MixSpeech: Cross-Modality Self-Learning with Audio-Visual Stream Mixup for Visual Speech Translation and Recognition

Xize Cheng*, Tao Jin* , Rongjie Huang* , Linjun Li
Wang Lin , Zehan Wang , Ye Wang , Huadai Liu , Aoxiong Yin , Zhou Zhao†
Zhejiang University
{chengxize,jint_zju,rongjiehuang,zhaozhou}@zju.edu.cn

Abstract

*Multi-media communications facilitate global interaction among people. However, despite researchers exploring cross-lingual translation techniques such as machine translation and audio speech translation to overcome language barriers, there is still a shortage of cross-lingual studies on visual speech. This lack of research is mainly due to the absence of datasets containing visual speech and translated text pairs. In this paper, we present **AVMuST-TED**, the first dataset for **A**udio-**V**isual **Mu**ltilingual **S**peech **T**ranslation, derived from **TED** talks. Nonetheless, visual speech is not as distinguishable as audio speech, making it difficult to develop a mapping from source speech phonemes to the target language text. To address this issue, we propose MixSpeech, a cross-modality self-learning framework that utilizes audio speech to regularize the training of visual speech tasks. To further minimize the cross-modality gap and its impact on knowledge transfer, we suggest adopting mixed speech, which is created by interpolating audio and visual streams, along with a curriculum learning strategy to adjust the mixing ratio as needed. MixSpeech enhances speech translation in noisy environments, improving BLEU scores for four languages on AVMuST-TED by +1.4 to +4.2. Moreover, it achieves state-of-the-art performance in lip reading on CMLR (11.1%), LRS2 (25.5%), and LRS3 (28.0%).*

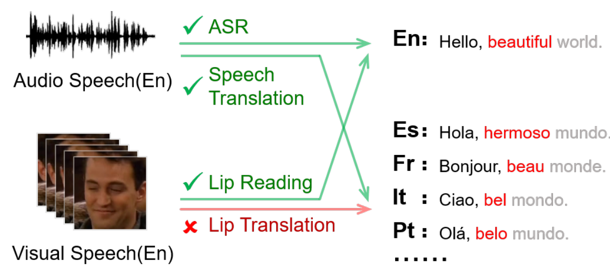


Figure 1. Diagram of speech tasks. Audio speech and visual speech are paired parallel speech streams which can be employed for speech recognition and speech translation. However, only Lip-Translation remains unexplored.

1. Introduction

Multi-media techniques, including Audio-Visual Speech Recognition (AVSR) [4, 1, 2, 53], Audio-Visual Speech Translation (AVST) [8, 38, 61], and Audio-Visual Speech Generation (AVSG) [48, 32, 24], are commonly employed in various online communication scenarios, such as conferences, education, and healthcare. As a tool for ultra-remote communication, many online interactions involve multiple languages, prompting the need for addressing cross-lingual challenges. Several works have attempted to tackle these challenges, including Machine Translation (MT) [9, 37, 15] for text utterance, Speech Translation (ST) [58, 19] for audio utterance, and Speech-to-Speech Translation (S2ST) [58, 19, 17, 33, 29] for simultaneous interpretation. However, research on cross-lingual visual speech is still limited, as illustrated in Figure 1. As an essential component of multi-media speech, visual speech can be combined with audio to enhance the recognition and understanding of speech content as audio-visual speech [1, 2, 54], and is the unique resource for speech content understanding in audio-disabled scenarios [36].

Visual speech translation has never been studied, mainly for the lack of visual speech datasets with translated texts in different languages. The few remaining works [57, 60, 44] also cannot be quantitatively verified for this reason, making them unconvincing. The available visual speech corpus is often very scarce compared to audio speech owing to the high demands of visual speech for model training, which requires mostly-frontal and high-resolution videos with a sufficiently high frame rate, such that motions around the lip area are clearly captured [23]. In this paper, we propose the

*Equal contribution.
†Corresponding author.

first Audio-Visual Multilingual Speech Translation dataset, AVMuST-TED. During the process of acquisition, we first screen out videos with professional translations in four different languages from TED talk which performs strict translation and review processes, and then determine the real speaker's talking head by checking whether each pair of visual speech (*i.e.*, talking head) and audio speech matches in the manner of [1, 2]. Incidentally, this dataset can also be used for quantitative evaluation of other multi-modality translation tasks, such as cross-lingua audio-visual speech generation [51, 60].

The cascaded model comprising of a speech recognition model and a machine translation model can handle speech translation tasks but suffers from error accumulation due to model cascades and cannot process languages without text (*e.g.*, Minnan). Our proposed end-to-end model, which can translate directly from source speech to target text, addresses the above issues. However, visual speech is less distinguishable than audio speech, making it difficult to develop a mapping from source speech phonemes to the target language text. To address this, we introduce MixSpeech, a method that first pretrains the decoder using high-discrimination audio speech to obtain a mapping from speech phonemes to text and then generalizes this mapping to the visual speech task through cross-modality self-learning. Furthermore, since audio speech and visual speech are two distinct modalities of speech, there is a significant modality gap between them that hinders knowledge transfer. To narrow this gap and improve knowledge transfer, we propose mixed speech, which is created by interpolating audio and visual streams, rather than relying solely on audio speech. We also propose a curriculum-learning [7] based strategy to adjust the mixing ratio as the training progresses and cross-modality integration deepens.

The code and dataset are available[1], the main contributions of this paper are as follows:

- We present the first lip-translation baseline and introduce the Audio-Visual Multilingual Speech Translation dataset, AVMuST-TED.
- We present a cross-modality self-learning framework that leverages distinguishable audio speech translation to regularize visual speech translation for effective cross-modality knowledge transfer.
- We present to adopt the mixed speech, interpolated from audio and visual speeches, and a curriculum-learning based mixing ratio adjustment strategy to reduce the inter-modality gap during knowledge transfer.
- We achieve state-of-the-art performance in lip translation for four languages on AVMuST-TED, with a +1.4 to +4.2 boost in BLEU scores and in lip reading on CMLR (11.1%), LRS2 (25.5%) and LRS3 (28.0%).

[1] https://github.com/Exgc/AVMuST-TED

2. Related work

2.1. Audio-Visual Speech

Audio and visual speeches are two separate modalities that convey speech content. Numerous works [45, 13, 1, 2, 47, 25, 28, 26, 34, 11] have explored ways to extract information from speech using these modalities. Speech recognition [45, 6, 22] is widely used in online meetings and social applications to recognize speech content. Speech translation [58, 66, 19] is commonly used in simultaneous interpretation applications for cross-lingual communication in cross-border travel and meetings. Keyword spotting [5, 52, 30] is employed in short video applications to quickly retrieve relevant content. Additionally, in noisy scenarios, relevant speech tasks [14, 21, 47, 42] rely on visual speech to avoid interference from surrounding speech and background noise. Despite the growing interest in speech tasks that rely on visual speech, researches [57, 60] on visual speech translation are limited and lacks validation due to the lack of multilingual audio-visual speech transcription datasets. This paper proposes a baseline for visual speech translation and introduces the first large-scale audio-visual multilingual translation dataset, AVMuST-TED, which includes 706 hours of audio-visual speech and translation pairs in Spanish, French, Italian, and Portuguese. AVMuST-TED lays a solid foundation or future cross-lingual audio-visual translation tasks, such as Cross-Lingual Talking Head Generation [44].

2.2. Transfer learning from Audio to Visual

Many researchers [50, 53, 39, 64] attempt to enhance the representation of visual speech by leveraging corresponding audio speech, as the two are paired parallel speech streams. Some [50, 53, 39] use knowledge distillation to bootstrap the training of visual speech models using audio speech models, while others [71, 39] have proposed various distillation strategies to optimize the representation of visual speech by mining the intrinsic connection between audio and visual speeches. Some [53] also use self-supervised learning, with audio as auxiliary supervision for visual utterances, to obtain fine-grained visual representations. The success of these works demonstrates the critical role of audio speech, which has a higher discrimination compared to visual speech, in training visual speech models. However, previous works face the modality shift problem during knowledge transfer because they start directly from speeches of two different modalities, audio and visual speeches, with a significant modality gap. In this paper, we propose an cross-modality self-learning framework MixSpeech, that uses synthetic mixed speech to regularize visual speech translation for effective cross-modality knowledge transfer, reducing the gap between the two modalities during knowledge transfer.

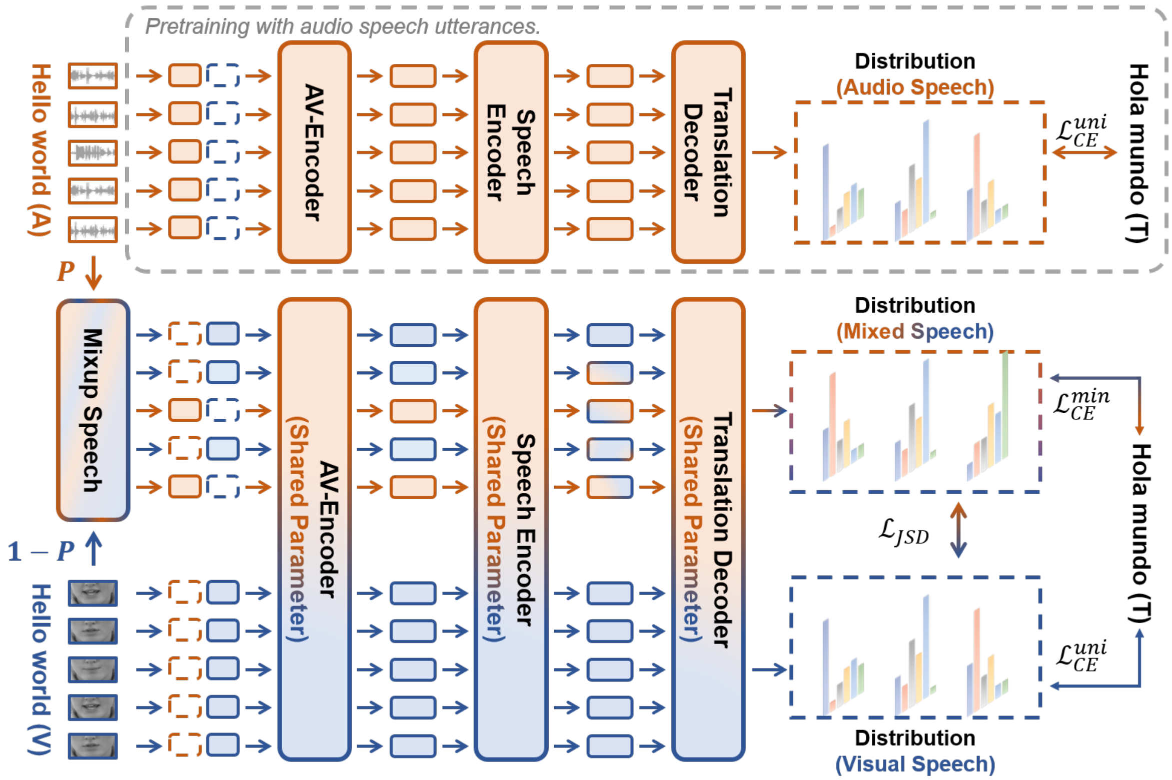


Figure 2. Illustration of our proposed MixSpeech. We first pretrain the model with audio speech translation as shown in the dashed boxed, and then train the visual speech translation with mixed speech regularization. The blank dashed boxes denote the modality missing speech.

2.3. Mixup for Cross-Modality Transfer

Many works [68, 20, 63, 19, 24] bridge the gap between modalities with mixup. [67] proposes mixup for data augmentation to improve model robustness. [10] suggests mixing at the representation-level to mine implicit associations between labeled and non-labeled sentences. Other works [63, 59, 18, 27] also use mixing to build bridges between different modalities. Some [63, 18] use CLIP [49] to retrieve semantically consistent images with text tokens and synthesize mixed sentences for text-visual consistency representation training. Others [59, 19] construct manifold mixup interpolations based on semantic consistency between audio and text to enhance understanding of audio with textual datasets. By implementing the mixup strategy, these studies have shown notable improvements across a range of tasks, highlighting its potential to facilitate knowledge transfer between different modalities. However, previous works use fixed hyperparameters [67] or mapping functions [19] for mixing ratios, which are typically not optimal and cannot be adapted to the training situation. In this paper, we propose an uncertainty-based [43] curriculum learning [7] strategy that gradually adjusts mixing ratios and apply mixup strategy for cross-modality knowledge transfer between audio and visual speeches for the first time.

3. Method

3.1. Task Formulation

As the twin task of speech recognition, speech translation involves translating source language speech into target language text. The speech translation model takes audio speech utterance $\mathbf{A}=\{\mathbf{A}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times D}$ or visual speech utterance $\mathbf{V}=\{\mathbf{V}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times D}$ as input and generates the target language text $\mathbf{w}=\{\mathbf{w}_i\}_{i=0}^{s}$, where $\mathbf{A}_t$ and $\mathbf{V}_t$ represent the t-th features in the audio and visual speeches, and $\mathbf{w}_i$ represents the i-th word in the target language translation with a total length of S. Note that, we stack 4 adjacent acoustic frames together for syncing with visual speech, both with T frames.

3.2. Overview

We propose a cross-modal self-learning framework for visual speech translation with audio speech regularization, named MixSpeech, as illustrated in Figure 2. This model consists of three modules – a feature extractor for extracting speech embeddings, a speech encoder for attending to the contextual dependencies of speech, and a target language-oriented translation decoder. We utilize the pre-trained feature extractor (`AV-Encoder`) and speech encoder (`Speech-Encoder`) from the AV-Hubert [53] to extract speech representations from both audio and visual speech utterances. Additionally, a randomly initialized translation decoder (`Trans-Decoder`) is used to autoregressively decode the speech representation into the target language text. MixSpeech is a two-stage training process: 1) Pretraining the translation decoder with high-discrimination audio speech utterances to learn inter-lingual mapping relations between source language phonemes and target language text, as detailed in subsection 3.3. 2) Aligning visual speech with audio speech to transfer the inter-lingual mapping from audio speech to visual in 3.5. The mixed speech 3.4 is synthesized by interpolating audio speech with visual speech in `MixupSpeech`, bridging the modality gap and enhancing knowledge transfer.

3.3. Pretraining with Audio Speech

For uni-modality audio speech $\mathbf{A} \in \mathbb{R}^{T\times D}$ or visual speech $\mathbf{V} \in \mathbb{R}^{T\times D}$, the uni-modality audio-visual feature $\mathbf{e}^u$=$\{\mathbf{e}_t^u\}_{t=1}^T \in \mathbb{R}^{T\times 2D}$ fed into feature extractor can be defined as:

$$\mathbf{e}_t^u = \begin{cases} \texttt{concat}(\mathbf{0}_D, \mathbf{V}_t), & \mathbf{V}_t \neq \text{None}, \\ \texttt{concat}(\mathbf{A}_t, \mathbf{0}_D), & \mathbf{A}_t \neq \text{None}, \end{cases} \tag{1}$$

where $\mathbf{0}_D$ denotes the feature of missing modality, following the practice of [53]. And then, we obtain the audio-visual fusion feature $\mathbf{e}^f \in \mathbb{R}^{T\times D}$ with `AV-Encoder`. The transformer-based `Speech-Encoder` allows us to obtain the phoneme embedding $\mathbf{e}^p \in \mathbb{R}^{T\times D}$ with the contextual speech details. A target language oriented translation decoder `Trans-Decoder` is appended to autoregressively decode the phoneme embedding $\mathbf{e}^p$ into the target probabilities P^u, where P^u=$\{P_t^u\}_{t=1}^S$=$\{p(\mathbf{w}_t|\{\mathbf{w}_i\}_{i=1}^{t-1}, \mathbf{e}^p)\}_{t=1}^S$ represents the probability of the t-th word being $\mathbf{w}_t$ when the previous $t-1$ predictions are $\{\mathbf{w}_i\}_{i=1}^{t-1}$ and s is the length of the target language translation. During the pretraining, the overall model is trained on audio speech with cross-entropy loss :

$$\mathcal{L}_{CE} = -\sum_{t=1}^{S} \log p(w_t|\{w_i\}_{i=1}^{t-1}, \mathbf{e}^p). \tag{2}$$

3.4. Audio-Visual Speech Mixing

Audio and visual speeches have a huge modality gap, which greatly impacts knowledge transfer across modalities. We attempt to employ mixed speech to bridge two different modalities of speech. Since the pair of audio and visual speeches is strictly temporally synchronous, we take advantage of this property to interpolate the mixed speech. For a pair of synchronized audio and video speech $(\mathbf{A}, \mathbf{V}) \in \mathbb{R}^{2\times T\times D}$, each visual feature $\mathbf{V}_t$ at t-th frame has its corresponding audio feature $\mathbf{A}_t$, representing the same phonetic content. We interpolate with probability ϕ to obtain a mixed speech $\mathbf{e}^m$=$\{\mathbf{e}_t^m\}_{t=1}^T \in \mathbb{R}^{T\times 2D}$ derived partly from audio speech and partly from visual speech:

$$\mathbf{e}_t^m = \begin{cases} \texttt{concat}(\mathbf{0}_D, \mathbf{V}_t), & p < \phi, \\ \texttt{concat}(\mathbf{A}_t, \mathbf{0}_D), & p \geq \phi, \end{cases} \tag{3}$$

where p is sampled from the uniform distribution $U(0, 1)$ and ϕ is the ratio of speech mixing. In particular, we propose a curriculum learning [7] based mixing ratio adjustment method that adapts the appropriate ϕ as the training progresses. The prediction uncertainty [43] indicates the confidence of the prediction (smaller is better), and we take it as a signal to adjust the mixing ratio:

$$\mathbf{u} = \frac{1}{S}\sum_{t=1}^{S} \texttt{Entropy}(P_t). \tag{4}$$

If the discrimination of mixed speech is insufficient to regularize visual speech translation and maintain n steps ($\Delta\mathbf{u}$=$\mathbf{u}^v$−$\mathbf{u}^m$<$k\mathbf{u}^v$, where $\mathbf{u}^v$ and $\mathbf{u}^m$ represent the uncertainty of uni-modality (visual) and mixed speech, respectively, and the threshold hyperparameter k is set to 0.05 with n set to 20 in our work), we gradually increase the proportion of audio at a rate of α (ϕ'=$\alpha\,\phi$). We initialize ϕ=0.1 to prevent excessive initial modality gap and maintain $\phi \in [0.1, 0.9]$ throughout the training process.

3.5. Cross-Modality Self-Learning for Speech

Since audio speech is more distinguished compared to visual speech, we intend to boost visual speech translation with the knowledge from audio speech. And the mixed speech bridges the gap between audio speech and visual speech, allowing us to boost cross-modality knowledge transfer with it. With audio speech feature $\mathbf{A} \in \mathbb{R}^{T\times D}$ and visual speech feature $\mathbf{V} \in \mathbb{R}^{T\times D}$ fed into the modules with shared parameters, the uni-modality visual speech feature $\mathbf{e}^u$ and the mixed speech feature $\mathbf{e}^m$ are decoded into the target probabilities P^u and P^m, respectively.

After the pre-training with audio speech translation, the model is promising enough for mixed speech containing partial audio speech, we adopt the Jensen-Shannon Divergence (JSD) [41] to regularize the probabilities of these two different speeches:

$$\mathcal{L}_{JSD} = \sum_{t=1}^{S} JSD(P_t^m \| P_t^u). \tag{5}$$

As this probability is across the entire training vocabulary, we are able to perform fine-grained regularization to enhance the training of visual speech. Meanwhile, we also minimize the cross-entropy loss between two speech translations and the real translation, $\mathcal{L}$=$\mathcal{L}_{CE}^{uni}$+$\lambda_1\mathcal{L}_{CE}^{mix}$+$\lambda_2\mathcal{L}_{JSD}$, where λ_1 and λ_2 are hyperparameters of loss weights, while λ_1=λ_2=1.0 in this work.

Dataset	Target Language Hours					# Lang	# $\sum$ Hrs	# $\sum$ Sents	# $\sum$ Tokens	
	En	Es	Fr	It	Pt				src	tgt
Audio-Only										
LibriSpeech [45]	960h	-	-	-	-	1	960h	180K	5.9M	5.9M
MuST-C [16]	-	504h	492h	465h	385h	8	3 617h	2 016K	38.1M	35.8M
VoxPopuli [62]	543h	441h	427h	461h	-	16	5 967h	2 045K	65.0M	60.1M
Audio-Visual										
LRS2 [1]	224h	-	-	-	-	1	224h	143K	2.3M	2.3M
LRS3 [2]	433h	-	-	-	-	1	433h	151K	4.2M	4.2M
AVMuST-TED (ours)	-	198h	185h	165h	158h	4	706h	925K	7.3M	7.0M

Table 1. Comparison of audio-visual speech recognition/translation datasets. #Lang denotes the number of target languages. **#$\sum$ Hrs** denotes the overall duration of speech in the dataset, #Sents and #Tokens denote the overall sentences and the overall token, respectively.

4. Experiments

4.1. Datasets

AVMuST-TED. To obtain a corpus for AVST, we screened a set of TED and TEDx talks with multilingual subtitles as the data source. All transcriptions and translations are performed strictly following the TED Translation Guidelines and require collaboration between at least one translator (or transcriber) and one reviewer. The prior lip-reading dataset acquisition pipeline is followed to crop face-tracks, and an audio-visual alignment network, SyncNet, is adopted for speaker proofreading. Table 1 compares AVMuST-TED with related datasets, and it is the first audio-visual speech translation dataset containing translations from English (En) to four target languages: Spanish (Es), French (Fr), Italian (It), and Portuguese (Pt). These four languages have the most translated subtitles in TED, and 1024/1536 pieces of data are randomly sampled for each language as the *test*/*validation* set. The information about AVMuST-TED is detailed in Appendix A.

LRS2&3 [1, 2], two commonly used publicly available English wild audio-visual speech recognition datasets, are adopted to demonstrate the lip-reading performance, containing 224 hours of video from BBC television shows and 433 hours of video from TED and TEDx talks. The training data in both datasets is divided into two partitions, namely *Pretrain* and *Train*, both of which are transcribed from videos to text at the sentence level. The only difference is that the video clips in the *Pretrain* partition are not strictly trimmed and sometimes longer than the corresponding text. In our experiments, we employ different amounts of training data from LRS2 and LRS3, including *Pretrain*+*Train* (224/433h) for high resource and *Train* (29/30h) for low resource.

CMLR [70], widely used dataset for Mandarin audio-visual speech recognition, contains 61 hours audio-visual speech utterances collected from Chinese TV stations. In our experiments, we adopt this dataset to demonstrate the performance of our proposed MixSpeech in low-resource languages such as Mandarin. Additionally, we sample a training set containing only 12 hours of utterances in the manner of [71] for low resource scenario.

Method	M	BLEU ↑			
		En-Es	En-Fr	En-It	En-Pt
Cascaded	V	12.7	11.3	11.5	13.2
AV-Hubert [53]	V	14.2	12.6	12.9	14.8
Cascaded	A `(+Noise)`	16.0	12.9	12.6	15.5
AV-Hubert [53]	A `(+Noise)`	17.6	14.5	14.1	17.1
MixSpeech(ours)	V	**18.5**	**15.1**	**14.3**	**17.2**

Table 2. Comparison of the performances of visual speech translation on AVMuST-TED with those of the noisy audio speech translation. The results of noisy audio speech translation are the mean value at five SNRs {-20, -10, 0, 10, 20}db.

4.2. Evaluation and Implementation Details

In this paper, we measure the performance of MixSpeech on two speech tasks, speech recognition and speech translation. For speech recognition, word error rate (WER) is adopted as the evaluation metric, which is defined as $\texttt{WER}=(S+D+I)/M$, where S, D, I, M represent the number of words replaced, deleted, inserted, and referenced. As for speech translation, the case-sensitive detokenized BLEU score is computed using SacreBLEU [46], following the same evaluation methodology as in previous speech translation works [16, 62]. The implementation details are provided in Appendix B due to page limitations.

4.3. Performance of Speech Translation

End-To-End Models VS. Cascaded Models. Table 2 presents a comparison of the lip translation performance between two representative methods: 1) an end-to-end model, implemented based on the state-of-the-art AV-Hubert [53] method for visual speech-related tasks, and 2) a cascaded model, combining a speech recognition model (*i.e.*, Lip-Reading or ASR) with a machine translation model. In the cascaded model, we use the speech recognition model trained by AV-Hubert on LRS3, which achieve the best lip-reading performance to date, and a transformer-based machine translation model trained on the paired translated text corpus in AVMuST-TED. Comparing the lip translation performance of the end-to-end model and the cascade model, we find that the BLEU score of the end-to-end model improved by +1.3 to +1.6. This result demonstrates that the end-to-end trained model can effectively prevent the accumulation of errors caused by the model cascade, and that lip translation cannot be simply disassembled as the superposition of lip reading and machine translation.

MixSpeech VS. Prior Methods. Due to the discrimination of speech between modalities, visual speech models are not able to translate speech content as accurately as audio speech models. To address the issue of low discrimination in visual speech, we propose MixSpeech, which is a cross-modality self-learning framework that employs mixed speech to transfer knowledge obtained from audio speech pre-training into the visual speech model. Our proposed MixSpeech significantly improves the BLEU score by another +1.4 to +4.3. Furthermore, the improvement from MixSpeech is related to the discrepancy in speech translation between audio and visual modalities. For example, En-Es exhibits a larger discrepancy of 14.7 between audio and visual speech translation, ranging from 28.9 to 14.2, and MixSpeech significantly improves it by +4.3. Conversely, Italian shows a smaller discrepancy of 10.9, ranging from 23.8 to 12.9, and improves only by +1.4. This highlights that the improvement in lip translation stems from the knowledge acquired from audio speech translation.

Visual Speech VS. Noisy Audio Speech. We also evaluate the performance of audio speech translation in noisy environments, by adding noise sampled from MUSAN [55] to the audio speech and measuring the performance at five SNR levels {-20, -10, 0, 10, 20}db. We compare the average BLEU scores of different SNRs and present the detailed performance in Appendix C.1. Our experiments show that although noisy audio speech performs better than visual speech, the translation performance is still significantly lower compared to noiseless audio speech. In contrast, MixSpeech, which fully leverages the knowledge of audio speech, greatly improves the visual speech translation performance, making it more reliable in noisy scenes. We also provide a comparison of translation with audio speech and audio-visual speech, demonstrating that visual speech enhances the ceiling and robustness of speech translation, but the details are only available in the Appendix C.1 since audio-visual speech does not require the cross-modality knowledge transfer proposed in this paper.

# RES	Method	WER$_{(Labeled\ Visual\ Utts\ Hrs)}$ ↓		
		CMLR	LRS2	LRS3
High	WAS [56]	$38.9_{(61)}$	$70.4_{(224)}$	-
	TM-seq2seq [1]	-	$49.8_{(698)}$	$59.9_{(698)}$
	CSSMCM [70]	$32.5_{(61)}$	-	-
	Conv-seq2seq [69]	-	$51.7_{(698)}$	$60.1_{(698)}$
	CTC+KD [3]	-	$51.3_{(224)}$	$58.9_{(433)}$
	LIBS [71]	$31.3_{(61)}$	$65.3_{(698)}$	-
	CTCH [40]	$22.0_{(61)}$	-	-
	Master [50]	-	$49.2_{(698)}$	$59.0_{(698)}$
	Sub-Word [47]	-	$28.9_{(698)}$	$40.6_{(698)}$
	†AV-Hubert [53]	$12.7_{(61)}$	$28.7_{(224)}$	$28.6_{(433)}$
	MixSpeech(ours)	$\mathbf{11.1}_{(61)}$	$\mathbf{25.5}_{(224)}$	$\mathbf{28.0}_{(433)}$
Low	LIBS [71]	$50.5_{(12)}$	-	-
	†AV-Hubert [53]	$25.8_{(12)}$	$31.4_{(29)}$	$32.5_{(30)}$
	MixSpeech(ours)	$\mathbf{18.5}_{(12)}$	$\mathbf{26.9}_{(29)}$	$\mathbf{28.6}_{(30)}$

Table 3. Comparison of lip reading methods under different resource conditions. # RES represents the amount of resources. (Hours) highlighted in blue are used for low resources. † For better comparison, we reproduce AV-Hubert on CMLR and LRS2.

4.4. Performance of Speech Recognition

As shown in Table 3, we compare the performance of MixSpeech on another visual speech task, lip reading (*i.e.*, Visual Speech Recognition), to highlight the mixspeech from more perspectives. MixSpeech obtain state-of-the-art performance on three datasets, two for English (25.5% on LRS2 and 28.0% on LRS3) and one for Chinese (11.1% on CMLR), demonstrating that this cross-modality self-learning framework can be applied for different languages to capture the intrinsic association between audio and visual speeches and thus effectively improve the understanding of visual speech. Since visual speech is relatively low-resource, we verify whether MixSpeech can effectively improve the performance of visual speech tasks in low-resource with audio speech. Compared with previous methods, MixSpeech boosts the WER of lip-reading by -3.9% to -7.3%, highlighting the critical role of high-resource audio speech in low-resource visual speech tasks. Specifically, on LRS2 and LRS3, the performance of Mixspeech in the low-resource scenario (26.9%/28.6% WER obtained with only 29h/30h visual utterances) outperforms the performances of prior methods in the high-resource scenario (28.7%/28.6% obtained with 224h/433h or even more visual utterances). Even though with only limited labeled visual corpus, our proposed MixSpeech performs no less than works with more. It is the bridge between two modalities of speech, which helps visual speech to access the knowledge stored in high-resource and high-discrimination audio speech without barriers.

4.5. Can MixSpeech bridge cross-modality speech?

Our proposed MixSpeech builds a bridge between cross-modality speech through cross-modality self-learning, with the properly mixes speech. The details are as follows:

Cross-Modality Self-Learning for Knowledge Transfer. The experiments in Figure 3 provide a positive answer to the question of whether MixSpeech can contribute to achieving knowledge transfer between audio and visual speeches. We evaluate the performance of visual speech translation with different regularization strategies: no audio speech regularization (*i.e.*, $\phi = 0$), mixed speech regularization with different mixing ratios (*i.e.*, $\phi \in (0, 1)$), audio speech regularization (*i.e.*, $\phi = 1$), and mixing ratio adjustable mixed speech regularization (*i.e.*, dashed lines). It is evident that the cross-modality self-learning framework significantly enhances visual speech translation, as all performances with audio speech regularization are noticeably better than those without self-learning ($\phi = 0$), demonstrating the effectiveness of our proposed MixSpeech.

Narrow the Cross-Modality Distance with Properly Mixed Speech. Moreover, the introduction of mixed speech facilitates smoother cross-modality knowledge transfer by narrowing the modality gap between speeches. Some segments in the mixed speech come from the visual speech, making it much closer to visual speech in terms of modality distance than audio speech. When regularizing with mixed speech in En-Es, the translation performance of visual speech improves further by +0.3 to +0.8 compared to audio speech regularization alone. Among them, bootstrapping with mixed speech of mixing ratio $\phi = 0.5$ achieves the highest BLEU score of 18.3. This demonstrates that a reasonably mixed ratio ensures that it is neither overly biased towards visual speech, leading to a lack of knowledge of audio speech, nor overly biased towards audio speech, leading to excessive cross-modality distances that affect knowledge transfer. The adjustable mixing ratio strategy based on curriculum learning further increases the applicability of mixed speech to cross-modality self-learning training, thereby boosting visual speech translation performance again.

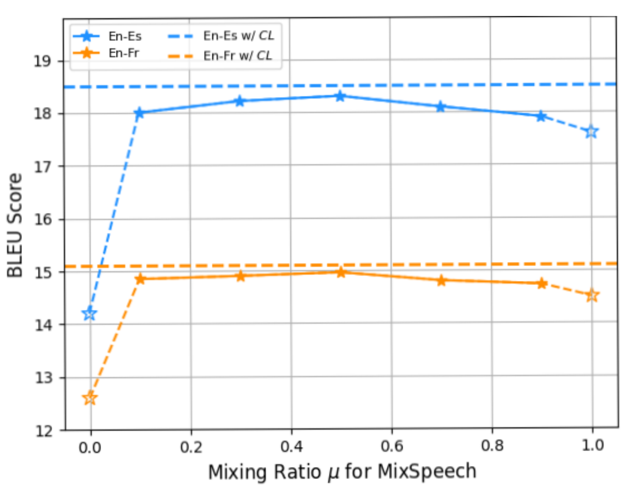


Figure 3. BLEU scores of MixSpeech with different speech regularization on En-Es and En-Fr. $\phi = 0$: no audio speech regularization, $\phi \in (0, 1)$: mixed speech regularization, $\phi = 1$: only audio speech regularization. The dashed lines represent the adjustable mixing ratio strategy based on curriculum learning.

ID	Method			BLEU ↑			
	$\mathcal{L}_{CE}^{mon}$	$\mathcal{L}_{CE}^{mix}$	$\mathcal{L}_{JSD}$	**En-Es**	**En-Fr**	**En-It**	**En-Pt**
#1	✔			14.2	12.6	12.9	14.8
#2	✔	✔		17.5	14.3	13.6	16.5
#3	✔		✔	18.1	14.8	14.1	16.9
#4	✔	✔	✔	**18.5**	**15.1**	**14.3**	**17.2**

Table 4. BLEU of different module combinations in MixSpeech.

4.6. What role does each part play in MixSpeech?

The effectiveness of MixSpeech, which is a cross-modality self-learning framework designed to improve visual translation performance, has been demonstrated. In this study, we investigate the role of each component in detail and present relevant experiments in Table 4:

Bridging the cross-modality gaps. We observe a significant improvement in the lip translation performance with the inclusion of $\mathcal{L}_{JSD}$ (ID: #3, #4) for regularizing the probabilities of visual speech and mixed speech, compared to without (ID: #1, #2). Specifically, experiment #3 with $\mathcal{L}_{JSD}$ outperform experiment #2 with $\mathcal{L}_{CE}^{mix}$ by +0.6 in lip translation performance on En-Es. This demonstrates that $\mathcal{L}_{JSD}$ is the main contributor to achieving cross-modality knowledge transfer by building a bridge between the two speeches and performing fine-grained regularization across the probability of each word.

Maintaining knowledge of audio speech . It is also important to note that during the regularization process, the representation of audio speech is also affected by visual speech, which can interfere with the knowledge of audio speech and ultimately harm the lip translation performance of MixSpeech. As evidenced by experiment #2, the lip translation performance on En-Es decrease by -0.4 compared to experiment #3 when $\mathcal{L}_{CE}^{mix}$ is not applied. To address this issue, $\mathcal{L}_{CE}^{mix}$ is introduced to enhance the training ceiling of the cross-modality self-learning framework. By maintaining the translation performance of mixed speech and preventing the excessive disturbance to audio speech knowledge, $\mathcal{L}_{CE}^{mix}$ helps to improve the overall performance of MixSpeech.

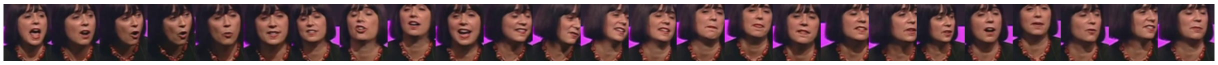
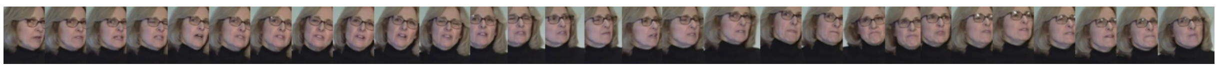
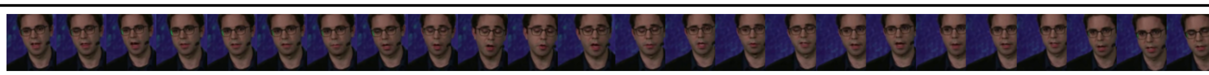
En-Es			
	En	Transcription:	and always as a child I had this fantasy that somebody would come and rescue me
		MixSpeech:	and always as a child I had this fantasy that somebody would come and rescue me
	Es	Ground Truth:	y de niña siempre tenía la fantasía de que alguien vendría a salvar me de
		MixSpeech:	y de (desde) niña ~~niño~~ siempre tenía la esta fantasía de que alguien vendría a salvar me (rescató) de
En-Fr			
	En	Transcription:	and solutions create new problems which have to be solved in their turn
		MixSpeech:	and solutions to create new problems which have to be solved in their turn ~~this year~~
	Fr	Ground Truth:	et les solutions créent de nouveaux problèmes devant être résolus à leur tour
		MixSpeech:	et les (des) solutions ~~pour~~ créer de nouveaux problèmes devant ~~qui doivent~~ être résolus à leur tour
En-It			
	En	Transcription:	one of the last 10,000 years, and the other certainly of the last 25 years
		MixSpeech:	one of the last 10,000 years, and the other certainly ~~assistance~~ of the last 25 years
	It	Ground Truth:	uno presente negli ultimi diecimila anni e l'altro certamente negli ultimi 25 anni
		MixSpeech:	uno presente negli ultimi diecimila (10000) anni e l'altro certamente negli ultimi 25 anni
En-Pt			
	En	Transcription:	has around 2,000 people descending on MIT's campus
		MixSpeech:	has around 2,000 people descending on MIT's campus
	Pt	Ground Truth:	congregam cerca de 2000 pessoas no campus do mit
		MixSpeech:	congregam ~~de~~ cerca de 2000 pessoas ~~a estudar~~ no campus do mit

Table 5. Qualitative performance of Visual Speech Recognition and Translation on AVMuST-TED. ~~Red Strikeout Words~~: mistranslated words with opposite meaning, (Blue Words in parentheses): mistranslated words with similar meaning, Gray Words: the absent words.

4.7. Qualitative results

We present several examples of lip translation in Table 5 to qualitatively evaluate the translation quality of MixSpeech. The translation results are very close to the ground truth, and the semantics are consistent. We observe two types of words that differ in translation: synonyms and context-sensitive translations. Synonyms that have different spellings but the same meaning, such as `salvar` and `rescató` in Spanish, both meaning 'rescue', and `diecimila` and `10000` in Italian, both meaning 'ten thousand', are commonly found in translation tasks and can affect translation consistency. Additionally, there are translations that require context information, such as when the speaker refers to themselves as a `child`, and the translation in Spanish needs to take into account the speaker's gender to choose between `niña` for girl' or `niño` for boy' and 'child'. The qualitative translation results of MixSpeech demonstrate its capability to achieve reliable cross-lingua lip translation. In Appendix C.2, we also provide translation results of noisy audio speech translation with visual speech translation and audio speech translation with audio-visual speech translation to highlight the importance of visual speech in speech translation.

5. Conclusion

With the advancement of online technologies, such as online healthcare and sales, language barriers often prevent these tools from reaching and benefiting disadvantaged areas. In light of this, we focus on visual speech, a branch of the speech stream, and aim to translate visual speech from source languages to other target languages for cross-linguistic communication, specifically through lip translation. We meticulously curate the AVMuST-TED dataset, consisting of 706 hours of speech clips with professional translations from TED, to facilitate cross-linguistic research on visual speech. We also introduce MixSpeech, a cross-modality self-learning framework that utilizes mixed speech to regularize visual speech translation and achieves state-of-the-art performance in lip translation on AVMuST-TED and lip reading on LRS2, LRS3, and CMLR datasets.

Moreover, our work on visual speech and AVMuST-TED lay a solid foundation for further research on visual speech in cross-lingual fields. There are numerous related tasks with great potential for practical applications, such as Cross-Lingual Talking Head Generation [44]. These tasks hold immense promise for breaking down language barriers and promoting communication across diverse communities.

Acknowledgements

This work was supported in part by 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

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 1, 2, 5, 6, 12, 13

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 1, 2, 5, 12, 13

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 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, 2020. 6

[4] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016. 1

[5] Kartik Audhkhasi, Andrew Rosenberg, Abhinav Sethy, Bhuvana Ramabhadran, and Brian Kingsbury. End-to-end asr-free keyword search from speech. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1351–1359, 2017. 2

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 2

[7] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 2, 3, 4

[8] Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*, 2016. 1

[9] Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990. 1

[10] Jiaao Chen, Zichao Yang, and Diyi Yang. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *arXiv preprint arXiv:2004.12239*, 2020. 3

[11] Xize Cheng, Tao Jin, Linjun Li, Wang Lin, Xinyu Duan, and Zhou Zhao. Opensr: Open-modality speech recognition via maintaining multi-modality alignment. *arXiv preprint arXiv:2306.06410*, 2023. 2

[12] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 13

[13] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017. 2, 12

[14] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006. 2

[15] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022. 1

[16] Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Must-c: a multilingual speech translation corpus. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2012–2017. Association for Computational Linguistics, 2019. 5

[17] Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, Qibing Bai, and Yu Zhang. Leveraging pseudo-labeled data to improve direct speech-to-speech translation. *arXiv preprint arXiv:2205.08993*, 2022. 1

[18] Qingkai Fang and Yang Feng. Neural machine translation with phrase-level universal visual representations. *arXiv preprint arXiv:2203.10299*, 2022. 3

[19] Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. Stemm: Self-learning with speech-text manifold mixup for speech translation. *arXiv preprint arXiv:2203.10426*, 2022. 1, 2, 3

[20] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3714–3722, 2019. 3

[21] Naomi Harte and Eoin Gillen. Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, 2015. 2

[22] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 2

[23] Wei-Ning Hsu and Bowen Shi. A single self-supervised model for many speech modalities enables zero-shot modality transfer. *arXiv preprint arXiv:2207.07036*, 2022. 1

[24] Chenxu Hu, Qiao Tian, Tingle Li, Wang Yuping, Yuxuan Wang, and Hang Zhao. Neural dubber: Dubbing for videos according to scripts. *Advances in Neural Information Processing Systems*, 34:16582–16595, 2021. 1, 3

[25] Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*, 2022. 2

[26] Rongjie Huang, Huadai Liu, Xize Cheng, Yi Ren, Linjun Li, Zhenhui Ye, Jinzheng He, Lichao Zhang, Jinglin Liu, Xiang

Yin, et al. Av-transpeech: Audio-visual robust speech-to-speech translation. *arXiv preprint arXiv:2305.15403*, 2023. 2

[27] 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*. 3

[28] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605, 2022. 2

[29] Rongjie Huang, Zhou Zhao, Jinglin Liu, Huadai Liu, Yi Ren, Lichao Zhang, and Jinzheng He. Transpeech: Speech-to-speech translation with bilateral perturbation. *arXiv preprint arXiv:2205.12523*, 2022. 1

[30] Taejun Kim and Juhan Nam. Temporal feedback convolutional recurrent neural networks for keyword spotting. *arXiv preprint arXiv:1911.01803*, 2019. 2

[31] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 12

[32] 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. 1

[33] Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, et al. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*, 2021. 1

[34] Linjun Li, Tao Jin, Xize Cheng, Ye Wang, Wang Lin, Rongjie Huang, and Zhou Zhao. Contrastive token-wise meta-learning for unseen performer visual temporal-aligned translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10993–11007, 2023. 2

[35] Rainer Lienhart. Reliable transition detection in videos: A survey and practitioner's guide. *International journal of image and graphics*, 1(03):469–486, 2001. 12

[36] Zhijie Lin, Zhou Zhao, Haoyuan Li, Jinglin Liu, Meng Zhang, Xingshan Zeng, and Xiaofei He. Simullr: Simultaneous lip reading transducer with attention-guided adaptive memory. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1359–1367, 2021. 1

[37] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. 1

[38] Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. End-to-end speech translation with knowledge distillation. *arXiv preprint arXiv:1904.08075*, 2019. 1

[39] Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W Schuller, and Maja Pantic. Lira: Learning visual speech representations from audio through self-supervision. *arXiv preprint arXiv:2106.09171*, 2021. 2, 12

[40] Shihui Ma, Shilin Wang, and Xiang Lin. A transformer-based model for sentence-level chinese mandarin lipreading. In *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*, pages 78–81. IEEE, 2020. 6

[41] ML Menéndez, JA Pardo, L Pardo, and MC Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997. 4

[42] Liliane Momeni, Triantafyllos Afouras, Themos Stafylakis, Samuel Albanie, and Andrew Zisserman. Seeing wake words: Audio-visual keyword spotting. *arXiv preprint arXiv:2009.01225*, 2020. 2

[43] David S Moore. Uncertainty. *On the shoulders of giants: New approaches to numeracy*, pages 95–137, 1990. 3, 4

[44] NVIDIA. Nvidia maxine: Reinventing real-time video communications with ai. 2022. 1, 2, 8

[45] 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. 2, 5

[46] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, Oct. 2018. Association for Computational Linguistics. 5

[47] KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman. Sub-word level lip reading with visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5162–5172, 2022. 2, 6

[48] 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. 1

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[50] Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, and Shengfeng He. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13325–13333, 2021. 2, 6

[51] 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

[52] Andrew Rosenberg, Kartik Audhkhasi, Abhinav Sethy, Bhuvana Ramabhadran, and Michael Picheny. End-to-end speech recognition and keyword search on low-resource languages. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 5280–5284. IEEE, 2017. 2

[53] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022. 1, 2, 3, 4, 5, 6, 12, 13

[54] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*, 2022. 1, 13

[55] David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *ArXiv*, abs/1510.08484, 2015. 6, 13

[56] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6447–6456, 2017. 6

[57] Hyoung-Kyu Song, Sang Hoon Woo, Junhyeok Lee, Seungmin Yang, Hyunjae Cho, Youseong Lee, Dongho Choi, and Kang-wook Kim. Talking face generation with multilingual tts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21425–21430, 2022. 1, 2

[58] Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. Neural lattice-to-sequence models for uncertain inputs. *arXiv preprint arXiv:1704.00559*, 2017. 1, 2

[59] Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. A general multi-task learning framework to leverage text data for speech to text tasks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6209–6213. IEEE, 2021. 3

[60] Alexander Waibel, Moritz Behr, Fevziye Irem Eyiokur, Dogucan Yaman, Tuan-Nam Nguyen, Carlos Mullov, Mehmet Arif Demirtas, Alperen Kantarcı, Stefan Constantin, and Hazım Kemal Ekenel. Face-dubbing++: Lip-synchronous, voice preserving translation of videos. *arXiv preprint arXiv:2206.04523*, 2022. 1, 2

[61] Chen Wang, Yuchen Liu, Boxing Chen, Jiajun Zhang, Wei Luo, Zhongqiang Huang, and Chengqing Zong. Discrete cross-modal alignment enables zero-shot speech translation. *arXiv preprint arXiv:2210.09556*, 2022. 1

[62] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*, 2021. 5

[63] Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, and Ping Luo. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. In *International Conference on Machine Learning*, pages 22680–22690. PMLR, 2022. 3

[64] Zehan Wang, Yang Zhao, Xize Cheng, Haifeng Huang, Jiageng Liu, Li Tang, Linjun Li, Yongqi Wang, Aoxiong Yin, Ziang Zhang, et al. Connecting multi-modal contrastive representations. *arXiv preprint arXiv:2305.14381*, 2023. 2

[65] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 2019. 12

[66] Rong Ye, Mingxuan Wang, and Lei Li. End-to-end speech translation via cross-modal progressive training. *arXiv preprint arXiv:2104.10380*, 2021. 2

[67] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3

[68] Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? *arXiv preprint arXiv:2010.04819*, 2020. 3

[69] Xingxuan Zhang, Feng Cheng, and Shilin Wang. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 713–722, 2019. 6

[70] Ya Zhao, Rui Xu, and Mingli Song. A cascade sequence-to-sequence model for chinese mandarin lip reading. In *Proceedings of the ACM Multimedia Asia*, pages 1–6. 2019. 5, 6

[71] Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song. Hearing lips: Improving lip reading by distilling speech recognizers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6917–6924, 2020. 2, 5, 6