

Face-Dubbing++: Lip-Synchronous, Voice Preserving Translation of Videos

Alexander Waibel^{1,2} Moritz Behr¹ Fevziye Irem Eyiokur¹
Dogucan Yaman¹ Tuan-Nam Nguyen¹ Carlos Mullov¹ Mehmet Arif Demirtas³
Alperen Kantarcı³ Stefan Constantin¹ Hazım Kemal Ekenel³
¹Karlsruhe Institute of Technology, ²Carnegie Mellon University, ³Istanbul Technical University
{firstname.lastname}@kit.edu, {demirtasm18, kantarcia, ekenel}@itu.edu.tr

Abstract

In this paper, we propose a neural end-to-end system for voice preserving, lip-synchronous translation of videos. 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, face video of the original speaker. The pipeline starts with automatic speech recognition including emphasis detection, followed by a translation model. The translated text is then synthesized by a Text-to-Speech model that recreates the original emphases mapped from the original sentence. The resulting synthetic voice is then mapped back to the original speakers' voice using a voice conversion model. Finally, to synchronize the lips of the speaker with the translated audio, a conditional generative adversarial network-based model generates frames of adapted lip movements with respect to the input face image as well as the output of the voice conversion model. In the end, the system combines the generated video with the converted audio to produce the final output. The result is a video of a speaker speaking in another language without actually knowing it. To evaluate our design, we present a user study of the complete system as well as separate evaluations of the single components. Since there is no available dataset to evaluate our whole system, we collect a test set and evaluate our system on this test set. 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. The collected dataset will be shared.

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 [14, 16, 26, 31, 36, 62, 63].

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, health-care 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 [36, 64], as they can be generated simultaneously [2, 38, 40, 41, 46] and do not create distractions during a speech or monologue. Still, when movies or offline 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 don't properly align. One proposed solution to improve on these problems is to apply *isometric* human or machine translation [2, 30], 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 [54], the integration was not smooth and unconvincing, and only a different synthetic voice could be generated. Important component technologies, however, have advanced considerably so that a complete more convincing integrated face translator architecture can now be realized that produces a more realistic and convincing experience. These include large vocabulary low latency simultaneous speech recognition and translation systems [36], voice conversion [65] and various forms of video manipulation [50] and lip syncing [11] methods. In the following, we propose an architecture that builds on these advances for an end-to-end speech translation system with voice conversion and lip synchronization that is able to take videos of English-speaking subjects in real-time and generates videos of these speakers with translated German audio and adapted lip movements while also preserving the original speaker’s voice characteristics, prosodic cues and emphases.

In the following, we do not only develop and explore the different component models, e.g., Automatic Speech Recognition (ASR), Machine Translation, Text-to-Speech Generation (TTS), Voice Conversion, and Lip Generation, but also investigate how to employ these models together to provide an effective and accurate end-to-end video experience to translate speech to a target language and generate the synchronized lips with respect to the translated audio data. Several challenging issues must be addressed. First, we need to provide robust ASR processing to prevent any loss in the original content but preserving details in prosody for better naturalness downstream. Second, we should have an effective translation system to translate the transcribed text from input language to the target language without error and missing content while moving emphasis information recognized during ASR to the appropriate parts of the translated speech. Another important dimension is the ability to generate natural, synthetic audio from the translated speech in another language, but preserving the original speaker’s voice. For this, a TTS system is designed to work accurately from translated text data while providing means for fine-grained prosody control. These means of prosody control are then used to add emphases to the generated speech that match the emphases in the original speech. After speech is generated, we must use voice conversion to adapt the TTS output back to the input speaker’s voice, since a TTS is trained on a single or multiple but fixed speakers and thus cannot generate speech with arbitrary voices. During this adaptation, voice conversion must not cause any degradation in speech quality. Finally, we need to generate the lips with respect to the translated, voice-converted speech. During this, speaker identity must be preserved which means the image generation model should not cause any degradation on the face and lips while generating output. Last but not the least, we need to run all these models sequen-

tially and in a pipelined fashion using the outputs of previous models as an input to the next with minimum delay and without degradation in performance of each model in order to provide a robust end-to-end system.

Our multimodal system includes two pipelines: a video pipeline for face detection and lip synchronization, and an audio pipeline for speech recognition, translation, speech synthesis, and voice conversion. The desired output of the audio pipeline is audio of the original speaker uttering a translation of the speech in the input video with properly aligned emphases if any are present in the original audio. This is achieved by pipelining multiple models. First, our ASR model with emphasis detection creates a transcript of the original speech with additional emphasis information. Then, the English transcript is translated to German by our translation model while any emphasis information is moved to the corresponding parts of the German translation. Now, our TTS model synthesizes German speech with appropriate emphases for the given translation and the voice conversion model adapts the synthesized speech to the voice characteristics of the original speaker. Meanwhile, the video pipeline gets the input video frames to detect the speaker’s face in them. Finally, the lip generation module employs the generated speech and detected faces to synthesize new frames of the speaker’s face with lips that are synchronized to the generated speech. To evaluate our system, we conducted comprehensive experiments to evaluate the performance of each module as well as the entire system.

To assess the effectiveness of the resulting system, we collected a test set that contains 262 videos belonging to 25 different speakers. We carried out a user study to study different aspects of output quality including intelligibility and naturalness of speech, synchronicity of lips and audio, and the credibility of the generated faces in the video.

In this paper, we propose a novel architecture that combines recent advances and new techniques in an end-to-end system that achieves the dream of language transparent communication, i.e. creating a video communication experience (in audio and video) between people speaking different languages that removes the language barrier. More specifically, :

- We propose an integrated neural end-to-end system to perform automatic video translation that creates the illusion of a speaker speaking another language. Given the video of a speaker, a translated and high-quality lip-synced version of the video is generated that preserves emphases, prosody, the face and voice characteristics of the original speaker.
- A real-time, low latency end-to-end speech translation system capable of translating speech from many languages to text in many others (subtitling) is extended

for synthetic speech output.

- 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.
- A voice conversion module is developed and deployed that maps the synthetic speech back to the voice of the original speaker, even though no data from that speaker is available in the target language.
- We develop a real-world dataset to evaluate the components and the overall system. It contains 262 videos of 25 different speakers. The dataset will be shared upon publication.

2. Related Work

2.1. ASR and MT

Attention-based models based on sequence-to-sequence (S2S) [3, 37, 40, 49] are currently one of the top-performing approaches to end-to-end ASR and MT. A significant amount of study has already been spent to improving the performance of S2S models. Attention-based S2S models, which use a neural network architecture to approximate the direct mapping from the input to the textual transcript, have become a very efficient approach for building high performance speech recognition systems or machine translation systems, with a very low real-time factor and a significantly lower word error rate in batch processing on GPUs. The S2S technique has the benefit of simplifying the training of a full end-to-end system, hence hiding the knowledge of complicated components, as in statistical ASR or MT systems. The detail of our S2S ASR and MT system is describe in Section 3.1.

2.2. TTS

Generating text from speech is a mature research field. Recent developments show that here too that deep learning approaches are effective to generate superior quality speech when compared with traditional approaches. Tacotron [67], FastSpeech [52] and others now outperform traditional approaches as general-purpose TTS systems and are trainable on raw speech data with transcripts. Since TTS is a hard task due to its inherent one-to-many mapping problem, most modern TTS models are using Mel spectrograms as an intermediate target. The one-to-many mapping problem in TTS refers to the fact that for a given text, there are a large number of possible audio sequences that can be considered fitting TTS outputs as there are many valid variations in voice, prosody, and background noise. To turn the Mel spectrogram outputs into audio waveforms, a vocoder model is

subsequently used. Tacotron 2 [56], and many subsequently published variations of it, remain widely used TTS models. Its architecture, an encoder-decoder sequence-to-sequence model based on recurrent units, has the drawback that it is auto-regressive, which makes it hard to parallelize the inference process. To accelerate inference, multiple alternative TTS models with non-autoregressive architectures have been proposed. One of these is FastSpeech 2 [52] which does not employ recurrent units, while providing slightly better audio quality than other state-of-the-art models like Tacotron 2 [56], as evaluated by a survey in the original paper. This also makes it easier to run it incrementally so as to minimize resulting latency. We have modified the FastSpeech 2 architecture to provide fine-grained prosodic control to be able to use information about emphases from the original speaker’s voice during speech synthesis.

2.3. Voice Conversion

The purpose of the Voice Conversion module is to make the resulting voice of the speaker in the target language sound like the original source speaker’s voice. Classical Gaussian Mixture Model-based strategies had been proposed and performed well, but modern Artificial Neural Network-based techniques outperform them. GAN, VAE, and seq2seq architectures have been utilized to overcome voice conversion challenges. Voice Conversion systems can be configured in a variety of ways, including one-to-one, one-to-many, many-to-many, any-to-any, and so on. The most challenging voice conversion scenario is given by any-to-any Voice Conversion systems, to convert any source voice to any target speaker, even one not seen in the training data. This has been attempted by several previous architectures such as VQMIVC [65], AutoVC [51], Adain [9], FragmentVC [32]. According to r benchmark comparisons, the VQMIVC is one of the best any-to-any voice conversion systems. We describe the VQMIVC in detail in Section 3.4. Finally, we train VQMIVC for converting our TTS output back to the voice of our input speaker.

2.4. LipSync Video Generation

Perhaps the earliest attempt at lip generation from text in a foreign language was presented by Ritter et al. [54]. In their work video was synthesized for speakers speaking another language with lip movement synchronous to the speech of the other language. However, rendering smooth lip motion and integration in a face for convincing natural videos was not yet possible. Articulation still appeared jumpy and unnatural and voices were synthetic and differed from the input speaker. Improvements in lip and face synthesis were necessary to create more natural synthetic videos reflecting the original speaker.

Neural techniques have re-energized new efforts and enabled considerable advances to synthesize video based on

arbitrary voice track or text. Many initial efforts were mostly focused on speaker-dependent approaches [15, 28, 54, 57, 59], but more recent efforts [7, 23] present fully speaker and language independent approaches. [68] proposed X2Face, which can regenerate the reference video using a variety of modalities, including an input clip or another video to be used as the pose reference. [8] separated the audio embedding network from the video generation network to reduce error accumulation. They proposed using attention mechanisms in video generation to achieve higher visual quality than previous methods. [72] used language-specific adversarial classifiers to disentangle audio-visual embedding to increase lip-sync quality. [24] built upon the architecture in [23] by implementing a language-independent lip synchronization discriminator. [61] employed a noise generator in addition to audio and identity encoders in order to capture minor changes in facial expressions that are not captured by the audio. [50] proposed the use of a pretrained classifier as the expert discriminator in [24] that provides supervision on the lip-sync accuracy. In our work, we inspire from [50] to design a lip generation model due to its outperforming lip-sync performance.

Recent studies focused on not only the lip synchronization but also providing variance in the head pose and head movements of the subject. [70] proposed a new GAN-based model that captures implicit attributes related to head pose in addition to lip-sync information. [73] introduced an additional pose source as input to add realistic movement to lip-synced talking heads. Another research direction is to utilize a single reference image instead of a short clip to synthesize videos [66, 71, 74]. Also, 3D model-based approach [29] and NeRF-based methods [19, 33, 69] are presented to allow for head or body rotation.

3. System Components

Our proposed system contains five different modules which are ASR, machine translation, TTS, voice conversion, and lip generation. The system takes an input video and then extracts the audio and video frames. While the ASR model uses extracted audio to transcribe it and detect emphases in it, the translation model receives the transcribed text and detected emphases to translate it to the target language and move the emphases to the appropriate words in the translation. Afterwards, the TTS model generates new audio with appropriate emphases using the translated text emphasis information and sends the output to the voice conversion model. Later, voice conversion adapts the TTS output to the speaker’s voice and provides the output to the lip generation model. Meanwhile, the face detection model runs on the video frames to extract faces. In the end, the lip generation system obtains consecutive face images and generated speech, which is the output of the voice con-

version system, to synthesize the output face that should have the synchronized lips. An high-level overview of the pipeline is illustrated in Figure 1.

3.1. ASR

First, we trained a sequence-to-sequence ASR model to transcribe audio of English (or other language) speech. At our laboratories, three architectures are under investigation: A long short-term memory (LSTM) based model, a Transformer, and a Conformer LSTM-based model. LSTM-based [39] models include 6 bidirectional layers for the encoder and 2 unidirectional layers for the decoder, with 1536 units in each. They have delivered superior recognition performance on the Switchboard conversational speech benchmark task [40]. The Transformer-based model proposed in [48] feature 24 encoder layers and 8 decoder layers. The Conformer-based model [18] consists of 16 encoder layers and 6 decoder layers. The size of each layer in both the Transformer-based and the Conformer-based models is 512, while the size of the hidden state in the feed-forward sub-layer is 2048. As explained in [39], the speech data augmentation approach was employed to reduce over-fitting. Also recent work on factorizing multilingual models delivered considerable improvements in view of broad multilingual expansion [47]. In the present implementation we used Stochastic Layers with a dropout rate of 0.5 on both Transformer-based and Conformer-based models to successfully train a deep network [48]. To classify a word as emphasized, we add a binary classifier layer to the end of the network. The ensemble of LSTM-based and Conformer-based sequence-to-sequence model provided the best results.

3.2. Translation

We translate from English to German (and, indeed, to many other languages) using a neural sequence-to-sequence model. More specifically, we employ a Transformer [60] model with the *base* configuration as described by [60], implemented in the NMTGMinor framework [45]. We train the model on 1.8 million sentences of Europarl data [25] and finally finetune on 150,000 sentences of TED data [6] for better adaptation towards spoken language.

For emphasis translation, we extract a 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. The word alignment we obtain by averaging the normalized attention scores from each head of the final encoder-decoder multihead-attention layer:

$$\alpha_{ji} = \frac{1}{h} \sum_{k=0}^h \alpha_{ji}^k \tag{1}$$

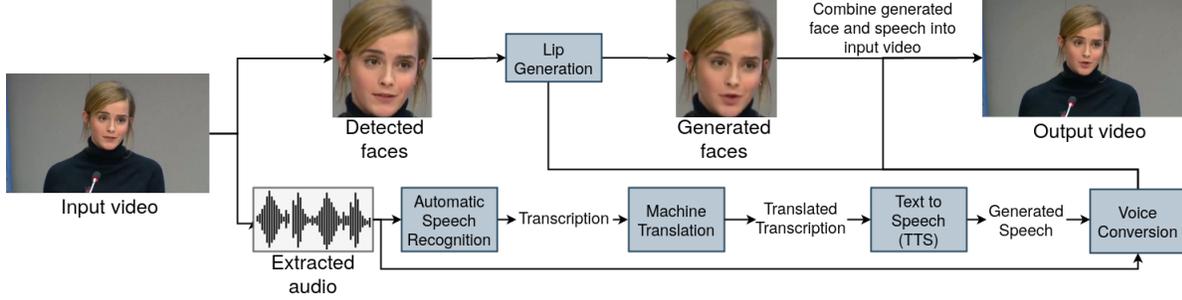


Figure 1. Pipeline of the proposed end-to-end speech-translated lipsync-video generation system. Our system first obtains the audio data from the video input and then extracts Mel spectrogram representation of the audio data. Afterwards, automatic speech translation system provides transcription of the audio to the machine translation system. Later, we acquire the translated text data and send it to the text-to-speech generation system to synthesize the output audio. In order to generate the speech with the same voice of the speaker in the input data, we utilize a voice conversion model and make the synthesized output speech the same with the speaker’s voice. Meanwhile, face detector captures faces from each frame and then we provide these faces to the lip generation model in conjunction with the synthesized speech. In the end, we create the video with the synthesized frames and combine with the generated speech to achieve the same video with the input video that has translated speech and synchronized lips.

$$\alpha^k = \text{softmax} \left(\frac{(QW_k^Q)(KW_k^K)^T}{\sqrt{d}} \right) \quad (2)$$

where $h = 8$ is the number of attention heads, $d = 512$ is the model size, and Q, W, W^K and W^Q follow the description in [60]. For each emphasized input token s_i emphasis is thus put on the output token t_j with $j = \text{argmax}_{k=1..|T|}(\alpha_{ki})$.

3.3. TTS

We are using a modified FastSpeech 2 [52] model for synthesizing Mel spectrograms of speech for a given text. We chose FastSpeech 2 over other popular TTS models like Tacotron 2 [56] as FastSpeech 2 allows for faster inference times due to its non-autoregressive design. Its architecture is based on an encoder-decoder architecture and employs multiple feed-forward Transformer blocks [53] that are made up of stacks of self-attention and TDNN/1-D-convolution layers.

To make non-auto-regressive 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 not as a sequence of graphemes but rather 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 [34].

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 [58] (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 [27] to generate audio wave-forms from the Mel spectrograms generated by the TTS model.

3.4. Voice Conversion

Following TTS in a standardized voice in the target language with the prosody projected from the source speech, 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

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 encoder and pitch encoder to extract content embedding and prosody embedding. To extract the target speaker embedding, the target speech is sent into the speaker encoder. Finally, the decoder reconstructs the converted speech using the source speech’s content embedding and prosody embedding and the target speech’s speaker embedding. We adapt the pretrained VQMIVC voice conversion on both German and English datasets to get better performance on both languages. Our VQMIVC model is fine-tuned with the same hyper-parameters as in the original papers. The evaluation of VQMIVC is presented in [65].

3.5. Lip Generation

We address the lip generation task as a conditional generative adversarial network-based [17, 35] image generation, since our goal is to generate lips with respect to the audio input and face input in order to make the generated lips synchronized with the audio. We design our GAN model with inspiration from [50]. First of all, we propose an audio-guided face generator G to synthesize a face image that is synchronized with the audio. For this, we first obtain Mel spectrogram representation of the audio data. Afterwards, we provide an audio sequence as a sequence of Mel spectrogram to our audio encoder as an input. The audio encoder is responsible for embedding the audio input in order to extract the embedded feature representation. Meanwhile, we utilize an image encoder to encode the input image. Our input image has six channels, namely the depth-wise concatenation of two separate images. While the first three channels contain a face of the corresponding subject from another time sequence or from another video of the same subject, namely reference image x_r , the second image is the masked version of the ground truth face, x_m . The task is to generate the masked area of x_m with respect to the audio sequence. We basically mask the half-bottom part of the face image. Since preserving the identity and details of the

face to improve the realism of the final image are crucial, the reference image x_r is necessary to inject these details to the G while the final face image is being generated. Otherwise, it would be challenging for our generator to preserve the identity and the details in the bottom part of the image. After we acquire the audio and face feature representations from audio and image encoders, we concatenate them along the depth. We further feed the face decoder with this concatenated feature representation.

We further utilize residual connections between the reciprocal layers of the image encoder and image decoder networks in our generator G . These connections allow us to transmit the output of encoder’s layers to the decoder’s layers in order to transfer the crucial details and identity of the input face images. We utilize ReLU activation function in our generator with instance normalization layers.

For the discriminator, we employ a binary classifier with a cross-entropy loss to distinguish real and fake images. This discriminator is responsible for the quality and realism of the generated image. In the discriminator, we benefit from spectral normalization to provide more stable training by normalizing the gradients. Besides, we employ Leaky ReLU and Instance normalization in discriminator. In addition to this, we must also control whether the prior condition is provided in the generated image as it is proposed in [50]. Therefore, we utilize a pretrained synchronization model [11, 50] to evaluate the coherence between the conditional input audio and the output face image. This synchronization network is also a binary classifier that classifies the image to produce output whether synchronization is provided or not. The whole lip generation model is illustrated in Figure 2.

To train our system, we employ a large-scale Oxford-BBC Lip Reading Sentence 2 dataset (LRS2) [1, 10, 12]. We feed our image generator with a set of five consecutive frames. We further send the audio data to audio encoder after we obtain Mel spectrogram representation of the corresponding audio sequence. During the experiments, we follow the proposed data splits to train, validate, and test our model. In order to calculate synchronization loss, we directly use the pretrained lip synchronization model [50], and we do not update this model during the training. Our overall loss function is as follows:

$$L = L_{cGAN} + \alpha * L_{img} + \beta * L_{sync} \quad (3)$$

where L_{cGAN} is a conditional adversarial loss, L_{img} is an image reconstruction loss, $\|y - y'\|$, that calculates the L1 distance between target face image and the generated face image in the pixel space. L_{sync} is a synchronization loss that provides feedback to the generator whether the synchronization between the lip and the audio input is able to be provided in the generated face image. α and β are coefficients that alter the effect of image reconstruction loss

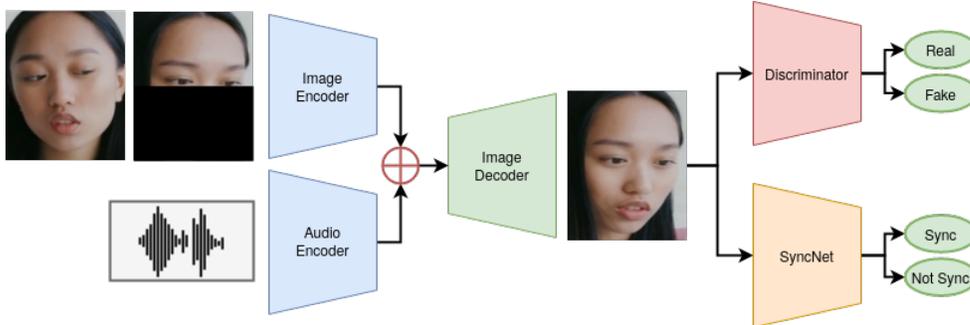


Figure 2. Illustration of the proposed lip generation model. We have two encoders, namely audio encoder and image encoder, and one face decoder to generate face images. After we extract features from the audio and the image, we concatenate them along the depth axis to provide input to the face decoder. Besides, we have a discriminator network to evaluate the quality of the generated face images and decide whether they are real or fake. Finally, we have a pretrained synchronization network that is classifying the generated face image to determine whether it is synchronized with the input audio. Please note that we train the whole video pipeline as end-to-end.

and synchronization loss on the total loss. According to the experimental results, we find the best α and β coefficients as 1 and 0.05.

4. System Integration

In order to combine the multitude of models for ASR, translation, TTS, voice conversion, and lip generation into a single system, we chose a cascade architecture. A diagram of the high-level architecture of our system is shown in Figure 1. Initially, the audio of the given video is extracted and converted to the expected waveform format of the ASR module which then creates an English transcription of the input speech with additional information regarding detected emphases. Then, the translation module produces a German translation of that transcript, including SSML tags for emphases at the parts of the text that correspond to words in the original English transcript that were marked as emphasized by the ASR module. Subsequently, the TTS module is given this translated text and generates German speech with emphases according to the SSML tags. The resulting Mel spectrogram is turned into a waveform file by the HiFi-GAN vocoder. Afterwards, the final audio is created by the voice conversion module which gets the waveform of German speech that the vocoder produced as input and uses the original English audio of the input video as target speaker. Meanwhile, the video pipeline starts by detecting faces in the input video to provide consecutive face images to the lip generation model. Besides, the lip generation module is given the speech produced by the voice conversion model to generate face images with the modified lips. In the end of the system, the generated faces and generated speech are combined to create the final output video. This whole pipeline allows us to acquire a video with translated speech of the original speaker in the target language and the adapted lips by only providing an arbitrary video.

5. Experimental Results

5.1. Dataset

We used various datasets to train and evaluate our models. Besides, we collected a test set to measure the performance of the entire system.

ASR and MT training dataset For training and evaluation of our ASR models, we used Mozilla Common Voice v6.1 [4], Europarl [25], How2 [55], Librispeech [42], MuST-C v1 [13], MuST-C v2 [5] and Tedlium v3 [20] datasets. We also collected the text parallel training data provided by WMT 2019, 2020, 2021 for training MT consisting of a total of 69.8 million sentences as shown on the right side of Table 1.

CSS10 German dataset [44]. CSS10 is a collection of single speaker speech datasets that contain ten different languages. It includes short audio clips and their aligned text data. Since we aimed to generate the audio in German, we utilized the CSS10 German dataset to train our TTS model as it provides 17 hours of high-quality single speaker audio data which is enough to train a single speaker TTS model.

LRS2. We employed the Oxford-BBC Lip Reading Sentences 2 (LRS2) dataset to train our lip generation model and also evaluate its performance. We followed the presented train, validation, and test setups to train the model as well as evaluate the performance. The training set contains 45839 utterances, while validation and test sets include 1082 and 1243 utterances respectively.

Our dataset. Since there is no suitable dataset to test our end-to-end video translation system in the literature, we collected 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 all speakers speak English since we evaluated our system for the combination of English input and German output.

Table 1. Summary of the English datasets used for speech recognition (left) and machine translation (right)

Corpus	Utterances	Speech data [h]	Dataset	Sentences
A: Training Data				
Mozilla Common Voice	1225k	1667	TED Talks (TED)	220K
Europarl	33k	85	Europarl (EPPS)	2.2MK
How2	217k	356	CommonCrawl	2.1M
Librispeech	281k	963	Rapid	1.21M
MuST-C v1	230k	407	ParaCrawl	25.1M
MuST-C v2	251k	482	OpenSubtitles	12.6M
Tedlium	268k	482	WikiTitle	423K
B: Test Data			Back-translated News	26M
Tedlium	1155	2.6		
Librispeech	2620	5.4		

5.2. Evaluation metrics

WER and BLEU The word error rate (WER) is a common metric for measuring speech recognition performance. The Levenshtein distance at the word level is used to calculate the WER. The WER of Librispeech test set represents the ASR’s performance on read speech, while the WER of Tedlium test set represents the ASR’s performance on spontaneous speech. The BLEU, or Bilingual Evaluation Understudy, is a score that compares a candidate translation of text against one or more reference translations [43].

LSE-D and LSE-C [50]. Since the FID, SSIM, and PSNR are not able to evaluate the synchronization of the lips and the synchronization is a crucial key-point in the lip generation task in addition to the quality of the generated face images, using Lip-Sync Error-Distance (LSE-D) and Lip-Sync Error-Confidence (LSE-C) provide more reliable representation about the synchronization. Therefore, as it is proposed in [50], we utilized LSE-D and LSE-C metrics to evaluate the synchronization performance of our lip generation model.

FID [21]. In order to evaluate the quality of the generated face images, we employed FID score [21] by providing the manipulated face images. Thus, FID basically calculates the distance between real samples and generated samples in the feature space. For this, Inceptionv3 image classification model, that was trained on ImageNet dataset, is utilized to extract features. In this metric, lower score indicates better quality for the generated images.

User study. For the evaluation of TTS model as well as the whole system there are no widely accepted computable quality metrics. So in order to evaluate the TTS model and the whole system, we conducted user studies and asked participants to evaluate the performance in several different aspects.

5.3. ASR and Translation

Our ASR and translation models are evaluated by employing computable metrics on standard datasets. For ASR, our ensemble of LSTM-based and Conformer-based sequence-to-sequence model achieves WERs of respectively 2.4 and 3.9 on the Libri and Tedlium test sets. In Table 2, we present the results of Conformer-based, Transformer-based, LSTM-based, and ensemble-based approaches. According to the table, ensemble-based method achieves the best results on Libri test, while it reaches the same performance with LSTM-based approach and surpass the Conformer-based and Transformer-based methods on TED-LIUM test set. Therefore, we decide to use our ensemble-based approach in the final proposed system. Besides, our translation model attains a translation score of 29.7 BLEU on the IWSLT *tst2010* test set.

Table 2. WER results on Libri and Tedlium test sets. While we obtain the best result with ensemble-based method on Libri dataset, we get the best results with ensemble-based and LSTM-based methods on Tedlium dataset.

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

5.4. TTS

We trained our TTS model on the CSS10 German dataset [44] which is a single-speaker dataset consisting of nearly 17 hours of German speech and on the LJSpeech [22] dataset, an English single-speaker dataset consisting of ap-

Table 3. MOS and 95% confidence intervals for ground truth samples and TTS syntheses by Tacotron 2 and modified FastSpeech 2.

	MOS
Ground Truth	4.21 ± 0.17
Tacotron 2	3.86 ± 0.21
Modified FastSpeech 2	3.87 ± 0.2

proximately 24 hours of speech. Montreal Forced Aligner was used to transform the grapheme inputs of the datasets to phoneme sequences and generate the text-audio alignments needed for training the variance adaptors. Training was done on a server with an Intel 4124 CPU, 32 gigabytes of memory, and a single NVIDIA RTX Titan GPU and took approximately 72 hours. A pretrained universal HiFi-GAN model was used as vocoder, no finetuning was necessary.

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 with Tacotron 2, we synthesized ten texts from the test set of the LJSpeech dataset with both Tacotron 2 and FastSpeech 2. For ground truth comparison we used the respective audio samples. 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. Table 3 shows the MOS and confidence intervals results from this survey. As the results show, our modified FastSpeech 2 model performs as well as Tacotron 2. This confirms the results of the FastSpeech 2 evaluation in [52] 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 capabilities 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 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 group of eight participants asked to rate the audio qual-

Table 4. MOS and 95% confidence intervals for ground truth and TTS samples.

	Naturalness	Intelligibility
Ground Truth	4.28 ± 0.12	4.82 ± 0.08
Synthesis	3.59 ± 0.28	4.69 ± 0.09

Table 5. MOS, Comparison, and 95% confidence intervals regarding naturalness, intelligibility and perceptibility of emphasis for TTS samples with randomly emphasized word.

	TTS with Emphasis	Change vs. Standard Synthesis
Naturalness	3.29 ± 0.16	- 0.3
Intelligibility	4.71 ± 0.13	+ 0.02
Emphasis	3.75 ± 0.28	-

ity considering two metrics, naturalness and intelligibility on a scale from 1 to 5. For the emphasized TTS outputs perceptibility of emphasis was additionally rated by the participants. Table 4 shows the MOS and confidence intervals for ground truth and unemphasized samples. Table 5 shows the MOS and confidence intervals for the synthesized samples with added random emphasis. Additionally, changes in naturalness and intelligibility scores when compared with non-emphasized TTS samples are shown.

The results show no clear difference between intelligibility scores of synthesized samples and ground truth samples. However, naturalness is rated worse for synthesized samples, implying a perceptible difference in audio quality or prosody when comparing ground truth and synthesized samples. But these differences do not seem to decrease intelligibility in any way. Adding emphases to the generated speech seems to slightly decrease naturalness, suggesting that the emphases, while being well perceptible, might not sound entirely natural.

The overall performance of our model, even when emphases are added, is comparable with the MOS results for Tacotron 2 and our modified FastSpeech 2 obtained in the first user study as shown in Table 3. However, there cannot be any conclusive comparison as these models have been trained for English speech and only a single MOS value was given by the participants.

5.5. Lip Generation Results

In order to evaluate the lip generation performance, we follow three different strategies. We first evaluate the quality of the generated images by using FID score [21]. We further consider the conditional image generation by measuring the synchronization between the generated lip and the audio input. For this, we benefit from recently proposed novel metrics, LSE-D and LSE-C [50], which are basically distance and confidence scores for the synchronization per-

Table 6. Evaluation of Wav2Lip and our model. We test both model on LRS2 test set and our test set. Since we do not have the ground truth outputs for the German language scenario, we could not calculate FID scores. Wav2Lip-GAN results on LRS2 datasets are taken from the corresponding paper [50].

Model	Data	Language	LSE-D	LSE-C	FID
Wav2Lip-GAN	LRS2	English	6.46	7.78	4.44
Ours	LRS2	English	6.98	6.93	8.86
Wav2Lip-GAN	Ours	English	8.35	6.40	19.62
Ours	Ours	English	8.11	6.52	21.15
Wav2Lip-GAN	Ours	German	7.93	7.18	-
Ours	Ours	German	7.90	7.17	-

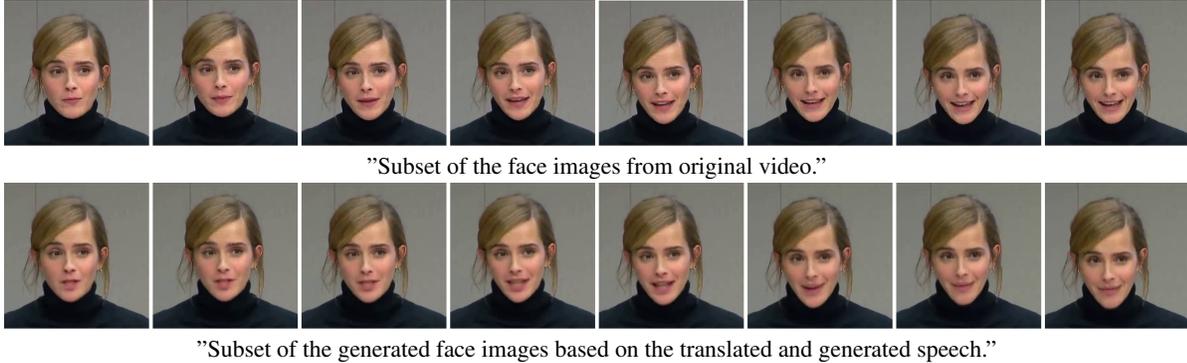


Figure 3. Sample face images from original video and generated video. In the first row, eight consecutive frames from original video are presented. In the second row, the same eight frames with new lips are presented. The images in the second row are synthesized by using generated German speech which is the translation of the input speech.

formance. Finally, we perform subjective tests to quantify the proposed system’s performance. In Table 6, we show LSE-D, LSE-C, and FID scores for LRS2 test set as well as our proposed test set. The LSE-D and LSE-C results on LRS2 dataset show that our model and Wav2Lip [50] achieve almost the same performance to provide synchronized lips, although Wav2Lip shows slightly better scores. On the other hand, on our dataset, we achieve a slightly better scores in English case and in German case, though the scores are quite similar. This outcome indicates that both models have an effective generalization capacity and they are robust against real-world challenges, since both models show a well performance on unseen dataset. Moreover, FID scores are again very close to each other. However, Wav2Lip achieves better FID scores on LRS2 and our dataset. This FID analysis indicates that the generation quality of our model should be improved. Please note that we could not calculate the FID score for our dataset, since we generate the faces based on the German audio input, therefore, there are no ground truth images.

5.6. System Evaluation

In order to evaluate the whole system, we conduct a user study with 25 participants. In this way, we aim to inves-

tigate 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 ask one question for each aspect. However, please note that only the participants who know the German language answered the questions related to intelligibility and the German translation of the speech. In the user study, we randomly choose 80 videos from the 262 videos of the dataset described in Subsection 5.1 and show the translated and lip-synced results to the participants with the transcribed original speech as well as the five questions. Sample evaluation videos can be found here ¹.

We illustrate results of quality of the generated faces, synchronization accuracy, and the translation accuracy in Figure 4. The results indicate that participants rate the quality of the generated faces as high. Similarly, the majority of the answers state that our system is provides accurate synchronization in the generated videos. For the translation accuracy, although a majority of the answers indicate that there are minor mistakes in the text, only 15% of the

¹<https://videospoeechtranslation.github.io>

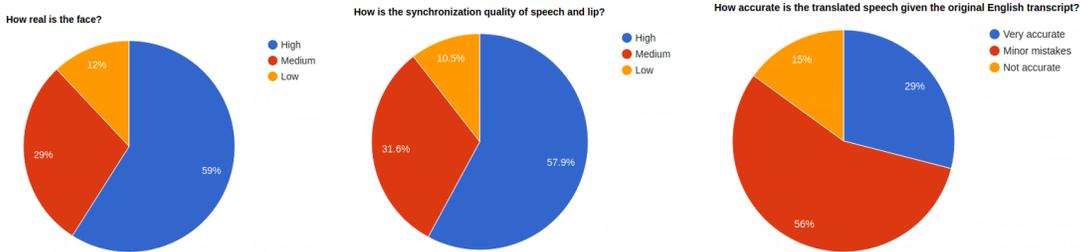


Figure 4. Subjective test results on our proposed test set. We ask participants to evaluate the generated videos in several different aspects, namely the quality of the generated face images, the synchronization quality of the lips and speech, the accuracy of the translated speech, naturalness of the generated speech, and the intelligibility of the generated speech.

Table 7. Subjective evaluation on our proposed test set. In the questions, the minimum score is 1 while the maximum score is 5. Scores show the mean value and standard deviation.

Measurement	Score
Naturalness	3.36 ± 0.98
Intelligibility	4.24 ± 0.86

answers find the results inaccurate. Moreover, we demonstrate naturalness and intelligibility results in Table 7. The results demonstrate 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 well intelligible. Sample images are shown in Figure 3. 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-synching model showed slight issues with bearded faces and also had some quality problems that must be addressed to improve the quality of the generated faces to make them more natural.

6. Conclusion

In this work, we proposed an end-to-end system for combined face, lip, audio translation from input video. Given a video of a speaker, our 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. The detailed experimental results of each module and user study for the system evaluation indicated that we achieve accurate modules for each task and acceptable performance

in the final system to do the video translation. To address remaining translation issues discovered in our experiments and to improve naturalness of the generated speech preserving pauses, we employ more advanced ASR models with punctuation capabilities and voice activity detection to mark pauses in the transcript. This information improves translation quality and naturalness. Improvements towards more robust voice conversion are also desirable as we still observe occasional robustness issues on long speech inputs. These issues are likely to improve with additional training data that specifically incorporates long sentence speech samples. Lastly, as the pipeline of our system contains many components, inference times and latency of the ensemble need to be improved. Ongoing work is devoted to improving speed and latency of the components, parallelizing component processing and a better pipelined architecture.

References

- [1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. In *arXiv:1809.02108*, 2018. 6
- [2] Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marceley Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, et al. Findings of the iwslt 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, 2022. 1
- [3] Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online, July 2020. Association for Computational Linguistics. 3
- [4] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lind

- say Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019. 7
- [5] Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155, 2021. 7
- [6] Mauro Cettolo, Christian Girardi, and Marcello Federico. Wit3: Web inventory of transcribed and translated talks. In *Conference of european association for machine translation*, pages 261–268, 2012. 4
- [7] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018. 4
- [8] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019. 4
- [9] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742*, 2019. 3
- [10] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3444–3453. IEEE, 2017. 6
- [11] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 2, 6
- [12] J. S. Chung and A. Zisserman. Lip reading in profile. In *British Machine Vision Association and Society for Pattern Recognition*, 2017. 6
- [13] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, June 2019. 7
- [14] Matthias Eck, Ian Lane, Ying Zhang, and Alex Waibel. Jibbig: Speech-to-speech translation on mobile devices. In *2010 IEEE Spoken Language Technology Workshop*, pages 165–166. IEEE, 2010. 1
- [15] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Transactions on Graphics*, 38(4):68:1–68:14, July 2019. 4
- [16] Christian Fügen, Alex Waibel, and Muntsin Kolss. Simultaneous translation of lectures and speeches. *Machine translation*, 21(4):209–252, 2007. 1
- [17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 6
- [18] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020. 4
- [19] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 4
- [20] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *International Conference on Speech and Computer*, pages 198–208. Springer, 2018. 7
- [21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 8, 9
- [22] Keith Ito and Linda Johnson. The lj speech dataset. <https://keithito.com/LJ-Speech-Dataset/>, 2017. 8
- [23] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You Said That?: Synthesising Talking Faces from Audio. *International Journal of Computer Vision*, 127(11):1767–1779, Dec. 2019. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 11 Publisher: Springer US. 4
- [24] Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. Towards Automatic Face-to-Face Translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 1428–1436, New York, NY, USA, Oct. 2019. Association for Computing Machinery. 4
- [25] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86, 2005. 4, 7
- [26] Muntsin Kolss, Matthias Wölfel, Florian Kraft, Jan Niehues, Matthias Paulik, and Alex Waibel. Simultaneous german-english lecture translation. In *Proceedings of the 5th International Workshop on Spoken Language Translation: Papers*, 2008. 1
- [27] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33, 2020. 5
- [28] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brebisson, and Yoshua Bengio. ObamaNet: Photo-realistic lip-sync from text. *arXiv:1801.01442 [cs]*, Dec. 2017. arXiv: 1801.01442. 4
- [29] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2755–2764, 2021. 4
- [30] Surafel M Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. Isometricmt: Neural machine translation for automatic dubbing. *arXiv preprint arXiv:2112.08682*, 2021. 1

- [31] Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, and Puming Zhan. Janus-iii: Speech-to-speech translation in multiple languages. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 99–102. IEEE, 1997. 1
- [32] Yist Y. Lin, Chung-Ming Chien, Jheng-Hao Lin, Hung yi Lee, and Lin shan Lee. Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention, 2021. 3
- [33] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-Aware Implicit Neural Audio-Driven Video Portrait Generation. *arXiv:2201.07786 [cs, eess]*, Jan. 2022. *arXiv: 2201.07786*. 4
- [34] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502, 2017. 5
- [35] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 6
- [36] Markus Müller, Thai-Son Nguyen, Jan Niehues, Eunah Cho, Bastian Krüger, Thanh-Le Ha, Kevin Kilgour, Matthias Sperber, Mohammed Mediani, Sebastian Stüker, et al. Lecture translator-speech translation framework for simultaneous lecture translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 82–86, 2016. 1, 2
- [37] Thai-Son Nguyen, Sebastian Stüker, and Alex Waibel. Super-human performance in online low-latency recognition of conversational speech. *CoRR*, abs/2010.03449, 2020. 3
- [38] Tuan Nam Nguyen, Thai Son Nguyen, Christian Huber, Ngoc-Quan Pham, Thanh-Le Ha, Felix Schneider, and Sebastian Stüker. KIT’s IWSLT 2021 offline speech translation system. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 125–130, Bangkok, Thailand (online), Aug. 2021. Association for Computational Linguistics. 1
- [39] Thai-Son Nguyen, Sebastian Stueker, Jan Niehues, and Alex Waibel. Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7689–7693. IEEE, 2020. 4
- [40] Thai-Son Nguyen, Sebastian Stüker, and Alex Waibel. Super-human performance in online low-latency recognition of conversational speech. *arXiv preprint arXiv:2010.03449*, 2020. 1, 3, 4
- [41] Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. Low-latency neural speech translation. *arXiv preprint arXiv:1808.00491*, 2018. 1
- [42] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015. 7
- [43] 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*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 8
- [44] Kyubong Park and Thomas Mulc. Cssl0: A collection of single speaker speech datasets for 10 languages. *Interspeech*, 2019. 7, 8
- [45] Ngoc-Quan Pham, Thanh-Le Ha, Tuan-Nam Nguyen, Thai-Son Nguyen, Elizabeth Salesky, Sebastian Stüker, Jan Niehues, and Alex Waibel. Relative Positional Encoding for Speech Recognition and Direct Translation. In *Proc. Interspeech 2020*, pages 31–35, 2020. 4
- [46] Ngoc-Quan Pham, Tuan Nam Nguyen, Thanh-Le Ha, Sebastian Stüker, Alexander Waibel, and Dan He. Multilingual speech translation KIT @ IWSLT2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 154–159, Bangkok, Thailand (online), Aug. 2021. Association for Computational Linguistics. 1
- [47] Ngoc-Quan Pham, Tuan-Nam Nguyen, Sebastian Stüker, and Alex Waibel. Efficient Weight Factorization for Multilingual Speech Recognition. In *Proc. Interspeech 2021*, pages 2421–2425, 2021. 4
- [48] Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, Sebastian Stüker, and Alexander Waibel. Very deep self-attention networks for end-to-end speech recognition. *arXiv preprint arXiv:1904.13377*, 2019. 4
- [49] Ngoc-Quan Pham, Alex Waibel, and Jan Niehues. Adaptive multilingual speech recognition with pretrained models. *arXiv preprint arXiv:2205.12304*, 2022. 3
- [50] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV 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*, pages 484–492, 2020. 2, 4, 6, 8, 9, 10
- [51] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss, 2019. 3
- [52] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*, 2020. 3, 5, 9
- [53] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32, 2019. 5
- [54] Max Ritter, Uwe Meier, Jie Yang, and Alex Waibel. Face translation: A multimodal translation agent. In *AVSP’99-International Conference on Auditory-Visual Speech Processing*, 1999. 2, 3, 4
- [55] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*, 2018. 7
- [56] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis

- by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018. 3, 5
- [57] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: learning lip sync from audio. *ACM Transactions on Graphics*, 36(4):95:1–95:13, July 2017. 4
- [58] Paul Taylor and Amy Isard. Ssml: A speech synthesis markup language. *Speech communication*, 21(1-2):123–133, 1997. 5
- [59] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural Voice Puppetry: Audio-Driven Facial Reenactment. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 716–731, Cham, 2020. Springer International Publishing. 4
- [60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 4, 5
- [61] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic Speech-Driven Facial Animation with GANs. *International Journal of Computer Vision*, 128(5):1398–1413, May 2020. 4
- [62] Alexander Waibel and Christian Fuegen. Simultaneous translation of open domain lectures and speeches, Jan. 3 2012. US Patent 8,090,570. 1
- [63] Alex Waibel, Ajay N Jain, Arthur E McNair, Hiroaki Saito, Alexander G Hauptmann, and Joe Tebelskis. Janus: a speech-to-speech translation system using connectionist and symbolic processing strategies. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, pages 793–796. IEEE Computer Society, 1991. 1
- [64] Naomi Aoki Waibel, Alexander Waibel, Christian Fuegen, and Kay Rottman. Hybrid, offline/online speech translation system, Aug. 30 2016. US Patent 9,430,465. 1
- [65] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng. Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion, 2021. 2, 3, 6
- [66] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. *arXiv:2107.09293 [cs]*, July 2021. arXiv: 2107.09293. 4
- [67] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *Proc. Interspeech 2017*, pages 4006–4010, 2017. 3
- [68] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018. 4
- [69] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. DFA-NeRF: Personalized Talking Head Generation via Disentangled Face Attributes Neural Rendering. *arXiv:2201.00791 [cs]*, Jan. 2022. arXiv: 2201.00791. 4
- [70] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3867–3876, October 2021. 4
- [71] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 4
- [72] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking Face Generation by Adversarially Disentangled Audio-Visual Representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9299–9306, July 2019. Number: 01. 4
- [73] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4176–4186, June 2021. 4
- [74] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. MakeltTalk: speaker-aware talking-head animation. *ACM Transactions on Graphics*, 39(6):221:1–221:15, Nov. 2020. 4