AV2AV: Direct Audio-Visual Speech to Audio-Visual Speech Translation with Unified Audio-Visual Speech Representation

Supplementary Material

	X-En			En-X		
Method	**AQ**	**VQ**	**R**	**AQ**	**VQ**	**R**
AVSR + NMT + TTS + TFG	3.79	4.50	3.30	3.60	3.89	3.20
Proposed Method (AV2AV)	3.97	4.60	4.00	4.51	4.11	4.07

Table 7. MOS scores on the translated AV output in terms of AQ (Audio Quality), VQ (Visual Quality), and R (Realness).

6. Human Subject Study on Audio-Visual Quality

In order to evaluate the synthesized quality of audio and video, we conducted a Mean Opinion Score (MOS) test. We gathered 20 participants and asked them to evaluate 32 generated samples of each method to rate in terms of Audio Quality (AQ), Visual Quality (VQ), and overall Realness (R). We presented the audio stream only to evaluate the AQ, the visual stream only for the VQ, and the audio-visual stream altogether to evaluate the R. We compare against the best-performing cascaded system, which is the 4-stage cascaded system of AVSR, NMT, TTS, and TFG. The MOS results are shown in Table 7. The result demonstrates that we can attain comparable performances with the proposed direct AV2AV approach as with the 4-stage cascaded method comprising state-of-the-art subsystems. Specifically, when both audio and visual streams are simultaneously presented, the participants assess the proposed method more seamlessly generates the two modalities than the cascaded method, as shown in the table (*i.e.*, Realness (R)).

7. Dataset Statistics

For training the m-AVHuBERT, we use a total of 7,011 hours of the following combined AV datasets.

LRS2 [14] is an English audio-visual dataset, containing 233 hours of training data from British TV shows.

LRS3 [63] is an English audio-visual dataset consisting of approximately 430 hours of video clips from TED and TEDx.

VoxCeleb2 [88] is a large-scale multilingual corpus for speaker recognition. It has over 1 million utterances from 6,112 celebrities of 145 different nationalities.

AVSpeech [90] is a large-scale multilingual corpus comprising 4700 hours of video clips with no interfering background noises sourced from a total of 290k YouTube videos.

mTEDx [89] is a multilingual corpus built for speech recognition and speech translation sourced from TEDx talks. It consists of 8 languages; Spanish (Es), French (Fr), Italian (It), Portuguese (Pt), Russian (Ru), Greek (El), Arabic (Ar), and German (De). We use cleaned data of Es, Fr, It, and Pt following [107].

En-Es	**En-Fr**	**En-It**	**En-Pt**
437	437	437	437
Es-En	**Fr-En**	**It-En**	**Pt-En**
178	176	101	153

Table 8. Finetuning dataset amount (hours) of MuAViC for each language pair.

En-Es	**En-Fr**
200	200

Table 9. Finetuning dataset amount (hours) of LRS3-T for each language pair.

For pretraining the AV2AV language translation model, we use a total of 12k hours of the following A2A datasets. Please refer to [11] for detailed data statistics for each translation pair.

Voxpopuli [35] is a multilingual A2A corpus from European Parliament even recordings. Following [11], we use translation from 15 source languages to 15 target languages, which results in 11.2k hours of translation data.

mTEDx [89] contains speech-to-text translation data from English (En) to Es, Fr, Pt, It, Ru, and El. Since it does not have target speech, we use generated speech from a pretrained TTS model which gives a total duration of 0.7k hours as in [11].

For evaluation, we finetune the AV2AV model on the following evaluation datasets. The detailed information for each translation pair is shown in Table 8 and Table 9

MuAViC [33] is a multilingual corpus for audio-visual speech recognition and audio-visual speech-to-text translation (AV2T). It reuses videos of LRS3 and mTEDx datasets including 1,200 hours of transcribed text from over 8000 speakers in 9 languages. Their transcriptions are generated by using an NMT model. Since it only provides target text, we generate the target speech by using pretrained TTS models, VITS [98] on each language. We utilize 4 En-to-X and 4 X-to-En paired data, where X is Es, Fr, Pt, and It, which gives a total of 2,356 hours.

LRS3-T [62] is an AV2A corpus curated from the LRS3 [63] dataset by converting the transcribed English text into the speech in target languages. It results in 200 hours of parallel AV2A translation pairs for En-to-Es and En-to-Fr.

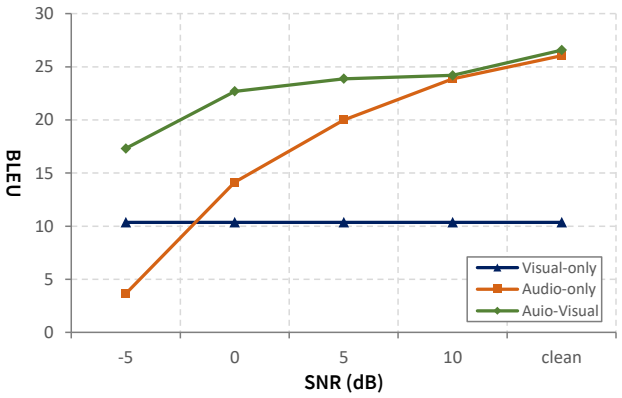

Figure 5. Evaluation of the robustness of the proposed AV2AV system using different input modalities on acoustic noise.

8. Additional Results

8.1. Generated Results

We show more generated translation results of the proposed AV2AV system in Fig. 6. The AV2AV system seamlessly translates the input video into the target language by transforming the mouth region while keeping the head pose and identity unchanged. The mouth movement aligns well with the corresponding phonemes as written under each video frame. The transcription obtained from ASR of the generated speech is also semantically coherent with the source speech. Moreover, since our model has been trained in many-to-many settings, our model supports translation into multiple different target languages. As shown in the second and third rows of (a-d), it can faithfully generate translated audio and visual speech in different target languages from a single source input.

Refer to the demo video for more demonstrations of the generated translation results. The demo consists of four parts: the first part shows En-X results on LRS3-T data, the second part shows X-EN results on mTEDx data, the third part compares the proposed method with the best-performing cascaded system, and the last part presents results on a completely different domain, the HDTF [47] data, The HDTF is a high-resolution (720P or 1080P) audio-visual dataset for TFG. For HDTF testing, please note that we have trained the face renderer on the HDTF dataset and attached a face enhancer [108] to support high-resolution synthesis of HDTF data.

8.2. Noise Robustness

We visualize the robustness of the proposed AV2AV system to acoustic noise. Specifically, we perturb the clean audio input with the babble noise of MUSAN [104] by varying Signal-to-Noise Ratio (SNR) levels from -5 dB to 10 dB. The BLEU score of the translated results on MuAViC Es-En data is shown in Fig. 5. As the visual stream is not affected by the acoustic noise, the visual-only case shows the constant BLEU score for all noise levels. In contrast, the audio-only model is mostly affected by the acoustic noise, and the performance is greatly decreased when the noise becomes strong. The interesting thing is that the audio-only model even shows worse performance than the visual-only model in the severely noisy case (*i.e.*, -5 dB). Therefore, it would be better to utilize a visual-only model in a very noisy situation. When we utilize multimodal inputs, audio-visual, the model shows robust performances to the acoustic noise. The model always shows the best performances for all noise levels. Even if the noise is strong, so that the SNR is -5 or 0 dB, the audio-visual model still can robustly translate the speech. The results show the importance of using multimodal inputs for practical usage of speech processing systems.

Hyper-parameter	Value
# steps	150k
# warmup steps	12k
LR scheduler	polynomial decay
peak learning rate	5e-4
max frames / GPU	1000
# GPUs	63
Adam (β_1, β_2)	(0.9, 0.98)

Table 10. Training hyper-parameters of the mAV-HuBERT.

Hyper-parameter	Value
# steps	30k
# warmup steps	10k
LR scheduler	polynomial decay
peak learning rate	3e-4
max tokens / GPU	1024
# GPUs	56
Adam (β_1, β_2)	(0.9, 0.98)

Table 11. Pretraining hyper-parameters of the AV2AV model.

Hyper-parameter	Value
# steps	30k
# warmup steps	3k
LR scheduler	polynomial decay
peak learning rate	1e-4
max tokens / GPU	1024
# GPUs	64
Adam (β_1, β_2)	(0.9, 0.98)

Table 12. Fine-tuning hyper-parameters of the AV2AV model.

9. Implementation Details

The mAV-HuBERT has the same architecture as the AV-HuBERT [52] large configuration which consists of 24 transformer encoder layers with 16 attention heads, a feed-forward dimension of 4,096, and an embedding dimension

of 1,024. We initialize the model with an AV-HuBERT pretrained on 1,759 hours of English subset of LRS3 and VoxCeleb2. Given the masked audio and visual streams as input, it aims to predict the corresponding target clusters extracted from a pretrained multilingual HuBERT [67]. During training, we apply modality dropout with a probability of 0.5. To extract the AV speech units, we cluster the unified AV representation into 1,000 clusters using k-mean clustering. Please refer to Table 10 for training configurations.

The AV2AV model is composed of an encoder embedding layer, 12 transformer encoder layers, 12 transformer decoder layers, and decoder embedding layers. The unit vocabulary size is 1,000 and the embedding dimension of each unit is 1,024. Both the unit encoder and the unit decoder have 8 attention heads and a feed-forward dimension of 4,096. The unit encoder is conditioned on the source language token and the unit decoder generates the target AV speech units conditioned on the target language token. We initialize the model from the mHuBERT unit-based A2A model [11] and pretrain it based on our AV speech units on parallel A2A translation data for 30k steps with a max token length of 1,024 using the training configuration settings as shown in Table 11. Then, the pretrained model is further fine-tuned on the evaluation datasets for 30k steps following the configuration settings in Table 12. The number of GPUs is adjusted according to the dataset size when fine-tuning on the LRS-T dataset and the model with the best performance in the validation set was used. During training, we utilized AV speech units extracted from audio-only, while at inference, we can use AV speech units extracted from any of audio-only, visual-only, and audio-visual data to perform A2AV, V2AV, and AV2AV.

The vocoder is based on the unit-based HiFi-GAN vocoder [92] with an additional speaker encoder [95] to extract the speaker embedding, a d-vector [97]. The AV speech units are embedded into 128-dimension through an embedding layer, and the d-vector is also embedded into 128-dimension with a linear projection. The two embeddings from the AVspeech units and the d-vector are channel-wise concatenated at each timestep. We train the vocoder with a length predictor for 1M steps on individual languages with 1 GPU and a batch size of 16. The face renderer is based on a TFG model, Wav2Lip [27]. The AV speech units are embedded by an embedding layer of dimension 512 and a single transformer Encoder layer, which are channel-wise concatenated with the corresponding identity features. We train the face renderer for 35k steps with 1 GPU and a batch size of 64. Learning rate of 1e-4 and Adam optimizer are used for both the vocoder and the face renderer.

10. Advantages of Direct AV2AV Approach

By employing the proposed direct AV2AV approach, we can gain several advantages compared to utilizing cascaded systems. **1) Inference speed.** As the proposed system does not go through with the intermediate text representations, the inference speed is faster than the cascaded system. We compared the inference speed with the best-performed cascaded system (*i.e.*, A2 in Table 2). On average, the proposed AV2AV approach required 1.36s, whereas the cascaded system took 2.42s to process 2.03s video (*i.e.*, audio-visual) for the complete pipeline. **2) Model size.** Compared to the cascaded systems, the proposed direct AV2AV approach has a smaller model size. The number of parameters of the cascaded system is 1,490M parameters in total. In contrast, the proposed AV2AV has 732M parameters. **3) Text-free ability.** As the proposed AV2AV can be trained without using text data, the system can be served even for languages having no writing systems, while the cascaded systems cannot.

Figure 6. (a-d) Translated results of the proposed AV2AV, each of which the first row is the source input and the second and third rows are the translated outputs in different target languages.