2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) | 979-8-3503-0261-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSPW59220.2023.10193719

FACE-DUBBING++: LIP-SYNCHRONOUS, VOICE PRESERVING TRANSLATION OF VIDEOS

Alexander Waibel[1,2] *Moritz Behr*[1] *Dogucan Yaman*[1] *Fevziye Irem Eyiokur*[1]
Tuan-Nam Nguyen[1] *Carlos Mullov*[1] *Mehmet Arif Demirtas*[3] *Alperen Kantarci*[3]
Stefan Constantin[1] *Hazım Kemal Ekenel*[3]

[1] Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology
[2] Carnegie Mellon University
[3] Department of Computer Engineering, Istanbul Technical University

ABSTRACT

In this paper, we propose a neural end-to-end system for voice preserving and lip-synchronous video translation. The system is designed to combine multiple component models and produces a video of the original speaker speaking in the target language that is lip-synchronous with the target speech, yet maintains emphases in speech, voice characteristics, and face video of the original speaker. The result is a video of a speaker speaking in another language without actually knowing it. For the evaluation, we present a user study of the complete system and separate evaluations of the single components. Since there is no available dataset to evaluate our whole system, we collect a test set to evaluate our system. The results indicate that our system is able to generate convincing videos of the original speaker speaking the target language while preserving the original speaker's characteristics.

Index Terms— end-to-end video translation, speech translation, text-to-speech, voice conversion, lip generation

1. INTRODUCTION

Speech-to-Speech translation systems have matured in recent years from early prototypes over mobile hand-held translators to fully integrated and operational simultaneous interpreting systems that have been deployed in lecture and video conferencing applications [1, 2, 3, 4, 5]. They have proven quite effective in practical deployments and commercial operations using different delivery mechanisms and modalities appropriate to their use case. In mobile consecutive translation of dialogues (travelers, healthcare providers, humanitarian missions, etc.) individual sentences are translated and the output is commonly synthesized in a target language. Simultaneous interpretation of lectures, movies and video conferences by contrast are best delivered by subtitling [5, 6], as they can be generated simultaneously [7, 8, 9, 10, 11] and do not create distractions during a speech or monologue. Still, when movies or off-line video recordings are to be produced, subtitling is sometimes tiresome and a distraction of its own. Movies, therefore, are sometimes also *dubbed* as an alternate form of delivery, where voice talents act out translated sentences in a target language to replace the original voice. So far such dubbing has been produced only for movies after the fact but it is costly, requires considerable human effort, and the result is frequently not convincing when the original video and the target voice and language do not properly align. One proposed solution to improve is to apply *isometric* human or machine translation [12, 11], where speech translation is performed on an original video source in a manner that optimizes a temporal match between the translator's generated output text and the original video. With isometric translation a better dubbing could thus be achieved, but the dubbed speech from a voice talent (or synthetic voice) in the output language still does not match well with the lip movement and the voice of the original speaker in the original video.

In this paper, we propose a different approach: Rather than inserting translated speech into the original video, we modify the original video in such a way that the resulting video shows lip movements corresponding to the *translated* speech, and the translated synthetic speech in the target language is also generated in a way that is preserving the original speaker's voice characteristics. The result is a more convincing video experience in the *target* language as *lips* and *voice* match *speech* and *speaker*. While this idea had already been proposed before in early work on a face translator [13], the integration was not smooth and unconvincing, and only a different synthetic voice could be generated. To overcome these problems and provide a complete more convincing system, we propose an integrated neural end-to-end system that generates a translated and high-quality lip-synced version of the given video by preserving emphases, prosody, the face and voice characteristics of the original speaker in a real-time with low latency. We present a variation to the FastSpeech 2 TTS model that generates synthetic speech but also permits fine-grained prosodic control for the synthesized speech so as to retain emphasis and prosody of the original speech. The

Authorized licensed use limited to: KIT Library. Downloaded on April 23,2024 at 15:58:14 UTC from IEEE Xplore. Restrictions apply.

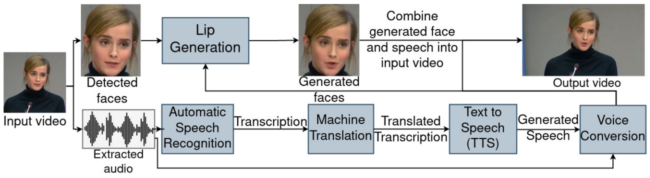

Fig. 1. Visualization of whole pipeline for video translation.

synthesized lips are mapped back to the speaker's voice by our voice conversion model, even though no data from that speaker is available in the target language. We also propose a modified Wav2Lip model to generate the synchronized lips more accurate. Finally, we collect a real-world test dataset to evaluate the components and the overall system.

2. SYSTEM COMPONENTS

In our proposed system, we use the ASR model as a first step to transcribe and detect emphasis from the input video. Next, we feed the transcription to Neural Machine Translation (NMT) and generate the translation with annotated emphasis tokens in the target language. Then, we pass the translation to the TTS model and synthesize audio while preserving emphasis using the labels. To match the speaker's voice, we convert the synthesized audio with voice conversion methods. Meanwhile, the face detection model runs on the video frames to extract faces and the lip generation model obtains consecutive face images and adapted speech to synthesize the output face that should have the synchronized lips. An high-level overview of the pipeline is illustrated in Figure 1.

ASR. In order to transcribe spoken language in offline setting, we train and compare the well-known models such as LSTM [14], Transformer [15], and Conformer [16]. Our LSTM-based model [14, 8] includes 6 bidirectional layers for the encoder and 2 unidirectional layers for the decoder, with 1536 units in each. On the other hand, while the transformer-based model has 24 encoder layers and 8 decoder layers, the Conformer-based model [16] consists of 16 encoder layers and 6 decoder layers. The size of each layer in both Transformer-based and Conformer-based model is 512, while the size of the hidden state in the feed-forward sub-layer is 2048. Besides, we apply the speech data augmentation method [14] to reduce overfitting. In the proposed method, we use Stochastic Layers with a drop rate of 0.5 on both Transformer-based and Conformer-based models to successfully train a deep network [15]. To classify a word as emphasized, we add a binary classifier layer to the end of the network. In the end, we find that the best performance is achieved by ensemble of LSTM-based and Conformer-based sequence-to-sequence model.

Translation. We employ a Transformer model [17] with the *base* configuration, implemented in the NMTGMinor framework [18] to translate from English to German. For preserving and mapping emphasis, we exploit word-alignments from the attention layers to extract source-to-target word alignment. For each emphasized input token, we then determine the matching output token and put emphasis on this corresponding output token. We obtain the word alignment by averaging the normalized attention scores from each head of the final encoder-decoder multihead-attention layer.

TTS. We employ a modified FastSpeech 2 [19] for synthesizing Mel spectrograms of speech for a given text since FastSpeech 2 allows for faster inference times due to its non-autoregressive design. It is based on an encoder-decoder architecture and employs multiple feed-forward Transformer blocks [20] that are made up of stacks of self-attention and TDNN/1D-convolution layers [21]. To make non-autoregressive TTS feasible, FastSpeech 2 employs variance adaptors which provide information on prosody to ease the one-to-many mapping problem inherent to TTS. The three variance adaptors enrich the hidden sequence by adding predicted pitch, duration, and energy information on phoneme-level to the hidden sequence thus helping the decoder by easing the one-to-many mapping problem of TTS. To further ease the training process of the model and make phoneme-level variance prediction possible, the model is given the input text as a sequence of phonemes. Consequently, prior conversion is needed for grapheme inputs. This is done by consulting a pronunciation dictionary and, for words not present in the dictionary, by employing a grapheme to phoneme model trained using the Montreal Forced Aligner [22]. Originally, the predictions of the variance adaptors can only be controlled by parameters for the entire utterance which would not allow for fine-grained prosody control. As we aim to add emphases to the synthesized speech that match the emphases in the original speech, we then add prosody controls at the word-level to the text input by way of Speech Synthesis Markup Language [23] (SSML) tags. Using SSML, we can now add emphasis tags to words in the translation that correspond to words in the original transcript that were emphasized by the speaker. In our system, this happens automatically as the ASR model adds emphasis tags to text sections where emphases were detected. The prosody predictions of the variance adaptors are then modified for the phonemes of that word to create an emphasis in the TTS output. The model varies duration and energy of the respective phonemes as well as increasing or decreasing pitch depending on the originally predicted pitch for the word. Finally, we use the HiFi-GAN vocoder [24] to generate audio wave-forms from the Mel spectrograms generated by the TTS model.

Voice conversion. We aim to revert the generated speech to the original speaker's voice. To accomplish this, we need to employ voice conversion from the synthetic TTS voice back to the original speaker's voice in our original videos. We use VQMIVC (Vector quantization mutual information voice conversion) as a method for this step. VQMIVC uses a straightforward but effective autoencoder architecture to

Authorized licensed use limited to: KIT Library. Downloaded on April 23,2024 at 15:58:14 UTC from IEEE Xplore. Restrictions apply.

perform voice conversion in a way that separates the effects of voice from prosody, content and emphasis. The framework consists of four modules: a content encoder that produces a content embedding from speech, a speaker encoder that produces a speaker embedding (D-vector) from speech, a pitch encoder that produces prosody embedding from speech, and a decoder that generates speech from content, prosody, and speaker embeddings, respectively. Phonetics and prosody are represented through content embedding and prosody embedding. The content embedding is discretized by a vector quantization module and used as target for the contrastive predictive coding loss. A mutual information (MI) loss measures the dependencies between all representations and can be effectively integrated into the training process to achieve speech representation disentanglement. During the conversion stage, the source speech is put into the content, pitch and speaker encoders to extract content, prosody and target speaker embeddings. Finally, the decoder reconstructs the converted speech using the source speech's embeddings. We adapt the pre-trained VQMIVC voice conversion on both German and English datasets to get better performance on both languages. We followed the original paper for the hyper-parameters [25].

Lip generation. We propose to use GAN [26] and design our model with inspiration from [27] by addressing the task as a conditional image generation. First, we propose an audio-guided face generator to synthesize a face image. For this, we obtain Mel spectrogram representation of the audio and provide it to our audio encoder to obtain the embedded features. Meanwhile, we utilize an image encoder to encode the input image. Our input has six channels, namely the depth-wise concatenation of two separate images. While the first image is a face of the corresponding subject from another time sequence, which is crucial to preserve the identity during the generation, the second image is the bottom half-masked ground truth face. Later, the encoded image and audio samples are concatenated along the depth to feed the face decoder in order to synthesize the face image; that is, half-masked area. We further utilize residual connections between the reciprocal layers of image encoder and decoder. This allows us to preserve the identity information. As a discriminator, we employ a binary classifier with a cross-entropy loss and spectral normalization to distinguish real and fake images. Besides, we benefit from a pretrained Syncnet [28, 27] to measure the coherence between the audio input and generated face image to check whether or not prior condition is ensured.

3. EXPERIMENTS

Dataset. For training and evaluation of our ASR models, we used the same datasets in [10]. For the translation, we train the model on 1.8 million sentences of Europarl data [29] and finally finetune on 150,000 sentences of TED data [30] for better adaptation towards spoken language. In order to synthesize speech, we train our TTS model on CSS10 German dataset [31]. Moreover, we train our lip generation model on LRS2 dataset [32] by following the presented training and test setups. Finally, we collected a test dataset to evaluate our end-to-end video translation system since there is no suitable dataset for this purpose. We gathered various videos from the internet to create a test set. Our test set contains 262 different video clips belonging to 25 different speakers. The duration of the test clips is about ten seconds. Please note that we selected all the recording videos in English and evaluated our system as English-to-German video translator.

Data	Libri	Tedlium
Conformer-based	3.0	4.8
Transformer-based	3.2	4.9
LSTM-based	**2.6**	**3.9**
Ensemble	**2.4**	3.9

Table 1. WER results on Libri and Tedlium test sets.

ASR and Translation. For ASR, our ensemble of LSTM-based and Comformer-based sequence-to-sequence models achieve WERs of, respectively, 2.4 and 3.9 on the Libri and Tedlium test sets. As shown in the Table 1, ensemble-based method achieves the best results on Libri test, while it reaches the same performance with LSTM-based approach and surpasses the Conformer-based and Transformer-based methods on TED-LIUM test set. Besides, our translation model attains a translation score of 29.7 BLEU on the IWSLT *tst2010* test set.

TTS. The evaluation of the TTS system was done in two user studies. A first study was conducted to compare the performance of our modified FastSpeech 2 architecture on the LJSpeech dataset with the widely used Tacotron 2 architecture to get a baseline. A second user study was done on our model, which was trained on the German CSS10 dataset, in order to evaluate its performance when applying fine-grained prosody control. For comparison, we synthesized ten texts from the test set of the LJSpeech dataset with both Tacotron 2 and FastSpeech 2, and also used the respective audio samples as ground truth. A group of eight participants was then asked to rate the quality of the audio samples on a scale from 1 to 5. After that, mean opinion scores (MOS) and confidence intervals were calculated. While we obtain 4.21 ± 0.17 score for ground truth data, the Tacatron 2 result is 3.86 ± 0.21. On the other hand, our modified FastSpeech attains 3.87 ± 0.2 performance and performs as well as Tacotron 2. This confirms the results of the FastSpeech 2 evaluation in [19] and suggests that our modifications to FastSpeech 2 did not decrease the quality of the synthesized speech.

For subjective evaluation of the German TTS system and the fine-grained prosody control capabilites of our model, speech was synthesized for texts randomly drawn from the test set of the CSS10 dataset. For ground truth comparison we further chose random audio samples from the test set. This

Authorized licensed use limited to: KIT Library. Downloaded on April 23,2024 at 15:58:14 UTC from IEEE Xplore. Restrictions apply.

Model	Naturalness	Intelligibility
Ground Truth	4.28 ± 0.12	4.82 ± 0.08
Ours	3.59 ± 0.28	4.69 ± 0.09
Ours with Emphasis	3.29 ± 0.16	4.71 ± 0.13

Table 2. MOS and 95% confidence intervals.

Model	Data	Language	LSE-D	LSE-C	FID
Wav2Lip	LRS2	English	6.46	7.78	**4.44**
Ours	LRS2	English	**6.30**	**7.83**	8.86
Wav2Lip	Ours	English	8.35	6.40	**19.62**
Ours	Ours	English	**8.11**	**6.52**	21.15
Wav2Lip	Ours	German	7.93	**7.18**	-
Ours	Ours	German	**7.90**	7.17	-

Table 3. GAN-based Wav2Lip and our model were trained on LRS 2 training set and tested on LRS 2 and our test sets.

time we also compared the quality of the generated speech when using default prosody with the quality of generated speech with added emphases. We conducted this additional comparison only on the German model as this is the model we also use in the final system evaluation. To evaluate the capability of the system to add emphases to the synthesized speech, the chosen text samples were synthesized again, this time with an emphasis added to a random word. To get a more differentiated view on quality differences between unemphasized and emphasized TTS outputs, the participants were asked to rate the audio quality considering two metrics, naturalness and intelligibility on a scale from 1 to 5. Table 2 shows the MOS and confidence intervals for ground truth, unemphasized, and emphasized samples.

Lip generation. We first evaluate the quality of the generated images by using FID score [33]. We further consider the conditional image generation by measuring the synchronization between the generated lip and the audio input using recently proposed novel metrics, LSE-D and LSE-C [27], which are basically distance and confidence scores. In Table 3, we show utilized metrics for LRS2 test set and our proposed test set. The LSE-D and LSE-C results on both test sets show that our model is able to surpass Wav2Lip [27] in terms of providing synchronized lips. Please note that both models were trained only on LRS 2 training set and tested on both datasets. The performance on our test set —unseen data and unknown language— indicates that our model has an effective generalization capacity in terms of the visual input and language than Wav2Lip model, and is robust against real-world challenges.

System evaluation. We conducted a user study with 25 participants and presented the results in Table 4. In this way, we aim to investigate the performance of the system by considering several different aspects: 1) realism of the generated face, 2) naturalness of the generated voice, 3) intelligibility of the speech, 4) synchronization quality of the speech and lip, 5) accuracy of the translated speech given the original English transcript. We asked one question for each aspect. In the user study, we randomly chose 80 videos from the 262 videos of our dataset and showed the translated and lip-synced results to the participants with the transcribed original speech as well as the five questions. Sample evaluation videos are available here [1].

Measurement	High	Medium	Low
Reality of the faces	59%	29%	12%
Synchronization	57.9%	31.6%	10.5%
Translation accuracy	29%	56%	15%
Naturalness	3.36 ± 0.98		
Intelligibility	4.24 ± 0.86		

Table 4. Subjective evaluation on our proposed test set.

The results in Table 4 indicate that participants rate the quality of the generated faces as high. Similarly, the majority of the answers state that our system provides accurate synchronization in the generated videos. Although a majority of the answers indicate that there are minor mistakes in the translated text, only 15% of the answers find the results inaccurate. Moreover, we demonstrate naturalness and intelligibility results in the last two rows of the same table as a MOS metric and they state that our system is successful in providing naturalness and intelligibility in the generated video. The evaluation showed that the faces and lip-synchronization in the generated videos were believable and the generated speech was well intelligible. However, we observed occasional problems with naturalness of the generated speech and inaccuracies in the translations due to lacking punctuation in the transcripts generated by the ASR model. Moreover, the lip-syncing model showed slight issues with bearded faces and also had some quality problems that must be addressed.

4. CONCLUSION

Given a video of a speaker, our proposed system can generate a convincing output video of that speaker uttering a translation of the original speech while adapting lip movements to the new audio and preserving voice characteristics. Additionally, emphases are preserved by emphasis detection in the ASR model, and modifications to the used FastSpeech 2 TTS model allow fine-grained prosody control which is used to create corresponding emphases in the synthesized speech. We also employed voice conversion to allow us to have the same speaker characteristic in the output video. Lip generation model manipulates the lips with respect to given translated audio to provide realistic synchronization in the final video. The detailed experiments indicated that we achieve accurate modules for each task and acceptable performance in the final system. Ongoing work is devoted to improving speed and latency of the components to obtain a better architecture.

[1] https://videospeechtranslation.github.io

Authorized licensed use limited to: KIT Library. Downloaded on April 23,2024 at 15:58:14 UTC from IEEE Xplore. Restrictions apply.

5. REFERENCES

[1] Alex Waibel et al., "Janus: a speech-to-speech translation system using connectionist and symbolic processing strategies," in *ICASSP*, 1991, pp. 793–796.

[2] Alon Lavie et al., "Janus-iii: Speech-to-speech translation in multiple languages," in *ICASSP*, 1997.

[3] Muntsin Kolss et al., "Simultaneous german-english lecture translation.," in *IWSLT*, 2008.

[4] Matthias Eck et al., "Jibbigo: Speech-to-speech translation on mobile devices," in *SLT*, 2010, pp. 165–166.

[5] Markus Müller et al., "Lecture translator-speech translation framework for simultaneous lecture translation," in *NAACL*, 2016, pp. 82–86.

[6] Naomi Aoki Waibel et al., "Hybrid, offline/online speech translation system," Aug. 30 2016, US Patent 9,430,465.

[7] Jan Niehues et al., "Low-latency neural speech translation," *arXiv preprint arXiv:1808.00491*, 2018.

[8] Thai-Son Nguyen et al., "Super-human performance in online low-latency recognition of conversational speech," *arXiv preprint arXiv:2010.03449*, 2020.

[9] Ngoc-Quan Pham et al., "Multilingual speech translation kit@ iwslt2021," in *IWSLT*, 2021, pp. 154–159.

[10] T. N. Nguyen et al., "Kit's iwslt 2021 offline speech translation system," in *IWSLT*, 2021, pp. 125–130.

[11] Antonios Anastasopoulos et al., "Findings of the iwslt 2022 evaluation campaign," in *IWSLT*, 2022.

[12] Surafel M Lakew et al., "Isometricmt: Neural machine translation for automatic dubbing," *arXiv preprint arXiv:2112.08682*, 2021.

[13] Max Ritter et al., "Face translation: A multimodal translation agent," in *AVSP*, 1999.

[14] Thai-Son Nguyen et al., "Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation," in *ICASSP*, 2020, pp. 7689–7693.

[15] Ngoc-Quan Pham et al., "Very deep self-attention networks for end-to-end speech recognition," *arXiv preprint arXiv:1904.13377*, 2019.

[16] Anmol Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[17] Ashish Vaswani et al., "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[18] Ngoc-Quan Pham et al., "Relative Positional Encoding for Speech Recognition and Direct Translation," in *Proc. Interspeech 2020*, 2020, pp. 31–35.

[19] Yi Ren et al., "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *ICLR*, 2020.

[20] Yi Ren et al., "Fastspeech: Fast, robust and controllable text to speech," *NeurIPS*, vol. 32, 2019.

[21] Alexander Waibel et al., "Phoneme recognition using time-delay neural networks," *IEEE transactions on acoustics, speech, and signal processing*, vol. 37, no. 3, pp. 328–339, 1989.

[22] Michael McAuliffe et al., "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, 2017, vol. 2017, pp. 498–502.

[23] Paul Taylor and Amy Isard, "Ssml: A speech synthesis markup language," *Speech communication*, vol. 21, no. 1-2, pp. 123–133, 1997.

[24] Jungil Kong et al., "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *NeurIPS*, vol. 33, 2020.

[25] Disong Wang et al., "Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," *arXiv preprint arXiv:2106.10132*, 2021.

[26] Ian Goodfellow et al., "Generative adversarial nets," *NeurIPS*, vol. 27, 2014.

[27] KR Prajwal et al., "A lip sync expert is all you need for speech to lip generation in the wild," in *ACM International Conference on Multimedia*, 2020, pp. 484–492.

[28] Joon Son Chung and Andrew Zisserman, "Out of time: automated lip sync in the wild," in *ACCV*. Springer, 2016, pp. 251–263.

[29] Philipp Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of machine translation summit x: papers*, 2005, pp. 79–86.

[30] Mauro Cettolo et al., "Wit3: Web inventory of transcribed and translated talks," in *EAMT*, 2012, pp. 261–268.

[31] Kyubyong Park and Thomas Mulc, "Css10: A collection of single speaker speech datasets for 10 languages," *Interspeech*, 2019.

[32] T. Afouras et al., "Deep audio-visual speech recognition," in *arXiv:1809.02108*, 2018.

[33] Martin Heusel et al., "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS*, vol. 30, 2017.

Authorized licensed use limited to: KIT Library. Downloaded on April 23,2024 at 15:58:14 UTC from IEEE Xplore. Restrictions apply.