IMPROVING LIP-SYNCHRONY IN DIRECT AUDIO-VISUAL SPEECH-TO-SPEECH TRANSLATION

Lucas Goncalves, Prashant Mathur[*], Xing Niu*[*], Brady Houston, Chandrashekhar Lavania, Srikanth Vishnubhotla*[†], Lijia Sun*[‡], Anthony Ferritto*[‡]

Amazon

ABSTRACT

Audio-Visual Speech-to-Speech Translation (AVS2S) typically prioritizes improving translation quality and naturalness. However, an equally critical aspect in audio-visual content is lip-synchrony—ensuring that the movements of the lips match the spoken content—essential for maintaining realism in dubbed videos. Despite its importance, the inclusion of lip-synchrony constraints in AVS2S models has been largely overlooked. This study addresses this gap by integrating a lip-synchrony loss into the training process of AVS2S models. Our proposed method significantly enhances lip-synchrony in direct audio-visual speech-to-speech translation, achieving an average LSE-D score of 10.67, representing a 9.2% reduction in LSE-D over a strong baseline across four language pairs. Additionally, it maintains the naturalness and high quality of the translated speech when overlaid onto the original video, without any degradation in translation quality.

Index Terms— Audio-visual speech translation, Lip synchrony, Automatic translation systems

1. INTRODUCTION

Traditionally, research has emphasized the importance of lip synchrony – the alignment of translated audio with the visible mouth movements of the original actors—as a key factor in maintaining the quality and realism of dubbed content [1, 2, 3]. However, improving lip synchrony should not compromise translation quality and naturalness [4, 5]. Simply achieving a trade-off between lip synchrony and translation quality may not significantly enhance the overall user experience. In this study, we aim to improve lip-synchrony in dubbed videos while ensuring that translated speech is natural and of high quality when overlaid on original videos, addressing the trade-off.

Recent advancements in audio-visual speech translation [6] have focused on generating translated audio-visual outputs from audio-visual inputs. These approaches typically generate visuals by modifying lips to match the audio [7] which can generate artifacts and lead to two main problems 1) generation of deepfake [8] which raises ethical concern [9] 2) generation of deepfake videos without safeguarding people's identities and personalities ('likeness') from being digitally recreated without their consent [10]. Modifying a speaker's lip movements in automatic translations could violate these concerns, potentially infringing on their image rights [11, 12]. Moreover, generating high-quality video outputs in audio-visual speech translation currently poses significant challenges. State-of-the-art video generation models often produce artifacts, such as issues with teeth appearance [13], which can distract viewers and negatively affect the overall viewing experience. In this work, we sidestep the problem of modifying videos and our project solely focuses on improving speech translation while maintaining the original visual content. This approach avoid introducing any visual artifacts and maintain the integrity of the original video.

While early works incorporated visual information to improve translation quality of speech to speech translation models, none have actively worked maintaining lip synchrony in the generated target speech and original video [14, 15]. Current works like [6] do not explicitly utilize lip-synchrony as a constraint in model training, and the synchrony evaluations are performed only on the outputs where both audio-visual are generated. In this work, we explore the idea of leveraging visual inputs to enhance lip synchrony between generated speech in target language and original video, at the same time, sidestepping generation of visuals along with the speech. Our contribution focuses on integrating a lip-synchrony loss into training process of audio-visual (AV) speech translation models. As discussed in [5], dubbers typically violate lip-sync in order to achieve better translation quality and naturalness in speech, so we explore ways to improve lip-synchrony in the training process without compromising on translation quality or naturalness of the translated speech.

2. RELATED WORK

Lip synchronization has emerged as a crucial research area with wide-ranging applications, particularly in automatic dubbing for translation [1]. Numerous studies have focused on improving synchrony in dubbing through various approaches, such as isometric translation, where translations are generated to match the length of the source text [16], and prosodic alignment, which seeks to synchronize translated text with the source speech [17, 18]. Additionally, joint training of translation and duration models has been explored to enable a single model to learn both translation and phone duration estimation [19, 20]. While these methods effectively replicate the prosody of the source speech (speech-pause structure), they often overlook the critical aspect of lip synchronization in dubbing.

Prajwal et al. [21] introduced an approach to incorporate a lip-sync discriminator, resulting in more accurate synchronization between arbitrary video and audio inputs. However, the aforementioned approaches suffer from issues such as mouth blurring and inconsistent rendering of teeth. Other approaches, instead of directly manipulating lip movements based on audio, Xie et al. [22] proposed a two-stage framework. In the first stage, a generator is trained to predict facial landmarks from the audio. In the second stage, these predicted landmarks, combined with the target frame, are used to generate the final output. However, this method has its limitations: the generated reference landmarks are often inaccurate, and the ap-

[*]Corresponding authors.

[†]Provided administrative and resourcing support.

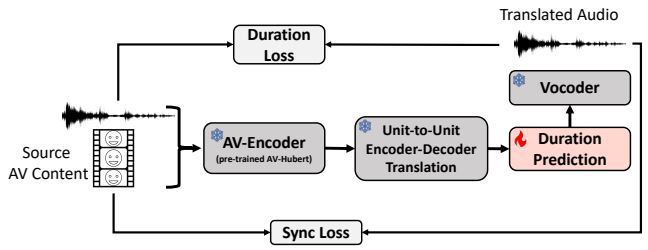

Fig. 1: AVS2S Framework Overview

proach may distort the actor's identity, similar to the method proposed by Choi et al. [6], where facial attributes are modified to match the generated audio content. One issue with these techniques is that there are ethical concerns about safeguarding individuals' identities and likenesses from being digitally recreated or modified without consent [10]. Unlike the previous approaches, we explore a research area with two realistic constraints 1) where the original videos are preserved 2) voice characteristics of original speakers are not mimicked.

3. METHODOLOGY

Our overall framework of *Audio-Visual Speech-to-Speech Translation* (AVS2S) system in depicted in Fig. 1 and is based off Choi *et al.* [6]. The framework consists of the following: visuals and speech content from the original video are fed to an Audio-Visual (AV) Encoder which processes lip region and speech content and convert them into discrete unified audio-visual units. AV-Encoder used in this work is a pre-trained multilingual AV-HuBERT [23], presented in the work of Choi *et al.*[6]. Next, we have a translation module which is a encoder-decoder network that translates source language AV unit to target language AV unit based on the work of Kim *et al.*[24]. Lastly, the vocoder is based on unit-based HiFiGAN framework [25, 26]. Essentially, we remove the visual generation component from the AV Renderer [6] and overlay the generated target language speech on the original video.

In Choi *et al.*[6], AV Renderer component generates both speech and visuals at the same time as they take the AV unit as input and feed it to vocoder and wav2lip [7] modules, which results in an inherent lip-synchrony as both modules are utilizing same AV unit as input. In our case, since we are not synthesizing face and instead overlaying the generated speech on original video, the same lip-synchrony element is lost. In this work, we aim to enhance lip-synchrony between the translated output audio stream of an input video (with faces) from one language to another using an AVS2S framework. The following section provides details about the AVS2S framework which adds two specific losses for lip-synchronization.

3.1. Duration Predictor

Since the unit-to-unit translation framework takes input as source AV units and outputs target AV units, a duration predictor is typically employed before the vocoder which pre-processes the units needed to generate the target speech content. This predictor estimates the duration of each speech unit and adjusts them accordingly based on the desired target duration [26, 24, 6]. This is done via a "length predictor" in AV2AV work, however, this model does not take into account duration of source language speech. In our initial experiments, we saw AV2AV model generating videos shorter than source and this problem can be attributed to length predictor not using source speech duration. In our work, since the output acoustic stream must be synchronized with the source video, we use a standard duration loss, computed between the source audio and the generated target speech, as defined in Eq. 1.

$$\mathcal{L}_{\text{dur}} = \frac{1}{N}\sum_{i=1}^{N}(\log d_i^p - \log d_i)^2 \tag{1}$$

where N is the total number of speech units in the sequence, d is the target duration, and d^p is the predicted duration.

In addition to duration loss, we also employ a synchronization loss, computed using the SyncNet model [27] as our AV sync expert. Similar to the approach used in Wav2Lip [7], the synchronization loss is defined as shown in Eq. 2.

$$\mathcal{L}_{\text{sync}} = \frac{1}{N}\sum_{i=1}^{N}\log\left(\text{SyncNet}^i(AV)\right) \tag{2}$$

where $\text{SyncNet}^i(AV)$ represents the synchronization score for the i-th AV pair (source video and generated audio).

The overall loss used to fine-tune the duration predictor depicted on the orange block in Fig. 1 is then formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sync}} + \lambda \cdot \mathcal{L}_{\text{dur}} \tag{3}$$

Set	#Videos	Total Duration	Avg Duration
Trainval	4,004	30 hr	3.42 sec
Test	412	51 min	2.32 sec

Table 1: LRS3 dataset statistics.

4. EXPERIMENTAL SETTINGS

4.1. Dataset

We leverage LRS3 [28] which is a large-scale video data consisting of thousands of spoken sentences collected from TED talks. The dataset statistics are provided in Table 1. Importantly, there is no overlap between the videos used in the test set and those used in the Trainval sets. In our experiments, we use the Trainval set to fine-tune the duration predictor, and perform evaluation on the test set.

4.2. Evaluation Metrics

We evaluate our system using several key metrics to ensure high-quality audio-visual lip-synchronization, accurate speech translation and audio quality (naturalness). The LSE-C (Lip Sync Error - Confidence) [7] metric measures the average confidence score, indicating the correlation between audio and video, with higher scores reflecting better AV synchronization. The LSE-D (Lip Sync Error - Distance) [7] metric calculates the distance between lip and audio representations, where lower scores signify better lip-synchronization. The Perceptual Evaluation of Speech Quality (PESQ) is an industry-standard metric for assessing audio quality, evaluating factors such as audio sharpness, call volume, background noise, clipping, and interference, with scores ranging from -0.5 to 4.5, where higher scores indicate better quality. This metric is essentially our proxy to measuring naturalness of speech. BLASER 2.0 [29] is a reference-free metric that evaluates end-to-end speech-to-speech translation by leveraging a multilingual multimodal encoder. It computes scores based on the similarity of input and output speech embeddings,

		En-Es			En-Pt			En-It			En-Fr		
	Source	**Synthetic**	**AV2AV**	**Ours**	**Synthetic**	**AV2AV**	**Ours**	**Synthetic**	**AV2AV**	**Ours**	**Synthetic**	**AV2AV**	**Ours**
LSE-C ↑	7.63	2.22	2.13	**2.45***	2.27	1.15	**3.43***	2.24	2.23	**2.97***	2.08	2.23	**2.46**
LSE-D ↓	6.88	12.08	11.6	**10.68***	12.12	12	**10.12***	11.92	11.67	**10.89***	11.97	11.74	**10.98***

Table 2: Lip-synchrony evaluations of our proposed method (Ours) against baseline models (Synthetic and AV2AV-Speech) across four language pairs (English to Spanish (Es), Portuguese (Pt), Italian (It) and French (Fr).

which correlate well with human judgment. Finally, ASR-BLEU is a metric that measures BLEU score [30] (translation quality) by comparing ASR outputs of generated speech to translations from Amazon Translate.[1]

4.3. Implementation Details

For visual feature pre-processing, we follow a similar approach to previous works by cropping the mouth region using a face detector and a facial landmark detector [6, 23]. The audio is sampled at 16kHz. In training, as shown in Fig. 1, we extract units for both inputs and targets using the frozen AV-Encoder, which can accept either audio-visual or audio/video-only inputs.

We fine-tune the duration predictor, as depicted in Fig. 1 within the orange box. At this stage, the unit-to-unit decoder is frozen, and the duration predictor is trained using the loss function from Eq. 3, with λ set to 10. The model is fine-tuned from the pretrained weights from [6], and we fine-tune duration predictor module for 200K iterations. To ensure accurate synchrony evaluation, the vocoder generates the entire audio output at each step. We process one sample at a time and accumulate the gradient to reach a batch size of 32. The model is optimized using AdamW [31] with a learning rate of 2E-4.

5. EXPERIMENTAL RESULTS

5.1. Baselines

We use the latest work of AV2AV [6] as a strong baseline for this work. This is the only open-source AV translation model that has research permissive license. In our experiments with AV2AV, we did not evaluate their full audio-visual outputs. Instead, we use the generated speech by AV renderer and overlay it on the original video, similar to our approach. This was done to ensure fairness to our approach and to avoid making any facial modifications to the original videos. We call this system as *AV2AV-Speech*.

As another baseline (*Synthetic*), we leveraged machine translations via Amazon Translate and generated speech via Amazon Polly[2], overlayed it on the original video to generate dubbed video. We make sure that the speech generated by Polly is of similar duration as that of source via Speech Markers feature in Polly.

5.2. Results

The results presented in Table 2 provide lip-synchrony scores of our proposed method (*Ours*) against two baseline approaches across multiple language pairs (English to Spanish, Portuguese, Italian, and French). The "Source" column contains LSE-C and LSE-D scores for the original video which serves as a reference point for the best achievable lip-synchrony alignment (i.e. upper bound) between the generated speech and the original video.

[1] https://aws.amazon.com/translate/

[2] https://aws.amazon.com/polly/

Language	System	PESQ	BLASER	ASR-BLEU
En-Es	**AV2AV**	1.047	0.798	35.13
	Ours	1.051	0.797	35.18
En-Pt	**AV2AV**	1.043	0.709	28.17
	Ours	1.050	0.708	28.53
En-It	**AV2AV**	1.037	0.732	25.42
	Ours	1.052	0.732	25.71
En-Fr	**AV2AV**	1.041	0.767	21.34
	Ours	1.044	0.764	21.34

Table 3: Speech naturalness and translation quality evaluations of our proposed method (Ours) against AV2AV-Speech baseline across four language pairs (English to Spanish (Es), Portuguese (Pt), Italian (It) and French (Fr).

From the results in Table 2, we see that the *Synthetic* generation of speech, when overlaid on original video, results in the worst alignment. This outcome is anticipated, as the synthetic speech is generated using off the shelf text-to-speech (TTS) technology, which lacks an inherent understanding of temporal information across modalities. While the lip-sync scores of AV2AV model are close to ground truth as shown in [6], the results for the same model differs in ours as we only leverage speech generated via *AV2AV-Speech* and skip the face synthesis. Consequently, this approach face the same problem as *Synthetic* as there is no notion of lip-synchrony between the synthesized speech and original video thus resulting in sub-optimal alignment. This underscores the importance of introducing synchrony constraints in the training process.

Our approach fine-tunes the duration predictor by adding the lip-synchrony and duration losses as described in Section 3.1. *Ours* significantly outperform both *Synthetic* and *AV2AV-Speech* approaches in terms of lip-synchrony scores (p-value < 0.05). This indicates that our methods are effective in enhancing the lip-synchrony of translated speech with the original video content.

While the motivation of our work is to improve lip-synchrony, it should **not** come at the cost of degraded naturalness in speech or translation quality [5]. Table 3 collects results for speech quality as measured by PESQ and translation quality as measured by BLASER-2.0 (reference-free metric) and ASR-BLEU (reference-based metric). We do not observe any degradation across four language pairs in either naturalness or translation quality.

6. ABLATIONS

6.1. Duration Prediction

This ablation study assesses the impact of lip-sync loss (LS loss), duration loss (D. loss), and model initialization on lip-sync performance for English to Spanish translation, measured using LSE-C (higher is better) and LSE-D (lower is better). As shown in Table 4, our model (*LS+D FT*), which incorporates both lip-sync and dura-

System	LS loss	D. loss	From Scratch?	LSE-C	LSE-D
LS+D FT	✓	✓	✗	2.45	10.68
LS+D	✓	✓	✓	1.96	11.03
LS FT	✓	✗	✗	1.76	11.70

Table 4: Ablation study to understand the affect of lip-sync loss (LS loss), duration loss (D. loss) and initializing the model from pre-trained checkpoint for English to Spanish language direction.

Prompt

I want to translate dialogues. Make sure the visemes of the source sentence matches the visemes of target translations. Please generate 3 paraphrases for the following sentence in {language}: "{sentence}" and make sure these dialogues can be spoken in similar duration as that of source. Try to make paraphrases similar length or shorter than the original sentence.

Fig. 2: Prompt for generating paraphrases using Claude 3.0 Sonnet.

tion loss and is fine-tuned with parameters initialized from AV2AV pre-trained checkpoint, achieves the best performance with the highest LSE-C (2.45) and lowest LSE-D (10.68). *LS+D* model, which is randomly initialized achieves worse results in comparison to ours (LSE-C: 1.96, LSE-D: 11.03). *LS FT* model which is fine-tuned only with lip-sync loss without duration loss performs the worst with LSE-C of 1.76 and LSE-D of 11.70. These results indicate that combining both losses and leveraging pre-trained models significantly enhances lip-sync quality in translation tasks.

6.2. Unit-to-Unit Translation with Paraphrasings

We sought to determine whether improved lip-synchrony could be achieved by using a unit-to-unit translation model that generates variations aligned with the original video. Training such a model requires a parallel corpus of translated speech. However, existing datasets, like LRS3, lack parallel source sentences with multiple target translations. To address this, we created multiple translation options from English (the source language) into our target languages (Spanish, Portuguese, Italian, and French). For each English sample, we first generated a target translation using Amazon Translate, which was then refined using a large language model (Claude 3.0 Sonnet) with a specific prompt in Figure 2. Finally, these translations were converted into speech using Amazon Polly.

We leveraged the pre-trained AV2AV encoder-decoder model [6] specifically designed for unit-to-unit translation and fine-tuned it on the (spoken) paraphrased translations. During the fine-tuning process, we ensure that each batch is organized such that four targets (3 paraphrases + 1 original) corresponding to a source input are contained within the same batch. The unit-encoder receives a source language token $\langle L_s \rangle$, which indicates the language to be comprehended, along with the source AV speech units $\mathbf{u}_s = \{u_s^i\}_{i=1}^{T_s}$, where T_s represents the number of units. The unit-decoder then takes a target language token $\langle L_t \rangle$, which determines the output language, and uses its previous predictions to autoregressively predict the next AV speech unit in the target language.

Although this new training approach of unit-to-unit translation model with paraphrases achieve better lip-synchrony scores over our strong system (Table 5), this comes at a cost to translation quality which violates our initial motivation of "not trading off lip-synchrony improvements over speech translation quality and naturalness". We hypothesize that the drop in translation quality is due to LRS3 containing many short sentences (c.f. avg. duration per video in Table 1). While generating paraphrases, LLM changes the meaning of the sentence for e.g. "And that's powerful" is translated as "Y eso es poderoso." (And that is powerful.) and paraphrased as "Es muy fuerte." (It is very strong.) which is not a direct translation of source.

	LSE-D (↑)	LSE-C (↓)	BLASER	ASR-BLEU
Ours	10.68	2.45	0.797	35.18
Ours + Updated TM	10.18	2.91	0.766	32.18

Table 5: Ablation study with the updated translation model (Ours + Updated TM) on translation variations. These results are on the English-Spanish language pair, but we observed similar trends across all four language pairs.

System	LSE-D ↑	LSE-C ↓
Ours + Updated TM	10.18	2.91
Length Match	10.35	2.66
Original Translation	10.67	2.60

Table 6: Comparison of *Ours+Updated TM* with length matching and original translation approaches for the English-Spanish setting.

To determine whether fine-tuning on translations that closely match the original speech length improves lip-synchrony, we selected the best length-matched paraphrase based on audio-visual (AV) units based on the principle of isometric translations producing better dubbing quality [16]. Instead of training on four target translations, we trained the model using only this selected paraphrase. Surprisingly, the results worsened compared to our *Ours+Updated TM* approach, as shown by the *Length Match* results in Table 6 (LSE-D: 10.18 vs. 10.35). Additionally, when we used speech generated via text-to-speech (TTS) from the original translation (i.e., translation from Amazon Translate), we observed a significant degradation in LSE* metrics compared to *Ours+Updated TM* – specifically, LSE-D: 10.18 vs. 10.67.

7. CONCLUSIONS

This study focuses on generating speech that aligns seamlessly with the original video, avoiding the need for facial synthesis and ensuring high-quality dubbing. Our AVS2S framework incorporates lip-synchrony and duration loss to enhance the alignment between speech and lip movements in audio-visual translation models. By concentrating solely on improving lip-synchrony without altering facial features, our approach demonstrates significant improvements. While our approach is effective, further research is required to refine paraphrase generation, explore viseme variations across languages, and extend evaluations. We also aim to explore longer speeches, allowing for more nuanced paraphrasing that closely mirrors the original sentences.

8. REFERENCES

[1] Frederic Chaume, *Audiovisual Translation: Dubbing*, Translation Practices Explained. St. Jerome Pub, Manchester, UK, 1st edition, 2012.

[2] Chenxu Hu, Qiao Tian, Tingle Li, Wang Yuping, Yuxuan Wang, and Hang Zhao, "Neural dubber: Dubbing for videos according to scripts," in *Advances in Neural Information Processing Systems*. 2021, vol. 34, pp. 16582–16595, Curran Associates, Inc.

[3] Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt, "Neural style-preserving visual dubbing," *ACM Transactions on Graphics*, vol. 38, no. 6, pp. 1–13, 2019.

[4] Elisa Perego, David Orrego-Carmona, and Sara Bottiroli, "An empirical take on the dubbing vs. subtitling debate: An eye movement study," *Lingue e Linguaggi*, vol. 19, pp. 255–274, 2016.

[5] William Brannon, Yogesh Virkar, and Brian Thompson, "Dubbing in practice: A large scale study of human localization with insights for automatic dubbing," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 419–435, 2022.

[6] Jeong Yun Choi, Se Jin Park, Minsu Kim, and Yong Man Ro, "Av2av: Direct audio-visual speech to audio-visual speech translation with unified audio-visual speech representation," in *CVPR 2024*, 2024.

[7] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[8] Soumyya Kanti Datta, Shan Jia, and Siwei Lyu, "Exposing lip-syncing deepfakes from mouth inconsistencies," 2024.

[9] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer, "The deepfake detection challenge dataset," *CoRR*, vol. abs/2006.07397, 2020.

[10] Ben Bariach, Bernie Hogan, and Keegan McBride, "Faces of the future: How generative ai is redefining likeness and identity in the age of artificial intelligence," Feb 2024.

[11] G Murphy, D Ching, J Twomey, and C Linehan, "Face/off: Changing the face of movies with deepfakes," *PLoS One*, vol. 18, no. 7, pp. e0287503, 2023.

[12] E Meskys, J Kalpokiene, P Jurcys, and A Liaudanskas, "Regulating deep fakes: legal and ethical considerations," *Journal of Intellectual Property Law & Practice*, vol. 15, no. 1, pp. 24–31, 2020.

[13] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo, "Vasa-1: Lifelike audio-driven talking faces generated in real time," *ArXiv abs/2404.10667*, 2024.

[14] Rongjie Huang, Huadai Liu, Xize Cheng, Yi Ren, Lin Li, Zhe Ye, Jinzheng He, Lichao Zhang, Jinglin Liu, Xiaoyue Yin, and Zhou Zhao, "Av-transpeech: Audio-visual robust speech-to-speech translation," *ArXiv abs/2305.15403*, 2023.

[15] Xize Cheng, Lin Li, Tao Jin, Rongjie Huang, Wang Lin, Zehan Wang, Huangdai Liu, Yejin Wang, Aoxiong Yin, and Zhou Zhao, "Mixspeech: Cross-modality self-learning with audio-visual stream mixup for visual speech translation and recognition," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 15689–15699.

[16] Surafel M Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico, "Isometric mt: Neural machine translation for automatic dubbing," *arXiv preprint arXiv:2112.08682*, 2021.

[17] Johanes Effendi, Yogesh Virkar, Roberto Barra-Chicote, and Marcello Federico, "Duration modeling of neural tts for automatic dubbing," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8037–8041.

[18] Yogesh Virkar, Marcello Federico, Robert Enyedi, and Roberto Barra-Chicote, "Prosodic alignment for off-screen automatic dubbing," 2022.

[19] Proyag Pal, Brian Thompson, Yogesh Virkar, Prashant Mathur, Alexandra Chronopoulou, and Marcello Federico, "Improving isochronous machine translation with target factors and auxiliary counters," 2023.

[20] Alexandra Chronopoulou, Brian Thompson, Prashant Mathur, Yogesh Virkar, Surafel M. Lakew, and Marcello Federico, "Jointly optimizing translations and speech timing to improve isochrony in automatic dubbing," 2023.

[21] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY, USA, 2020, MM '20, p. 484–492, Association for Computing Machinery.

[22] Tianyi Xie, Liucheng Liao, Cheng Bi, Benlai Tang, Xiang Yin, Jianfei Yang, Mingjie Wang, Jiali Yao, Yang Zhang, and Zejun Ma, "Towards realistic visual dubbing with heterogeneous sources," in *Proceedings of the 29th ACM International Conference on Multimedia*, New York, NY, USA, 2021, MM '21, p. 1739–1747, Association for Computing Machinery.

[23] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *International Conference on Learning Representations*, 2021, pp. 2–15.

[24] Minsu Kim, Jeongsoo Choi, Dahun Kim, and Yong Man Ro, "Many-to-many spoken language translation via unified speech and text representation learning with unit-to-unit translation," *arXiv preprint arXiv:2308.01831*, pp. 1–15, 2023.

[25] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 17022–17033.

[26] Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, et al., "Direct speech-to-speech translation with discrete units," *arXiv preprint arXiv:2107.05604*, pp. 3–5, 2021.

[27] J.S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[28] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.

[29] Loïc Barrault and Others, "Seamlessm4t: Massively multilingual and multimodal machine translation," 2023.

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[31] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," 2019.