

A. AVMuST-TED

A.1. Details of AVMuST-TED

The dataset consists of over 706 hours of video, extracted from 4598 TED and TEDx talks in English. The visual speech corpus is provided as face-centered video in `.avi` files with a resolution of 224×224 and a frame rate of 25fps. The audio speech corpus is provided as the single-track, 16-bit 16kHz `.wav` files. Each pair of audio and video speech has its corresponding translation into other languages. Following the previous workflow [1, 65, 2] of visual-speech dataset acquisition, we fetch the complete face track from the massive data [35] and perform audio-visual synchronization testing to determine whether it is the face track of the speaker [13]. We take the four most amount of translation pairs, En-Es, En-Fr, En-It and En-Pt, from the numerous translation combinations of TED, and the detailed statistics in four different languages at AVMuST-TED are shown in Table 6.

Target Language	Hours	Sents	Vocab	Tokens
Spanish (Es)	198h	258K	95K	2.0M
French (Fr)	185h	244K	91K	1.9M
Italian (It)	165h	218K	95K	1.6M
Portuguese (Pt)	158h	205K	84K	1.5M

Table 6. Statistics in four different languages at AVMuST-TED.

A.2. Quality of Translated Texts

The translations in the AVMuST-TED dataset are taken directly from the high reliability translated subtitles in TED. TED has a very well-defined translation workflow to ensure that the translation accurately conveys the meaning, and we will now introduce it in detail. They recruit a total of 45,735 volunteers in 115 languages from all around the world, requiring each volunteer to be fluently bilingual in both source and target languages, fluent in the transcription language, and knowledgeable about what expressions are appropriate for subtitling. To ensure the quality of each assignment, each volunteer could apply for up to three editing assignments at the same time. Each volunteer can claim up to three editing assignments at a time to ensure the quality of each assignment. Each translation goes through three steps

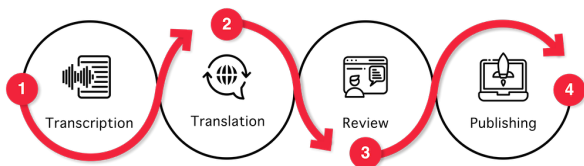


Figure 4. The TED translation workflow before publication.

En	It's a shared database
Es	Es una base de datos compartida
Fr	C'est une base de données partagée
It	È una base dati condivisa
Pt	É uma base de dados partilhada
En	That object was about 10 kilometers across
Es	Ese objeto tenía un diámetro de 10 km
Fr	Cet objet mesurait dix kilomètres de largeur environ
It	Quell'oggetto aveva un diametro di circa 10 chilometri
Pt	Esse objeto tinha cerca de 10 km de diâmetro
En	Can I correct my boss when they make a mistake?
Es	¿Puedo corregir a mi jefe cuando comete un error?
Fr	Puis-je corriger mon patron quand il fait une erreur ?
It	Posso correggere il mio capo quando fa un errore?
Pt	Posso corrigir o meu chefe quando ele comete um erro?
En	Now this turns out to be surprisingly common
Es	Ahora bien, esto resulta ser sorprendentemente común
Fr	Il s'avère que cela soit surprenamment commun
It	Ora questo risulta essere sorprendentemente comune
Pt	Isto parece ser surpreendentemente comum

Table 7. Examples of the source language transcription (En) and target language translation (Es, Fr, It, Pt) for audio-visual speeches (En) in AVMuST-TED.

of transcription, translation and review before publishing, as shown in Figure 4. TED provides an original transcript for all TED and TED-Ed content. For TEDx talks, volunteers are able to utilize auto-generated transcriptions as a base, or create their own from scratch. Subtitles are then translated from the original language into the target language, using a dynamic subtitle editor. Finally, before publication, subtitles are further reviewed by an experienced volunteer. In Table 7, we present some sample translations of AVMuST-TED.

B. Implementation Details

Audio and Visual Speeches Preprocessing. We follow the data preprocessing process in the prior work [53, 1] for audio and visual speeches. For visual speech, we only extract the lip region as visual speech input, first detecting 68 facial keypoints using dlib [31], and then aligning each face with the faces of its neighboring frames. From each visual speech utterance, we crop a 96×96 region-of-interest (ROI) lip-centered talking head video, representing the video speech. And for the audio speech, we also keep the same processing steps as the previous works [53, 39]. We extract 26-dimensional log filterbank energy feature from the raw waveform and stack 4 adjacent acoustic frames

Target Language	Method	Modality	BLEU						
			SNR					clean	
			-20 db	-10 db	0 db	10 db	20 db	Avg.	+∞
En-Es	Cascaded	A _(+Noise)	1.4±0.1	5.8±0.2	21.1±0.3	25.5±0.3	26.3±0.2	16.0	26.6
	AV-Hubert [53]	A _(+Noise)	1.5±0.2	6.7±0.2	22.3±0.4	27.7±0.2	28.6±0.3	17.6	28.9
	Cascaded	AV _(+Noise)	6.7±0.2	15.3±0.4	24.6±0.4	26.3±0.2	26.7±0.2	19.9	26.9
	AV-Hubert [53]	AV _(+Noise)	6.9±0.3	16.4±0.5	26.6±0.3	28.7±0.1	28.9±0.2	21.5	29.1
En-Fr	Cascaded	A _(+Noise)	1.3±0.2	4.5±0.3	16.6±0.4	20.9±0.3	21.3±0.1	12.9	21.7
	AV-Hubert [53]	A _(+Noise)	1.4±0.2	5.5±0.3	18.5±0.4	23.2±0.2	23.6±0.1	14.5	23.9
	Cascaded	AV _(+Noise)	4.6±0.1	11.4±0.5	19.4±0.3	21.5±0.2	22.0±0.2	15.8	22.3
	AV-Hubert [53]	AV _(+Noise)	4.9±0.2	12.1±0.3	21.6±0.4	23.7±0.3	24.3±0.1	17.3	24.6
En-It	Cascaded	A _(+Noise)	0.9±0.3	4.0±0.3	16.1±0.2	20.7±0.1	21.2±0.2	12.6	21.5
	AV-Hubert [53]	A _(+Noise)	1.0±0.2	5.1±0.5	18.3±0.3	22.7±0.2	23.6±0.2	14.1	23.8
	Cascaded	AV _(+Noise)	4.8±0.3	11.8±0.4	19.5±0.3	21.4±0.2	22.1±0.1	15.9	22.3
	AV-Hubert [53]	AV _(+Noise)	5.0±0.4	12.4±0.6	21.9±0.3	23.7±0.1	24.1±0.2	17.4	24.5
En-Pt	Cascaded	A _(+Noise)	1.1±0.3	5.4±0.5	20.1±0.4	24.9±0.2	26.0±0.1	15.5	26.2
	AV-Hubert [53]	A _(+Noise)	1.2±0.2	6.3±0.4	22.2±0.3	27.4±0.3	28.4±0.1	17.1	28.6
	Cascaded	AV _(+Noise)	5.8±0.4	13.8±0.6	23.5±0.4	25.8±0.2	26.3±0.1	19.0	26.4
	AV-Hubert [53]	AV _(+Noise)	6.1±0.3	15.5±0.4	26.0±0.3	28.2±0.3	28.6±0.2	20.9	28.8

Table 8. BLEU scores of audio speech translation and audio-visual speech translation with different noise SNRs.

together for syncing with visual speech. we randomly crop a region of 88×88 from the entire ROI and perform a horizontal flip with probability 0.5 for data enhancement. we also apply noise with a probability of 0.25 to each audio utterance from [55] as steps in the prior works [53, 1] for audio speech enhancement.

Training Details of MixSpeech. Our work is developed on the basis of the publicly available pre-trained model Transformer-Large of AV-Hubert [53], which has 24 Transformer-LARGE with the embedding dimension/feed-forward dimension/attention heads of 1024/4096/16. Concretely, we adopt here the Transformer-LARGE model trained on LRS3 [2] and VoxCeleb2 [12], augmented with noise. Correspondingly, for the translation decoder, we follow the same setup as AV-Hubert, with a 9-layer transformer decoder for easy comparison with it. During training, on one single 3090 GPU, we train 160K steps with labeled audio corpus, 80K of which are warmup steps; then we tune 40K steps with labeled visual corpus in the self-learning framework.

C. Experiment

C.1. Speech Translation with Noise

In this section, we show the detailed performance of speech recognition in noisy environments in Table 8. Although the discrimination of audio speech is excellent and

the performance of audio speech translation is outstanding, it is easily interfered by noise and the performance of audio speech translation decreases rapidly with the enhancement of noise interference. Following the previous works [1, 54], we add noise randomly sampled from MUSAN [55] to the audio speech and check the performance at five SNR levels $\{-20, -10, 0, 10, 20\}$ db. For each experiment, we performed ten times, calculating the mean and the error to avoid interference from random sampling. The experimental results show that the performance of audio-visual speech translation is better than that of speech translation with audio speech only on all four languages in the noise-free environment (*i.e.*, clear), demonstrating that visual speech further boosts the ceiling of speech recognition. Meanwhile, with the increase of noise interference (the smaller the SNR, the stronger the noise), the performance of audio speech translation decreases rapidly, especially during the process of SNR from 0db to -10db, the audio speech translation performance decreases most quickly, and the BLEU score decreases by -13.0 to -15.8. In contrast, speech translation with audio visual speech is significantly more resistant to noise, with the BLEU score decreasing by only -9.5 to -10.5 when SNR from 0db to -10db. At the same time, in terms of translation performance, all the audio-visual speech performances are better than those with only audio speech at the same SNR, and the audio-visual speech translation still performs well even at SNR = -10db, improving the robustness of the speech translation.





	
En-Es	En TRXN: that's why people often confuse me with a GPS.
	GT: por eso la gente me confunde a menudo con un gps
	A _(+N) : por eso la gente ayúdame a lo que me confunde a menudo con un gps alegra por favor
	Es V: por eso la gente a menudo me confunde con un gps los chimpancés
	A: por eso la gente a menudo me confunde con un (el) gps
AV: por eso la gente a menudo me confunde con un gps	
	
En-Fr	En TRXN: you need to understand that everyone who helps you on your journey
	GT: vous devez comprendre que tous ceux qui vous aident durant votre voyage
	A _(+N) : vous devez comprendre que tous ceux qui vous avoir partagé avec un adolescent et aident (aidé) durant ...
	Fr V: vous devez il faut comprendre que tous ceux (chacun) vous aident duran (aide) à votre voyage
	A: vous devez comprendre que tous ceux partout qui vous aident durant (aide dans) votre voyage (parcours)
AV: vous devez comprendre que tous ceux (chaque personne) qui vous aident durant (aide dans) votre voyage	
	
En-It	En TRXN: and one of our litigation strategies
	GT: e una delle nostre strategie in tribunale
	A _(+N) : e in una delle nostre strategie in tribunale di queste acque calde
	It V: e una delle nostre strategie in tribunale future eliminazioni
	A: e una delle nostre strategie in tribunale di contenzione
AV: e una delle nostre strategie in tribunale di (litigazione)	
	
En-Pt	En TRXN: and both of the finalists for the Democratic nomination
	GT: e ambos os finalistas para a nomeação democrática
	A _(+N) : e ambos os finalistas tenho estado à espera de um minuto para ereseer no meio duma pessoa a ...
	Pt V: e ambos (ambas) os finalistas as famílias democrática para a nomeação democracia
	A: e ambos (os dois) finalistas para a nomeação nação democrática
AV: e ambos (os dois) finalistas para a nomeação democrática	

Table 9. Qualitative performance of the four target languages on the AVMuST-TED. Among them, A_(+N) for noisy audio in the SNR of -10db, V for visual, A for audio and AV for audio-visual. **Red-Strikeout Words**: mistranslated words with opposite meaning, **(Blue Words in parentheses)**: mistranslated words with similar meaning, Gray Words: the absent words. TRXN: transcript in English. GT: Ground Truth in the target language.

C.2. More Qualitative Analysis

To further quantitatively demonstrate the enhancement of visual speech to speech translation, we show more samples from AVMuST-TED and their outcomes with different modality speech translation in Table 9.

Visual Speech VS Audio Speech with Noise Although the discrimination of visual speech is not as good as au-

dio speech, it is not interfered by noise, and we choose the translation of audio speech in the SNR of -10db to compare with that of visual speech.

Audio-Visual Speech VS Audio Speech The robustness of speech translation can be further enhanced with the visual speech based on audio speech in the manner of audio-visual speech translation.

D. Discussion

Ethical Discussion Based on audio speech translation, visual speech for translation further enriches the application scenarios of speech translation technology (in silent or noise-bearing scenarios), while increasing the reliability of speech translation with the manner of audio-visual speech translation. As a cross-lingual translation technology, speech translation can be applied to many online applications (*e.g.*, online medical, online education, *etc.*), contributing to the fairness of technology in disadvantaged areas. However, for visual speech, there could be some concerns about information leakage. But in fact, as we have mentioned before, lip reading and lip translation can only perform with high-definition, high-frame-rate frontal face videos that ensures clear visibility of lips and lip movements. Typically, only specially recorded videos, such as those from online meetings and public presentations, meet the strict video conditions that guarantee the unavailability of visual speech from videos such as surveillance for information leakage.

Limitations Discussion In this paper, we focus on the association between audio-visual speech and do not discuss the effect of machine translation datasets on lip translation yet. Many previous speech translation works have sufficiently demonstrated the enhancement of machine learning datasets for audio speech translation, and we have reasons to believe that it can also greatly improve the performance of lip translation, so there is no detailed discussion about it in this paper. Correspondingly, this paper focuses on a topic that has never appeared in other speech translation tasks, the interaction between audio-visual speech. Our follow-up work will address the blanks of this work.