# AV-TranSpeech: Audio-Visual Robust Speech-to-Speech Translation

**Rongjie Huang**[1]*, **Huadai Liu**[1]*, **Xize Cheng**[1]*, **Yi Ren**[2], **Linjun Li**[1], **Zhenhui Ye**[1], **Jinzheng He**[1], **Lichao Zhang**[1], **Jinglin Liu**[2], **Xiang Yin**[2], **Zhou Zhao**[1]†

Zhejiang University[1], ByteDance[2]

{rongjiehuang,liuhuadai,zhaozhou}@zju.edu.cn

{ren.yi,liu.jinglin,yinxiang.stephen}@bytedance.com

## Abstract

Direct speech-to-speech translation (S2ST) aims to convert speech from one language into another, and has demonstrated significant progress to date. Despite the recent success, current S2ST models still suffer from distinct degradation in noisy environments and fail to translate visual speech (i.e., the movement of lips and teeth). In this work, we present AV-TranSpeech, the first audio-visual speech-to-speech (AV-S2ST) translation model without relying on intermediate text. AV-TranSpeech complements the audio stream with visual information to promote system robustness and opens up a host of practical applications: dictation or dubbing archival films. To mitigate the data scarcity with limited parallel AV-S2ST data, we 1) explore self-supervised pre-training with unlabeled audio-visual data to learn contextual representation, and 2) introduce cross-modal distillation with S2ST models trained on the audio-only corpus to further reduce the requirements of visual data. Experimental results on two language pairs demonstrate that AV-TranSpeech outperforms audio-only models under all settings regardless of the type of noise. With low-resource audio-visual data (10h, 30h), cross-modal distillation yields an improvement of 7.6 BLEU on average compared with baselines.[1]

## 1 Introduction

Speech-to-speech translation (S2ST) models (Tjandra et al., 2019; Zhang et al., 2020; Jia et al., 2021) relying on speech data have achieved high performance and significantly broken down communication barriers between people not sharing a common language, which attracts broad interest in the machine learning community (Huang et al., 2022c; Huang et al.). Among them, direct systems (Lee et al., 2021a,b; Huang et al., 2022d) leverage recent progress on self-supervised discrete units learned from unlabeled speech for building textless S2ST, further supporting translation between unwritten languages (Chen et al., 2022).

As speech production (Huang et al., 2023; Lam et al., 2021; Huang et al., 2022b) is accompanied by the movement of lips and teeth, it can be visually interpreted to understand speech. In recent years, significant research (Shi et al., 2022a; Prajwal et al., 2022) has introduced joint modeling of spoken language and vision: Shi et al. (2022b) investigate to learn lip-based audio-visual speaker embeddings, where the speaker's mouth area is used alongside speech as inputs. Chern et al. (2022) focus on audio-visual speech enhancement and separation which better integrates visual information. Despite their success, it is unclear how lip can contribute to audio-based S2ST, and how to incorporate visual modality as auxiliary information in S2ST. A visual translator may open up a host of practical applications: improving speech translation in noisy environments, enabling dictation, or dubbing archival films.

Despite the benefits of audio-visual approaches, training direct speech translation models without relying on intermediate text typically requires a large amount of training data, while there are very few resources providing parallel audio-visual speech due to the heavy workload. To mitigate the data scarcity, researchers have leveraged multitask learning (Lee et al., 2021a), data augmentation (Popuri et al., 2022), and weekly-supervised data with synthesized speech (Jia et al., 2022a) in audio S2ST.

In this work, we propose AV-TranSpeech, introducing the first AV-S2ST system without using text. As illustrated in Figure 1, our textless AV-TranSpeech inherits speech-to-unit translation (S2UT) framework (Lee et al., 2021b; Huang et al., 2022d), which consists of an audio-visual speech-to-unit translation (AV-S2UT) model followed by

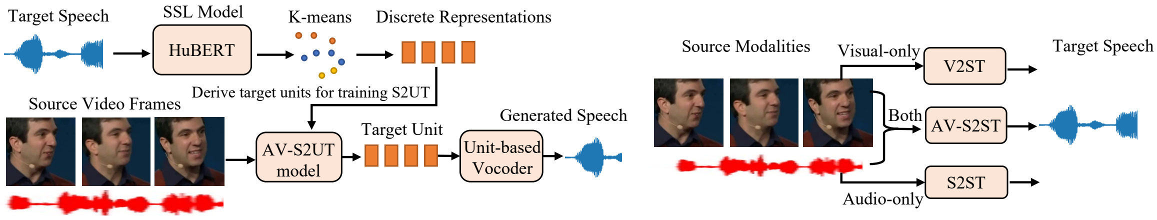
*Equal contributions

†Corresponding author

[1]Audio samples are available at https://AV-TranSpeech.github.io/.



*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 8590–8604
July 9-14, 2023 ©2023 Association for Computational Linguistics




(a) Direct audio-visual speech-to-speech translation (AV-S2ST) system (b) Speech-to-speech translation with varying input modalities

Figure 1: Beyond speech-to-speech translation (S2ST), we introduce visual-to-speech translation (V2ST) and audio-visual speech-to-speech translation (AV-S2ST), unlocking the ability for high-quality translation given a user-defined modality input.

a unit-based vocoder that converts discrete units to speech. AV-TranSpeech complements the audio stream with the auxiliary visual information, which is invariant to speaking environments and promotes system robustness. To tackle the challenges of data shortage, we 1) build upon the recently introduced Audio-Visual HuBERT (AV-HuBERT) which learns contextual representations through self-supervised masked prediction, and show that large-scale pre-training benefits AV-S2ST training; 2) introduce cross-modal distillation and leverage S2ST models trained on the audio-only corpus, which further reduces the requirements of visual data and boosts the performance of visual systems in low-resource scenarios.

Experimental results on two language pairs demonstrate the robustness of AV-TranSpeech in noisy scenarios, outperforming audio-only S2ST under all settings regardless of the SNR and the type of noise. With low-resource audio-visual data (10h, 30h), cross-modal distillation yields an improvement of 7.6 BLEU on average compared with baselines. The main contributions of this work include:

- We propose the first textless audio-visual speech-to-speech (AV-S2ST) translation model AV-TranSpeech and collect a benchmark dataset LRS3-T which we plan to release.
- We leverage the recent success of audio-visual self-supervised learning for contextual representations and show that large-scale pre-training alleviates the data scarcity issue in AV-S2ST systems.
- We introduce cross-modal knowledge distillation, which further reduces the requirements of visual data and boosts AV-S2ST performances in low-resource scenarios.
- Experimental results on two language pairs demonstrate the robustness of AV-TranSpeech in the noisy environment under all settings regardless of the type of noise, and cross-modal distillation yields a significant improvement compared with baselines.

## 2 Related Work

### 2.1 Speech-to-Speech Translation

Direct speech-to-speech translation has made massive progress to date. Translatotron (Jia et al., 2019) is the first direct S2ST model and shows reasonable translation accuracy and speech naturalness. Translatotron 2 (Jia et al., 2021) utilizes the auxiliary target phoneme decoder to promote translation quality but still needs phoneme data during training. UWSpeech (Zhang et al., 2020) builds the VQ-VAE model and discards transcript in the target language, while paired speech and phoneme corpora of written language are required. Most recently, a textless S2ST system (Lee et al., 2021a) takes advantage of self-supervised learning (SSL) and leverages speech-to-unit translation (S2UT) model followed by a unit-based vocoder that converts discrete units to speech, demonstrating the results without using text data. Popuri et al. (2022) show that self-supervised encoder and decoder pre-training with weakly-supervised data improves model performance. Huang et al. (2022d) apply speech normalization on rhythm, pitch, and energy to create deterministic training targets.

Despite their recent success, current S2ST models still suffer from distinct degradation in noisy scenarios and fail to translate visual speech. In this work, we complement the audio stream with the visual information, opening up a host of practical applications (improving speech translation in noisy environments, enabling silent dictation, or dubbing silent archival films), which has been relatively overlooked.

### 2.2 Audio-Visual Speech Self-Supervised Learning

It has been an increasing interest in self-supervised learning in the machine learning and speech-processing community. Wav2vec 2.0 (Baevski et al., 2020) trains a convolutional neural network to distinguish true future samples from random distractor samples using a contrastive predictive coding (CPC) loss function. HuBERT (Hsu et al., 2021) is trained with a masked prediction with masked continuous audio signals. For audio-visual representation learning, they rely on spatio-temporal CNNs consisting of multiple 3D convolutional layers or a single 3D convolutional layer followed by 2D ones. Stafylakis and Tzimiropoulos (2017) propose a residual network with 3D convolutions to extract more powerful representations.

Very recently, Shi et al. (2022a) propose AV-HuBERT with a self-supervised representation learning framework by masking multi-stream video input and predicting automatically discovered multimodal hidden units. It has been demonstrated to learn discriminative audio-visual speech representation, and thus we leverage the contextual representations to enhance the AV-S2ST performance.

### 2.3 Transfer Learning

Transferring knowledge across domains is a promising machine learning methodology for solving the data shortage problem. Zhang et al. (2021) perform transfer learning from a text-to-speech system to voice conversion with non-parallel training data. Afouras et al. (2020) apply cross-modal distillation from ASR for learning audio-visual speech recognition, where they train strong models for visual speech recognition without requiring human annotated ground-truth data. Cai et al. (2020) enhance the knowledge transfer from the speaker verification to the speech synthesis by engaging the speaker verification network. Popuri et al. (2022) perform transfer learning from a natural language expert mBART for faster coverage of training speech translation models.

Our approach leverages networks trained on one modality to transfer knowledge to another. In this way, the dependence on a large number of parallel audio-visual data can be reduced for constructing AV-S2ST systems.

## 3 AV-TranSpeech

In this section, we first overview the encoder-decoder architecture for AV-TranSpeech, following which we introduce the cross-modal distillation procedure for few-shot transfer learning with low-resource data. The overall architecture has been illustrated in Figure 1, and we put more details on the encoder and decoder block in Appendix A.

### 3.1 Overview

The overall AV-S2ST pipeline has been illustrated in Figure 1, where we 1) use the SSL HuBERT (Hsu et al., 2021) to derive discrete units of target speech; 2) build the audio-visual speech-to-unit translation (AV-S2UT) and 3) apply a separately trained unit-based vocoder to convert the translated units into waveform.

For audio-visual speech-to-unit translation, we adopt the encoder-decoder sequence-to-sequence model as the backbone. The audio-visual speech samples first pass through the multi-layer audio and video feature extractors, which are then fused and fed into the backbone conformer encoder. In the following, the unit decoder autoregressively predicts unit sequences corresponding to the target speech.

Training direct textless AV-S2ST models typically requires a large amount of parallel training data (Duquenne et al., 2022; Lee et al., 2021b), while resources providing parallel multimodal data could be limited due to the heavy workload. To alleviate the issue of data scarcity, we 1) build upon the recently introduced Audio-Visual HuBERT (AV-HuBERT) which learns contextual representations through self-supervised masked prediction, and show that large-scale pre-training benefits AV-S2ST training; 2) introduce the cross-modal distillation with S2ST models trained on the audio-only corpus, which further reduces the requirements of visual data and boosts the performance of visual systems in low-resource scenarios (10h, 30h).

### 3.2 Pre-Trained Encoder

Audio-Visual Hidden Unit BERT (AV-HuBERT) is a self-supervised model that learns from unlabeled audio-visual speech data. AV-HuBERT comprises four modules: a feed-forward network (FFN) audio feature extractor, a modified ResNet (Stafylakis and Tzimiropoulos, 2017; Martinez et al., 2020) video feature extractor, a fusion module, and a conformer (Gulati et al., 2020) backend.

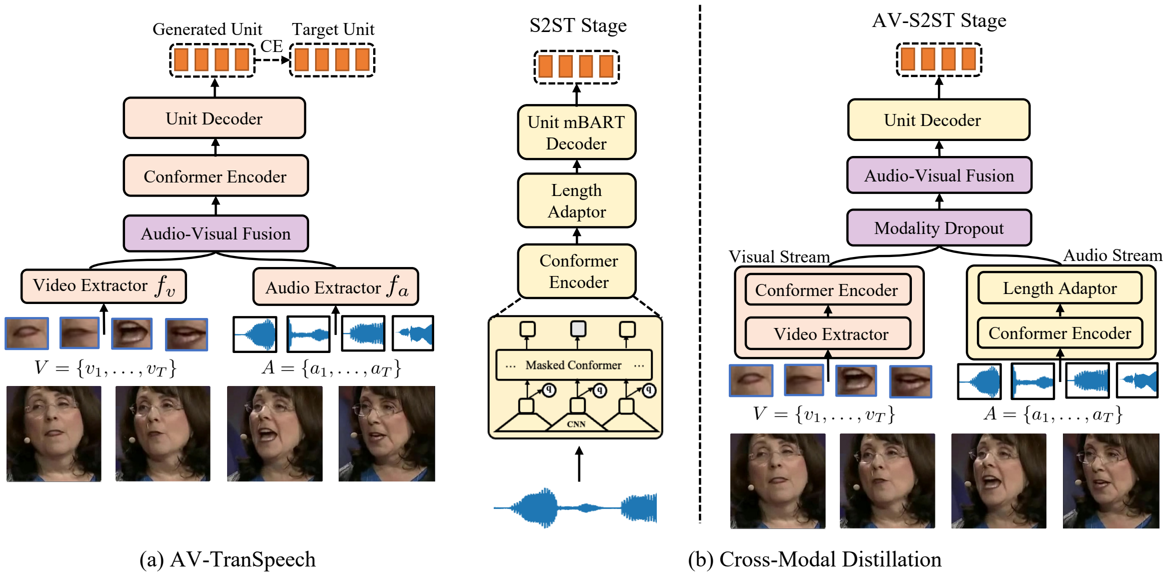



Figure 2: In subfigure (a), we compute the cross-entropy loss (denoted as "CE") during training. In subfigure (b), we initialize the visual and audio streams from AV-Hubert and S2ST models for cross-modal distillation in a low-resource setup. The modality dropout denoted with dotted lines is excluded during inference.

Denote the domain of visual and audio samples by $V, A \subset \mathbb{R}$ respectively. The source language is therefore a sequence of visual $V = \{v_1, \dots, v_T\}$ and speech $A = \{a_1, \dots, a_T\}$ samples for $T$ time-steps. The multi-layer audio feature extractor $f_a$ and the video feature extractor $f_v$ respectively take audio $A$ and visual frames $V$ as input, which are then fused (i.e., element-wise addition) and fed into the backbone conformer encoder and generates contextual representations $X = \{x_1, \dots, x_T\}$.

### 3.3 Unit Decoder

The autoregressive decoder is assisted with an attention module, which takes the encoder output as the source values for the attention, and predicts unit sequences corresponding to the target translated speech. We use a stack of transformer layers as the decoder, along with a multi-head attention (Vaswani et al., 2017). Given the $T$-frame contextual representations from source speech $X = \{x_1, \dots, x_T\}$, autoregressive model $\theta$ factors the distribution over possible outputs $Y = \{y_1, \dots, y_N\}$ by:

$$p(Y \mid X; \theta) = \prod_{i=1}^{N+1} p(y_i \mid y_{0:i-1}, x_{1:T}; \theta), \quad (1)$$

where the special tokens $y_0(\langle bos \rangle)$ and $y_{N+1}(\langle eos \rangle)$ are used to represent the beginning and end of all target units.

### 3.4 Cross-Modal Distillation

In this part, we investigate the transfer learning from orders of magnitude audio data to boost the performance of visual systems. Specifically, we leverage the S2ST model trained on a large-scale audio-only corpus and perform **cross-modal distillation** with low-resource audio-visual data.

**S2ST Model.** We adopt the current state-of-the-art S2ST model (Popuri et al., 2022) with pre-trained wav2vec 2.0 (Baevski et al., 2020) and mBART (Liu et al., 2020). Wav2vec 2.0 is a self-supervised framework to learn speech representations from unlabeled audio data, which is trained via contrastive loss with masked spans on the input to the context encoder. mBART has been originally proposed for denoising autoencoder over text sequences, which predicts the original text $z$ given its noisy version, $g(z)$, created by random masking. As such, the powerful S2ST model provides a significant initialization for training audio-visual systems.

**Modality Adaptor.** The encoder in S2ST model $f_a$ encodes speech representation with $A = \{a_1, \dots, a_{T'}\}$ with a stride of about 20ms, while the video stream $f_v$ generates visual feature sequence $V = \{v_1, \dots, v_T\}$ at a stride of 10 ms from the raw waveform. To alleviate this length mismatch between the audio and visual representations, we add a randomly initialized modality adaptor layer consisting of a single 1-D convolutional layer with stride 2 between the audio and video streams.

**Modality Dropout.** Since S2ST models provide a strong initialization for our AV-S2ST, and thus it can relate audio input to lexical output more effortlessly than the visual input stream, leading to the domination of audio modality in model decisions. To prevent the model's over-reliance on the audio stream in our joint model, we include a modality dropout with $p = 50\%$ probabilities to mask the full features of one modality before fusing audio and visual inputs, forcing the visual encoder to learn contextual representations. We show feature fusion in our cross-modal distillation with modality dropout:

$$f_{av} = \begin{cases} f_a + f_v & \text{with } p = 0.5 \\ f_a + \mathbf{0} & \text{with } p = 0.25 \\ \mathbf{0} + f_v & \text{with } p = 0.25 \end{cases}$$

As such, modality dropout (Chern et al., 2022; Zhang et al., 2019) prevents the model from ignoring video input and encourages the model to produce the prediction regardless of what modalities are used as input.

### 3.5 Model Training

In training AV-TranSpeech, we compute the cross-entropy loss (denoted as "CE") between generated and reference units. For low-resource scenarios, we group our distillation strategies into two categories: 1) for AV-S2ST, we adapt the speech encoder and the unit decoder in S2ST to the visual system; 2) for V2ST with visual-only input, we only transfer the knowledge from unit decoder in S2ST to avoid noisy and incomplete decoding. In this way, the dependence on a large number of parallel audio-visual data can be reduced for constructing visual systems.

## 4 Experiments

### 4.1 Experimental Setup

Following the common practice in the direct unit S2ST pipeline, we apply the publicly-available pre-trained multilingual HuBERT (mHuBERT) model and unit-based HiFi-GAN vocoder (Polyak et al., 2021; Kong et al., 2020), leaving them unchanged.

#### 4.1.1 Dataset

To evaluate the performance of the proposed model, we conduct experiments on two language pairs, including English-Spanish (En-Es), and English-French (En-Fr).

**LRS3-T.** We construct our translation dataset by converting the transcribed English text from LRS3 (Afouras et al., 2018b) into target language using cascaded neural machine translation (NMT) and text-to-speech (TTS) systems. We remove short clips (less than 2 seconds) and discard the non-vocal segments with voice activation detection (VAD). To this end, we collect 200-hour parallel audio-visual translation data (with source videos and target speech), namely LRS3-T which we plan to release.

**CVSS-C.** For training S2ST models, we use the benchmark dataset CVSS-C (Jia et al., 2022b), which is derived from the CoVoST 2 (Wang et al., 2020) speech-to-text translation corpus by synthesizing the translation text into speech using a single-speaker TTS system.

**Noise.** For evaluating our AV-S2ST models under different noise categories, we prepare noise audio clips in the categories of "music" and "babble" sampled from MUSAN dataset (Snyder et al., 2015), and create "speech" noise samples following Popuri et al. (2022).

The total duration of each dataset is shown in Table 1.

| Dataset | Subset | Modality | En-Es | En-Fr |
|---|---|---|---|---|
| LRS3-T | Normal<br>Small<br>Tiny | AV | 200<br>30<br>10 | |
| CVSS-C | / | A | 69.5 | 170 |
| Noise | Music<br>Babble<br>Speech | A | 35<br>20<br>50 | |

Table 1: Total duration in hours of samples in different datasets.

#### 4.1.2 Model Configurations and Training

For training S2ST models, we adopt Wav2vec 2.0 LARGE pre-trained on Libri-light dataset (Kahn et al., 2020) as audio encoder and unit mBART pre-trained on VoxPopuli dataset (Wang et al., 2021) as the decoder. Following the practice in unit-based S2ST (Lee et al., 2021a), we use the k-means algorithm to cluster the representation given by the normalized mHuBERT (Huang et al., 2022d) into a vocabulary of 1000 units as training targets. The inputs to AV-TranSpeech are lip Regions-Of-Interest (ROIs) for the visual stream and 80-dimensional

| ID | Pre-Training | Modality | En-Es | | En-Fr | |
|---|---|---|---|---|---|---|
| | | | BLEU | MOS | BLEU | MOS |
| 1 | / | AV | 0.67 | / | 1.01 | / |
| 2 | AVHubert | AV | 45.2 | 3.82±0.09 | 33.6 | 3.98±0.08 |
| 3 | / | A | 0.51 | / | 0.90 | / |
| 4 | AVHubert | A | 43.1 | 3.80±0.09 | 31.6 | 3.90±0.07 |
| 5 | / | V | 0.18 | / | 0.32 | / |
| 6 | AVHubert | V | 25.0 | 3.94±0.11 | 19.9 | 3.95±0.10 |
| 7 | Enhanced S2ST | A | 42.5 | 3.88±0.10 | 32.0 | 3.91±0.09 |
| 8 | Unit TTS | / | 67.7 | 4.04±0.07 | 54.1 | 4.09±0.10 |
| 9 | NMT+TTS | / | 76.0 | 4.15±0.08 | 63.9 | 4.20±0.10 |

Table 2: Translation quality (BLEU (↑)) and speech naturalness (MOS (↑)) comparison with baseline systems. We set the beam size to 10 in autoregressive decoding.

mel-filterbank features at every 10-ms for the audio stream. As the image frames are sampled at 25Hz, we stack the 4 neighboring acoustic frames to synchronize the two modalities. The encoders in AV-TranSpeech follow AV-HuBERT LARGE configuration with 24 transformer blocks, each with 16 attention heads and 1024/4096 embedding/feed-forward dimensions. We remove the auxiliary tasks for simplification and follow the unwritten language scenario (Lee et al., 2021b). AV-TranSpeech is trained until convergence for 20k steps using the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-8}$) with 6 Tesla V100 GPU. A comprehensive table of hyperparameters is available in Appendix A.

#### 4.1.3 Evaluation

We compare AV-TranSpeech with other systems using the publicly-available *fairseq* framework (Ott et al., 2019), including, 1) NMT+TTS cascaded system to simulate the construction of LRS3-T, where we adopt MMT to convert transcribed English text to target languages (regarded as *reference text*) and then apply TTS model for speech generation, following we transcribe the speech and compute the BLEU; 2) Unit TTS, where we first synthesize speech samples with target units, and then transcribe the speech and compute BLEU; 3) Moreover, we compare the performance of AV-TranSpeech in S2ST with baseline model (Popuri et al., 2022) (denoted as *Enhanced S2ST*).

For translation accuracy, we use open-sourced ASR models in *fairseq* framework to transcribe the audios and then calculate the BLEU score (Papineni et al., 2002) between the generated and the reference text. To evaluate the naturalness of the speech output, we conduct crowd-sourced human evaluations with MOS, rated from 1 to 5 and reported with 95% confidence intervals (CI) via Amazon Mechanical Turk. More details on evaluation have been attached in Appendix B.

### 4.2 Translation Accuracy and Speech Naturalness

Table 2 summarizes the translation accuracy and speech naturalness among all systems, and we have the following observations: 1) **Large-scale multimodal pre-training (1 vs. 2)** improves performance by a large margin, while the naive model fails to work without the self-supervised pre-training strategy. It is mainly because LRS3-T is a challenging unconstrained dataset with a large proportion of videos collected from TED talks, showing the difficulty (Zhang et al., 2020; Jia et al., 2019) of direct speech-to-speech translation without relying on intermediate texts or auxiliary multitask training. In contrast, with a pre-trained AVHubert encoder and a randomly initialized decoder, AV-TranSpeech is efficient in learning contextual representations from audio-visual signals. 2) **Visual modality (2 vs. 4)** has brought a gain of 2.0 BLEU points on average. It complements the audio stream with visual information, opening up a host of practical applications: enabling silent dictation or dubbing archival silent films. 3) We further compare AV-TranSpeech in S2ST with the baseline model (4 vs. 7), showing that AV-TranSpeech with audio-only input is on-par with the current state-of-the-art speech model in terms of translation accuracy. 4) For speech quality, AV-TranSpeech produces natural speech regardless of modalities input competitive with the baseline S2ST system. Since we apply the publicly-available pre-trained unit vocoder which mainly controls the naturalness of output speech and leave it unchanged, we ex-

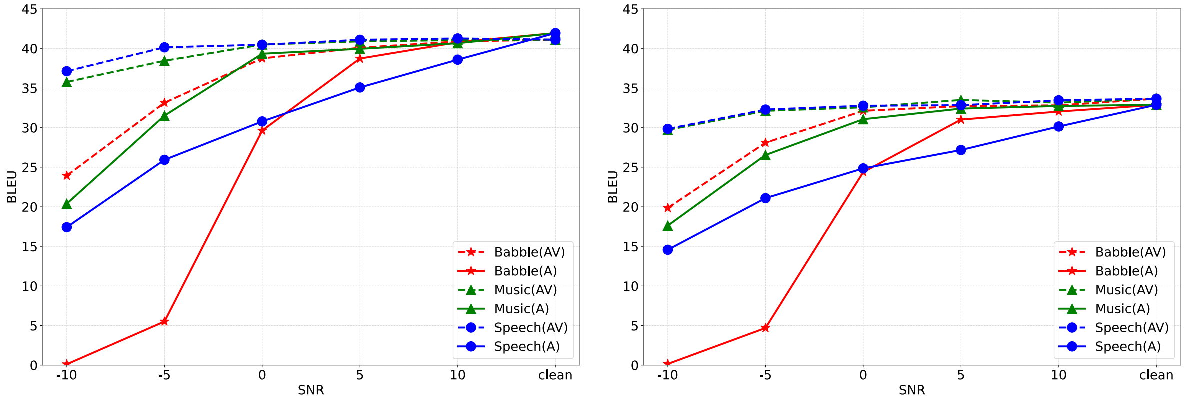



Figure 3: Illustration of model performance with different noise configurations and input modalities.

| Modality | Babble (SNR) | | | | | Music (SNR) | | | | | Speech (SNR) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | -10 | -5 | 0 | 5 | 10 | -10 | -5 | 0 | 5 | 10 | -10 | -5 | 0 | 5 | 10 |
| **En-Es Translation** | | | | | | | | | | | | | | | |
| AV | 23.9 | 33.1 | 38.7 | 40.0 | 40.8 | 35.7 | 38.4 | 40.4 | 40.8 | 41.0 | 37.1 | 40.1 | 40.4 | 41.0 | 41.2 |
| A | 0.1 | 5.5 | 29.6 | 38.7 | 40.7 | 20.3 | 31.5 | 39.3 | 39.9 | 40.6 | 17.4 | 25.9 | 30.7 | 35.0 | 38.5 |
| **En-Fr Translation** | | | | | | | | | | | | | | | |
| AV | 19.8 | 28.0 | 32.1 | 32.7 | 32.8 | 29.7 | 32.1 | 32.5 | 33.4 | 33.5 | 29.8 | 32.2 | 32.7 | 32.8 | 33.4 |
| A | 0.1 | 4.6 | 24.3 | 31.0 | 31.2 | 17.6 | 26.5 | 30.8 | 31.0 | 31.2 | 14.5 | 21.0 | 24.8 | 27.1 | 30.1 |

Table 3: Translation accuracy (BLEU scores (↑)) comparison among models with different noise configurations and input modalities.

pect AV-TranSpeech exhibits high-quality speech generation as baselines.

| ID | Audio | Modality | En-Es | En-Fr |
|---|---|---|---|---|
| **Finetune with 10 hours data** | | | | |
| 1 | / | AV | 7.2 | 6.0 |
| 2 | | V | 3.7 | 4.6 |
| 3 | Covost | AV | 21.5 | 24.4 |
| 4 | | V | 5.8 | 6.4 |
| **Finetune with 30 hours data** | | | | |
| 5 | / | AV | 13.0 | 11.5 |
| 6 | | V | 8.9 | 7.5 |
| 7 | Covost | AV | 22.2 | 28.3 |
| 8 | | V | 10.4 | 9.6 |

Table 4: Leveraging audio-only data for boosting the performance of visual systems (AV or V) in low-resource scenarios. Audio: S2ST model trained on audio-only data.

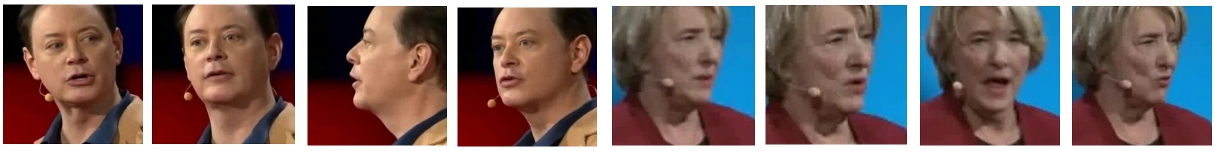
### 4.3 Visual Modality Evaluation

The benefit of incorporating the visual stream is more apparent in challenging scenarios (Afouras et al., 2018a; Popuri et al., 2022), and thus we evaluate our models in the noisy setting to examine the impact of input modality (audio or audio-visual). A noise category with an audio clip has been sampled each time, following which we randomly mix the sampled noise with varied probabilities at five SNR levels: $\{-10, -5, 0, 5, 10\}$ dB. For easy comparison, the results are presented in Table 3 and visualized in Figure 3, and we have the following observations:

1) AV-S2ST consistently outperforms audio-only S2ST under all settings regardless of the SNR and the type of noise. AV-TranSpeech complements the audio stream with visual information, which is invariant to speaking environments and promotes robustness. 2) As the volume of the noise increase with lower SNR, both languages have presented a degradation in translation accuracy. Informally, AV-S2ST models show a relatively slower BLEU drop (42% drop in SNR-10 babble noise), while a distinct decrease could be witnessed in audio-only S2ST models (99.9% drop in SNR-10 babble noise).

### 4.4 Low Resource Evaluation

Training direct AV-S2ST models without relying on intermediate text typically requires a large amount of parallel visual speech (i.e., lip) training data, while there may be very few resources due to the heavy workload. In this section, we prepare low-



| | |
|---|---|
| Source: | and the soldier on the front tank said we have unconditional orders to destroy this. |
| Target: | le soldat sur le char de front a dit que nous avions des ordres inconditionnels pour détruire cela. |
| V2ST: | le soldat **a** dit que nous avons des ordres **conditionnels à** détruit pour détruire ce. |
| S2ST: | le soldat sur le **premier ban** a dit que nous avons des ordres **non** conditionnels pour détruire cela. |
| Noisy S2ST: | **nous avons également des ordres constamment** pour détruire cela. |
| AV-S2ST: | le soldat sur le premier **réservant** a dit que nous avons des ordres inconditionnels pour détruire cela. |
| Noisy AV-S2ST: | le soldat sur **ledexpert** a dit que nous avons des ordres **nom policières** pour détruire cela. |
| Source: | we met men where they were at and we built a program. |
| Target: | nous avons rencontré des hommes là où ils étaient et nous avons construit un programme. |
| V2ST: | nous avons rencontré des hommes **lorsquils sétaient agis et que** nous avons construit un programme. |
| S2ST: | nous avons rencontré des hommes **quand** ils étaient **atteurs** et nous avons construit un programme. |
| Noisy S2ST: | **intéressante de la classe de mathématiques et de leur données de maternelle.** |
| AV-S2ST: | nous avons rencontré des hommes où ils étaient **atés** et nous avons construit un programme. |
| Noisy AV-S2ST: | nous avons rencontré des hommes **quand ils** étaient **atteurs** et nous avons construit un programme. |

Table 5: Two examples comparing translations produced by AV-TranSpeech with different modalities. The left video frames refer to the first example. We use the bond fonts to indicate the the issue of **noisy and incomplete translation.** We use SNR=0 with babble noise for both noisy scenarios.

resource audio-visual data (LRS3-T 10h, 30h) and leverage large-scale audio-only data (Covost) to boost the performance of visual systems (AV-S2ST, V2ST), to investigate the effectiveness of our cross-modal distillation. The results are compiled and presented in Table 4, and we have the following observations:

1) In consistent with previous practice (Duquenne et al., 2022; Tjandra et al., 2019), training speech models are faced with the significant issue of data scarcity. As training data is reduced in the low-resource scenario, a distinct degradation in translation accuracy could be witnessed in both modalities (AV-S2ST or V2ST). 2) Leveraging orders of magnitude audio-only data with cross-modal distillation, the visual systems achieve BLEU scores of 21.5 and 22.2 respectively in En-Es and En-Fr AV-S2ST, showing a significant promotion regardless of the modalities and languages. In this way, the dependence on a large number of parallel audio-visual data can be reduced for constructing visual systems.

### 4.5 Case Study

We present several translation examples sampled from the En-Fr language pair in Table 5, and have the following findings: 1) With the complemental visual information brought in, the results produced by AV-TranSpeech are noticeably more literal. 2) Moreover in challenging noisy scenarios, S2ST models suffer severely from the issue of *noisy and incomplete translation*, which is largely alleviated in AV-S2ST. AV-S2ST consistently outperforms audio-only S2ST in a noisy environment.

## 5 Conclusion

In this work, we proposed AV-TranSpeech, the first audio-visual speech-to-speech translation (AV-S2ST) model without relying on intermediate text. AV-TranSpeech complemented the audio stream with the visual information to promote robustness in noisy environments, opening up a host of practical applications: silent dictation or dubbing archival films. To mitigate the data scarcity for training AV-S2ST models, we 1) built upon the AV-HuBERT with a self-supervised learning framework fir contextual representations, showing that large-scale pre-training benefited the AV-S2ST training; 2) leveraged cross-modal distillation with S2ST models trained on the audio-only corpus, which further reduced the visual data requirements and boosted performance in low-resource scenarios. Experimental results on two language pairs demonstrated that AV-TranSpeech achieved significant robustness in noisy environments, outperforming audio-only S2ST models under all settings regardless of the type of noise. With low-resource audio-visual data (10h, 30h), cross-modal distillation yielded an improvement of 7.6 BLEU on average compared with baselines. We envisage that our work will serve as a basis for future audio-visual speech-to-speech translation studies, unlocking the ability for high-quality translation given a user-defined modality input.

## 6 Limitation and Potential Risks

As mentioned in our experimental setup, we provide results of AV-S2ST in LRS3-T with synthesized target speech, similar to the pioneer literature (Jia et al., 2022b) in S2ST. One of our future directions is to develop a better benchmark dataset (e.g., mined or human-annotated data) to improve translation performance.

As mentioned in our results analysis, the BLEU scores heavily depend on the ASR quality, which may not accurately reflect the speech translation performance. Future directions could be improving ASR quality or exploring other evaluation metrics without reliance on ASR models.

AV-TranSpeech lowers the requirements for audio-visual speech-to-speech translation, which may cause unemployment for people with related occupations such as interpreter and translator. In addition, there is the potential for harm from non-consensual voice generation or fake media. The voices of the speakers in the recordings might be overused than they expect.

## Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant No.2022ZD0162000, National Natural Science Foundation of China under Grant No.62222211, Grant No.61836002 and Grant No.62072397.

## References

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018a. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018b. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2020. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2143–2147. IEEE.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Zexin Cai, Chuxiong Zhang, and Ming Li. 2020. From speaker verification to multispeaker speech synthesis, deep transfer with feedback constraint. *arXiv preprint arXiv:2005.04587*.

Peng-Jen Chen, Kevin Tran, Yilin Yang, Jingfei Du, Justine Kao, Yu-An Chung, Paden Tomasello, Paul-Ambroise Duquenne, Holger Schwenk, Hongyu Gong, et al. 2022. Speech-to-speech translation for a real-world unwritten language. *arXiv preprint arXiv:2211.06474*.

I Chern, Kuo-Hsuan Hung, Yi-Ting Chen, Tassadaq Hussain, Mandar Gogate, Amir Hussain, Yu Tsao, Jen-Cheng Hou, et al. 2022. Audio-visual speech enhancement and separation by leveraging multi-modal self-supervised embeddings. *arXiv preprint arXiv:2210.17456*.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswani, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. 2022. Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations. *arXiv preprint arXiv:2211.04508*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2021. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3945–3954.

Rongjie Huang, Chenye Cui, Feiyang Chen, Yi Ren, Jinglin Liu, Zhou Zhao, Baoxing Huai, and Zhefeng Wang. 2022a. Singgan: Generative adversarial network for high-fidelity singing voice generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2525–2535.

Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. 2022b. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2023. Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*.

Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. In *Advances in Neural Information Processing Systems*.

Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. 2022c. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605.

Rongjie Huang, Zhou Zhao, Jinglin Liu, Huadai Liu, Yi Ren, Lichao Zhang, and Jinzheng He. 2022d. Transpeech: Speech-to-speech translation with bilateral perturbation. *arXiv preprint arXiv:2205.12523*.

Ye Jia, Yifan Ding, Ankur Bapna, Colin Cherry, Yu Zhang, Alexis Conneau, and Nobuyuki Morioka. 2022a. Leveraging unsupervised and weakly-supervised data to improve direct speech-to-speech translation. *arXiv preprint arXiv:2203.13339*.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2021. Translatotron 2: Robust direct speech-to-speech translation. *arXiv preprint arXiv:2107.08661*.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022b. Cvss corpus and massively multilingual speech-to-speech translation. *arXiv preprint arXiv:2201.03713*.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE.

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

Max WY Lam, Jun Wang, Rongjie Huang, Dan Su, and Dong Yu. 2021. Bilateral denoising diffusion models. *arXiv preprint arXiv:2108.11514*.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, et al. 2021a. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2021b. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2020. Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323. IEEE.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*.

Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. 2022. Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation. *arXiv preprint arXiv:2204.02967*.

KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman. 2022. Sub-word level lip reading with visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5162–5172.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022a. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*.

Bowen Shi, Abdelrahman Mohamed, and Wei-Ning Hsu. 2022b. Learning lip-based audio-visual speaker embeddings with av-hubert. *arXiv preprint arXiv:2205.07180*.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Themos Stafylakis and Georgios Tzimiropoulos. 2017. Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. Speech-to-speech translation between untranscribed unknown languages. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 593–600. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu. 2020. Uwspeech: Speech to speech translation for unwritten languages. *arXiv preprint arXiv:2006.07926*, 59:132.

Mingyang Zhang, Yi Zhou, Li Zhao, and Haizhou Li. 2021. Transfer learning from speech synthesis to voice conversion with non-parallel training data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1290–1302.

Shiliang Zhang, Ming Lei, Bin Ma, and Lei Xie. 2019. Robust audio-visual speech recognition using bimodal dfsmn with multi-condition training and dropout regularization. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6570–6574. IEEE.

## A Model Architectures

In this section, we list the model hyper-parameters of AV-TranSpeech in Table 6.

## B Subjective Evaluation

Following (Huang et al., 2021, 2022a), all our Mean Opinion Score (MOS) tests are crowd-sourced and conducted by native speakers. The scoring criteria have been included in Table 7 for completeness. The samples are presented and rated one at a time by the testers, each tester is asked to evaluate the subjective naturalness of a sentence on a 1-5 Likert scale. The screenshots of instructions for testers are shown in Figure 4. We paid $8 to participants hourly and totally spent about $500 on participant compensation.

Table 7: Ratings that have been used in the evaluation of speech naturalness of synthetic and ground truth samples.

| Rating | Naturalness | Definition |
|---|---|---|
| 1 | Bad | Very annoying and objectionable dist. |
| 2 | Poor | Annoying but not objectionable dist. |
| 3 | Fair | Perceptible and slightly annoying dist |
| 4 | Good | Just perceptible but not annoying dist. |
| 5 | Excellent | Imperceptible distortions |

| Hyperparameter | | AV-TranSpeech |
|---|---|---|
| Conformer Encoder | Encoder Layer | 24 |
| | Encoder Input/Output Dim | 1024 |
| | Encoder FFN Embed Dim | 4096 |
| | Encoder Attention Heads | 16 |
| | Encoder Dropout | 0.1 |
| Length Adaptor | Conv1d Layer | 1 |
| | Conv1d Kernel | 3 |
| | Conv1d Stride | 3 |
| Unit Decoder | Decoder Layer | 12 |
| | Decoder Input/Output Dim | 1024 |
| | Decoder FFN Embed Dim | 4096 |
| | Decoder Attention Headers | 16 |
| | Decoder Dropout | 0.1 |
| Total Number of Parameters | | 827 M |

Table 6: Hyperparameters of AV-TranSpeech.

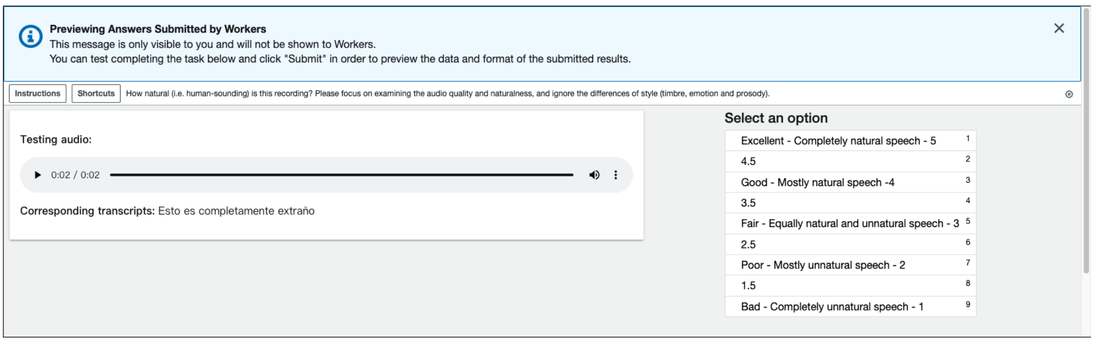



Figure 4: Screenshot of MOS testing.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*See Section 6.*

☑ A2. Did you discuss any potential risks of your work?
*See Section 6.*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*See Section 1.*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☒ Did you use or create scientific artifacts?

*Left blank.*

☐ B1. Did you cite the creators of artifacts you used?
*No response.*

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*No response.*

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*No response.*

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*No response.*

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*No response.*

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*No response.*

### C ☑ Did you run computational experiments?

*See Section 4.*

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*See Section 4.1.2 and Appendix A*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*See Section 4.1 and Appendix A.*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*See Section 4.1.2. We report BLEU on average and MOS with 95% confidence intervals (CI).*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*See Section 4.1*

**D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*See Section 4.1.3 and Appendix B.*

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*See Section 4.1.3 and Appendix B.*

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*See Section 4.1.3 and Appendix B.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Not applicable. Left blank.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Not applicable. Left blank.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Not applicable. Left blank.*