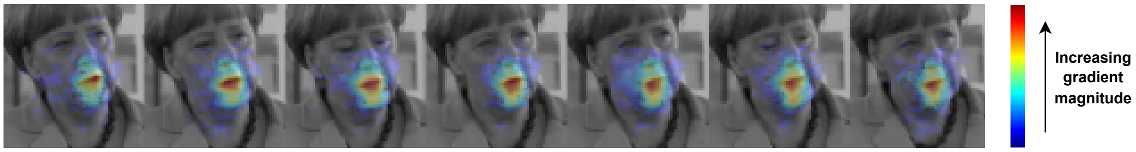

Fig. 3. Gradients obtained from Integrated-Gradients overlaid on the video (Reddish tones signify larger gradients)

performed by skipping and interpolating frames. All processing uses a granularity of 40 ms (based on the frame rate of 25 fps) and asynchronies of 1-6 frames are introduced (2-8 frames for the *mixed* case which combines audio and video asynchronies in opposite directions). Furthermore, we also introduce synchronous manipulations by shifting both audio and video in the same direction (termed the *multi* condition). Asynchronies are strongest at the location of the specified phoneme, weaker in the close proximity to it and non-existent in the rest of the video sequence. The *multi* condition does not introduce asynchronies but only the artifacts stemming from AV manipulation.

We gathered human ratings of synchrony with a MUSHRA test design [40] using a web-based experiment platform [22]. 83 participants provided 9234 observations out of which 70 (6930) passed our quality control (26 female, 39 male, 5 no info) As part of this quality control, we removed observations from tests that where performed in an unreasonably short time and also duplicates, created by participants restarting the session.

As the data indicates and the direct comparison shows (Figure 2), the general tendency of humans while assessing lip-synchrony in video material, corresponds to the LSE-D scores. Notable here is the stark contrast in regard of synchrony impairments introduced by editing the audio part of the video. While humans and SyncNet both tend to give a lower rating for stronger modified videos, SyncNet drastically penalizes even small impairments compared to humans, who show a higher tolerance in this case. This pronounced sensitivity reveals a level of detail that is not observed in human viewers and might lead to undesired consequences when using SyncNet LSE-D as a target function.

5 SYNCNET FOCUS IN SPACE AND TIME

In Section 4, we evaluated SyncNet's alignment with human evaluations. Detecting asynchronies is an inherent quality of human beings who develop varying sensitivities based on the type of audio-visual content they are exposed to over their lifetime. While an agreement empirically shows good performance for SyncNet, in this section, we attempt to understand the aspects of audio and video responsible for affecting the synchrony scores using neural network explainability and adversarial techniques [15, 41]. We believe that, in addition to understanding the inner working of SyncNet, this analysis helps to further our understanding of human synchrony perception. For convenience and comparability, we perform the following experiments on the Merkel Corpus [34] as well. However, based on the arguments provided in Section 3, we believe our findings likely generalise to other datasets.

Model Explainability. As a naive approach, similar to the Fast Gradient Sign Method (FGSM) attack [16], we obtain the gradients at the inputs by backpropagating the final predicted distance through each layer of the network down to the input features, i.e., each individual video pixel and audio sample. We also examine the gradients obtained on the video features using Integrated Gradients [26], a more reliable and powerful explainability technique which works by approximating the integral over the gradient, per feature. We demonstrate the image gradients overlaid onto the original video frames in Figure 3. The figure shows that SyncNet focuses near the mouth region of the speaker videos to capture the lip movements essential for determining synchrony (and this also becomes apparent from aggregations of many frames). This leads us to the conclusion that SyncNet is correctly focusing on the relevant parts of the image when determining synchrony.

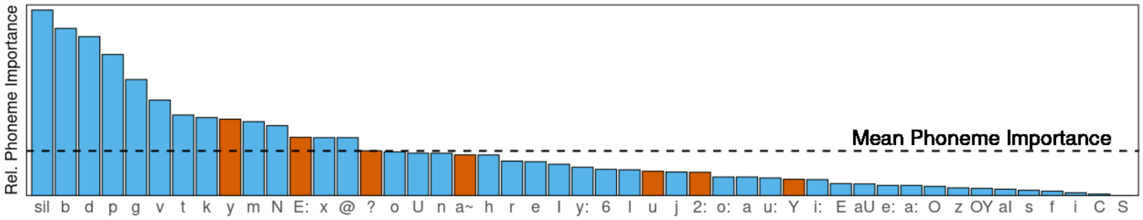

Fig. 4. Phoneme importance for SyncNet scores (phonemes coded in SAMPA, rare ones shown in orange).

In addition to the spatial locality of SyncNet gradients in the video, we analyze *when* in the audio SyncNet's synchrony decisions are predominantly determined. Using the audio gradients, we identify almost continuous intervals of high gradient values and relate these to the phonemes spoken at that point in time (based on MAUS alignments [35]). Based on the human heuristic that some phoneme types are more relevant for dubbing synchrony, we define the relevance of each phoneme type as the duration of its overlap with high gradient regions relative to the total duration of this phoneme in the entire corpus. We show the phoneme importance for each phoneme in Figure 4.

We find that SyncNet by-and-large follows the human heuristic: it pays very close attention to silences (between words and sentences) which leads us to conclude that it aims to check for "phantom movements" when no speech is present. Furthermore, bilabial and labio-dental phonemes (particularly plosives) are of highest importance among the speech sounds, which matches human intuition that lip closures are of particular importance when assessing lip synchrony. It is very interesting to note that the heuristic of matching jaw movements (by matching the degree of openness of vowels) is disregarded by the learnt model and this likely matches the relative unimportance of this heuristic as compared to matching quantity (by observing silences) and matching lip closures.

Adversarial Attacks. For FGSM attacks, the mode of application is taking one gradient update along the direction of the sign of gradient at each pixel. On performing FGSM attacks on SyncNet, we find that with a pixel perturbation magnitude of 3 for video input, we obtain an average increase of 1.5 in the LSE-D (i.e., the difference between 'well' and 'badly' judged synchrony in the perceptual evaluation) with no perceptible change in video quality. This proves that SyncNet is sensitive to noise-induced adversarial attacks, which must be kept in mind while leveraging SyncNet predictions for learning deep models for AV tasks. Furthermore, running FGSM attacks with the opposite objective of improving SyncNet scores was found to be a much harder task. Very insignificant decreases in LSE-D were observed, that too just with sub-pixel level perturbations. With large gradient update steps (>1), an increase in LSE-D was observed instead. A possible explanation for this is that ground-truth audio-video pairs lie near, if not exactly in the global minimum of LSE-D scores. So it is expected that large gradient updates can cause overshooting of the minima value.

6 SYNCNET ON DUBBED DATA

We here analyze SyncNet's effectiveness as an AV-sync metric for dubbing data. We use the Heroes television series as transcribed and made available [27, 28] and further labeled with each of the English utterances as being spoken on-screen, where the actor's face is visible on screen while speaking the utterance, and off-screen where it is not [20]. The original corpus does not contain video, so we use a DVD release with English and German dubbed video. For each of the utterances labeled on/off-screen, we get the corresponding video snippet in English and German, and analyze LSE-D and LSE-C scores on original as well as dubbed videos. We crop on-screen videos to the faces visible and resize to a 224x224 size. For off-screen, we directly resize to 224x224. We then run the SyncNet model and obtain LSE-D and LSE-C scores for each of these videos.

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)

Table 2. SyncNet scores for on/off screen, English/German.

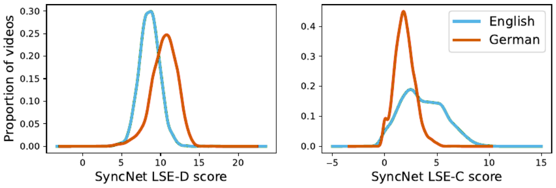

Fig. 5. Density plot for LSE-D and LSE-C scores for on-screen original English and dubbed German videos.

On-screen Analysis. Table 2 shows that for on-screen, English videos have a lower LSE-D and higher LSE-C scores compared to German dubbing. Dubbed audio does not follow lip movements accurately and SyncNet notices this. However, scores for English in Heroes are worse than all other datasets (Table 1) indicating an impact of the large variety of pose, lighting, perspective and possibly encoding artifacts. Further, the density plots in Fig. 5 show quite a clear demarcation between English original and German dubbing. We manually analyzed videos where LSE-D is lower for German dubbing as compared to English original, i.e., the videos that lie in the overlap of the two density plots. We observe that these are very short on average (1.17 s; $\sigma = 0.73$, median 1 s) and as described in Section 3, scores show high variance for very short videos. This may indicate that these scores are not systematically better.

Off-Screen Analysis. Table 2 shows that the SyncNet scores for both English and German dubbed off-screen videos are almost equal. This is expected since in these videos there typically is no face whose lips are moving. It is remarkable, however, that scores for on-screen dubbed German are worse than for off-screen. This leads us to believe that SyncNet penalizes more when lip motions are incorrect than for no lip motion at all (or no lips, for that matter), and this despite the fact that SyncNet was shown to pay close attention to "phantom speech" in Section 5.

It will therefore be crucial to combine SyncNet scoring with on/off-screen analysis when assessing dubbing quality.

7 CONCLUSION

We have analyzed to what extent SyncNet is applicable to dubbing automation. Dubbing poses unique challenges such as the target never being perfectly alignable given that the target language's phonetic realization differs from the source. 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 (pauses to avoid "phantom speech" and plosives to match lip closure). At the same time, we find evidence that opening angle of the jaw is less relevant than was previously hypothesized in the literature on dubbing.

However, dubbed speech is rated as badly as completely asynchronous speech in the Heroes corpus. This can be seen as an indication that 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.

We believe that future improvements to AV-sync assessment for dubbing can be made if it inludes dubbing and, foremost, if human judgements are better taken into account (e.g. by including high-quality dubbed TV in the target language in the training data).

There do exist some alternatives to SyncNet that show slightly better results on audiovisual time offset detection tasks [12, 19, 21, 36]. However, as these are not specifically trained for the dubbing task, our results are likely to carry over to these methods as well.

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.

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR* abs/1809.00496 (2018). arXiv:1809.00496 http://arxiv.org/abs/1809.00496

[2] Rehan Ahmad, Syed Zubair, Hani Alquhayz, and Allah Ditta. 2019. Multimodal speaker diarization using a pre-trained audio-visual synchronization model. *Sensors* 19, 23 (2019), 5163.

[3] Stefano Arduini and Robert Hodgson. 2007. *Similarity and Difference in Translation.* Ed. di Storia e Letteratura.

[4] R. Barsam and D. Mohanan. 2010. Looking at Movies: An Introduction to Film. 3 rd ed. New York:W. W. Norton & Company..

[5] Julie N. Buchan, Martin Paré, and Kevin G. Munhall. 2008. The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research* 1242 (Nov. 2008), 162–171. https://doi.org/10.1016/j.brainres.2008.06.083

[6] Dick Bulterman. 2008. *Synchronized Multimedia Integration Language (SMIL 3.0).* W3C Recommendation. W3C. https://www.w3.org/TR/2008/REC-SMIL3-20081201/.

[7] Frederic Chaume. 2018. An overview of audiovisual translation: Four methodological turns in a mature discipline. *Journal of Audiovisual Translation* 1 (Nov. 2018), 40–63. https://doi.org/10.47476/jat.v1i1.43

[8] Frederic Chaume Varela. 2004. Synchronization in dubbing: A translational approach. In *Benjamins Translation Library*, Pilar Orero (Ed.). Vol. 56. John Benjamins Publishing Company, Amsterdam, 35–52. https://doi.org/10.1075/btl.56.07cha

[9] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH.*

[10] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip Reading Sentences in the Wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 3444–3453.

[11] Joon Son Chung and Andrew Zisserman. 2017. Lip Reading in the Wild. In *Computer Vision – ACCV 2016*, Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato (Eds.). Springer International Publishing, Cham, 87–103.

[12] Joon Son Chung and Andrew Zisserman. 2017. Out of Time: Automated Lip Sync in the Wild. In *Computer Vision – ACCV 2016 Workshops*, Chu-Song Chen, Jiwen Lu, and Kai-Kuang Ma (Eds.). Springer International Publishing, Cham, 251–263.

[13] Martin Cooke, Jon Barker, Stuart P. Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America* 120 5 Pt 1 (2006), 2421–4.

[14] Yifan Ding, Yong Xu, Shi-Xiong Zhang, Yahuan Cong, and Liqiang Wang. 2020. Self-Supervised Learning for Audio-Visual Speaker Diarization. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), 4367–4371.

[15] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (2018), 80–89.

[16] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. *CoRR* abs/1412.6572 (2015).

[17] Michael Hassid, Michelle Tadmor Ramanovich, Brendan Shillingford, Miaosen Wang, Ye Jia, and Tal Remez. 2021. More than Words: In-the-Wild Visually-Driven Prosody for Text-to-Speech. *ArXiv* abs/2111.10139 (2021).

[18] Chenxu Hu, Qiao Tian, Tingle Li, Wang Yuping, Yuxuan Wang, and Hang Zhao. 2021. Neural Dubber: Dubbing for Videos According to Scripts. *Advances in Neural Information Processing Systems* 34 (2021).

[19] Venkatesh S. Kadandale, Juan F. Montesinos, and Gloria Haro. 2022. VocaLiST: An Audio-Visual Synchronisation Model for Lips and Voices. *ArXiv* abs/2204.02090 (2022).

[20] Alina Karakanta, Supratik Bhattacharya, Shravan Nayak, Timo Baumann, Matteo Negri, and Marco Turchi. 2020. The Two Shades of Dubbing in Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 4327–4333. https://doi.org/10.18653/v1/2020.coling-main.382

[21] You Jin Kim, Hee-Soo Heo, Soo-Whan Chung, and Bong-Jin Lee. 2021. End-To-End Lip Synchronisation Based on Pattern Classification. *2021 IEEE Spoken Language Technology Workshop (SLT)* (2021), 598–605.

[22] Sebastian Kraft and Udo Zölzer. 2014. *BeaqleJS: HTML5 and JavaScript based Framework for the Subjective Evaluation of Audio Quality.*

[23] Junchen Lu, Berrak Sisman, Rui Liu, Mingyang Zhang, and Haizhou Li. 2022. Visualtts: TTS with Accurate Lip-Speech Synchronization for Automatic Voice Over. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2022), 8032–8036.

[24] Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264, 5588 (Dec. 1976), 746–748. https://doi.org/10.1038/264746a0 Number: 5588 Publisher: Nature Publishing Group.

[25] Paul Mermelstein. 1976. Distance Measures for Speech Recognition – Psychological and Instrumental. In *Pattern Recognition and Artificial Intelligence, Proceedings of the Joint Workshop on Pattern Recognition and Artificial Intelligence*, C. H. Chen (Ed.). 374–388.

[26] A. Natarajan, M. Motani, B. de Silva, K. Yap, and K. C. Chua. 2007. Investigating Network Architectures for Body Sensor Networks. In *Network Architectures*, G. Whitcomb and P. Neece (Eds.). Keleuven Press, Dayton, OH, 322–328. arXiv:960935712 [cs]

[27] Shravan Nayak, Timo Baumann, Supratik Bhattacharya, Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. See me Speaking? Differentiating on Whether Words are Spoken On Screen or Off to Optimize Machine Dubbing. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. ACM, Virtual Event Netherlands, 130–134. https://doi.org/10.1145/3395035.3425640

[28] Alp Öktem, Mireia Farrús, and Antonio Bonafonte. 2018. Bilingual Prosodic Dataset Compilation for Spoken Language Translation. In *Proceedings of IberSPEECH 2018* (Barcelona, Spain, 21-23 November 2018). 20–24. https://www.isca-speech.org/archive/IberSPEECH_2018/pdfs/IberS18_P1-1_Oktem.pdf

[29] Margaret H. Pinson. 2011. Audiovisual Quality Components: An Analysis. NA (Nov. 2011). https://www.its.bldrdoc.gov/publications/details.aspx?pub=2565 Publisher: ITS.

[30] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and C. V. Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. *28th ACM International Conference on Multimedia (ACM MM)* (Oct. 2020). https://doi.org/10.1145/3394171.3413532 Publisher: Association for Computing Machinery.

[31] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. *A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild*. Association for Computing Machinery, New York, NY, USA, 484–492. https://doi.org/10.1145/3394171.3413532

[32] Prajwal K R, Rudrabha Mukhopadhyay, Vinay Namboodiri, and C. V. Jawahar. 2020. Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 13793–13802.

[33] Ashutosh Saboo and Timo Baumann. 2019. Integration of Dubbing Constraints into Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Association for Computational Linguistics, Florence, Italy, 94–101. https://doi.org/10.18653/v1/W19-5210

[34] Debjoy Saha, Shravan Nayak, and Timo Baumann. 2022. Merkel Podcast Corpus: A Multimodal Dataset Compiled from 16 Years of Angela Merkel's Weekly Video Podcasts. *ArXiv* abs/2205.12194 (2022).

[35] Florian Schiel. 2004. MAUS Goes Iterative. In *LREC*.

[36] Yoav Shalev and Lior Wolf. 2020. End to End Lip Synchronization with a Temporal AutoEncoder. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2020), 330–339.

[37] Shijing Si, Jianzong Wang, Xiaoyang Qu, Ning Cheng, Wenqi Wei, Xinghua Zhu, and Jing Xiao. 2021. Speech2Video: Cross-Modal Distillation for Speech to Video Generation. In *Interspeech*.

[38] Yaroslav V. Sokolovsky. 2010. On the Linguistic Definition of Translation. *undefined* (2010). https://www.semanticscholar.org/paper/On-the-Linguistic-Definition-of-Translation-Sokolovsky/b08bccc1d956ed35b5d1c5f89d7e9972cd3532ae

[39] Joon Son Son and Andrew Zisserman. 2017. Lip Reading in Profile. In *Proceedings of the British Machine Vision Conference (BMVC)*, Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk (Eds.). BMVA Press, Article 155, 11 pages. https://doi.org/10.5244/C.31.155

[40] International Telecommunication Union. 2015. *Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems*. Technical Report. International Telecommunication Union.

[41] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems* 30 (2019), 2805–2824.

[42] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 4174–4184.