XLAVS-R: Cross-Lingual Audio-Visual Speech Representation For Noise-Robust Speech Perception

Anonymous ACL submission

Abstract

In this paper, we present XLAVS-R, a cross-lingual audio-visual speech representation for noise-robust speech perception in over 100 languages. On the MuAViC benchmark, its largest model outperforms the previous state of the art by average 24.9% on speech recognition over 9 languages and average 34.5% on speech-to-text translation over 6 directions into English. XLAVS-R is based on a variant of AV-HuBERT, which requires only one training round and has better performance. Besides audio-visual speech data, we also leverage nearly half a million hours of audio-only speech data and scale model size up to 2B parameters.

1 Introduction

Audio-visual speech processing denotes a range of tasks that processes human speech based on both acoustic and visual signal (e.g., lip movement). It is an extension of traditional speech processing from unimodal to multimodal setting. The inclusion of visual modality brings a few benefits, notably noise robustness. For instance, speech recognition and translation models based only on acoustics can be sensitive to environmental noise. Even state-of-the-art speech recognizers (Radford et al., 2022) suffer from huge performance drop in a noisy environment, while an audio-visual system presents much higher resilience under such a setting (Anwar et al., 2023). Furthermore, generating speech based visual movement is also an essential component in autonomous avatars ().

Speech recognition and speech-to-text translation as two core speech perception tasks, and have witnessed rapid developments in the past two years. Although high benchmark performance are achieved in clean speech settings, it is shown that state-of-the-art models, such as Whisper (Radford et al., 2022) and SeamlessM4T (Communication, 2023b), suffer from significant performance degradation in noisy environments (Anwar et al., 2023; Communication, 2023a).

	Hours		Languages	
	A	AV	A	AV
Audio-Only Pre-Training Data				
AVFormer (Seo et al., 2023)	60K	0	$1^{\star}$	0
FAVA (May et al., 2023)	2.8K	0	$>100^{\dagger}$	0
Audio-Visual Pre-Training Data				
AV-HuBERT (Shi et al., 2022)	0	1.8K	0	$1^{\star}$
AV-data2vec (Lian et al., 2023)	0	1.8K	0	$1^{\star}$
AV2AV (Choi et al., 2023)	0	7K	0	$>100^{\dagger}$
Audio-Only & Audio-Visual Pre-Training Data				
u-HuBERT (Hsu and Shi, 2022)	0.5K	1.8K	$1^{\star}$	$1^{\star}$
VATLM (Zhu et al., 2023b)	4.3K	1.8K	$1^{\star}$	$1^{\star}$
XLAVS-R (this work)	436K	3.5K	128	$>100^{\dagger}$

Table 1: Pre-training data type, amount and language coverage in audio-visual speech perception models (A: audio, AV: audio+video). Our work, XLAVS-R efficiently makes better use of large-scale massively multilingual audio-only speech data than prior work. * English-only. † Estimated by a speech language identification model.

There has been many efforts on improving the noise-robustness of speech perception systems. For instance, prior work (Ng et al., 2023; Zhu et al., 2023a) adapts self-supervised speech representation learning (SSL) methods with noisy training data, which is usually simulated by adding noises to clean speech. A key limitation of the audio-only approach is its susceptibility to disruptions in challenging noisy settings (Shi et al., 2022), such as intensive babble noise and overlapped speech, which are common in real-life recordings. Adopting audio-visual approaches (Shi et al., 2021, 2022) naturally mitigates these issues. However, prior works either were English-only or used only limited amount of audio-visual data and did not leverage large-scale audio-only data.

In this paper, we present XLAVS-R, a cross-lingual audio-visual speech representation for noise-robust speech perception in over 100 languages. It achieves state-of-the-art performance on the MuAViC benchmark (Anwar et al., 2023) for speech recognition in 9 languages and speech-to-text translation in 6 languages. The closest work to ours is FAVA (May et al., 2023), which requires labeled audio-visual speech data for supervised fine-tuning after the audio-only self-supervised pre-training. In contrast, our approach does not require labels for audio-visual speech data. Table 1 provides the comparison of our model with other audio-visual speech perception models on pre-training data type, amount and language coverage.

2 Related Work

Self-supervised audio-only speech representation. Self-supervised learning (SSL) for speech aims to model a general speech representation that can be used for various downstream applications such as speech recognition, spoken language understanding tasks (). Among the first SSL models, wav2vec (Schneider et al., 2019) learns the representation through contrastive predictive coding (), which is to maximize the similarity between anchor and positive samples while minimizing the similarity between anchor and negative samples. The speech utterance is encoded with two fully-convolutional neural networks, where the samples and anchor are respectively drawn from. wav2vec 2.0 (Baevski et al., 2020) combines the contrastive approach with span masking, where contextualization is built upon the quantized masked features, and further scales up the pre-training data to 60K hours of unlabeled speech. Masked prediction is a common technique used in the pretext task of other speech SSL approaches. HuBERT (Hsu et al., 2021), another popular SSL framework, takes the masked speech feature and predicts hidden units, which are generated by applying k-means clustering on MFCC and iteratively refined through the layerwise features. BEST-RQ (Chiu et al., 2022) adopts the same BERT-style () objective on units from a quantization module learned in an end-to-end way. Instead of using discrete units, Data2vec (Baevski et al., 2022) directly regresses the dense features from an exponential moving average (EMA) teacher. SSL models greatly reduced the need for labeled speech data in various speech tasks (). Notably in speech recognition, it matches or even surpasses fully supervised models on LibriSpeech () with much fewer transcriptions.

Compared to supervised models, SSL approaches are less label-intensive, thus being easy to extend to low-resource languages. There has been sustaining efforts building multilingual SSL models, many of which are based on the wav2vec 2.0 (Baevski et al., 2020). Popular frameworks include XLSR-53 (Conneau et al., 2021), XLS-R (Babu et al., 2022) and MMS (Pratap et al., 2023), which extends language coverage to 53, 128 and over 1000, respectively.

On the axis of data, scaling training data shows additionally significant gains to SSL methods. USM (Zhang et al., 2023b) pre-trains a 2B conformed speech encoder with Best-RQ (Chiu et al., 2022) objective using 12M hours of speech data, and shows strong performance on speech recognition and translation for multiple datasets in multiple domains. SeamlessM4T v2 (Communication, 2023a) leverages 4.5M hours of speech data to train a w2v-BERT 2.0 (Communication, 2023b) encoder, leading to coverage of over 143 languages.

Separately, there has been also works (Zhu et al., 2023a; Ng et al., 2023) for improving noise robustness of SSL models. Robust data2vec (Zhu et al., 2023a) trains a data2vec model (Baevski et al., 2022) with noise augmentation. DeHuBERT (Ng et al., 2023) introduces auxiliary losses to HuBERT (Hsu et al., 2021) driving correlation matrices between pairwise noise-distorted embeddings towards identity.

Self-supervised audio-visual speech representation. Audio-visual SSL approaches are heavily inspired from the audio-only counterpart. AV-HuBERT (Shi et al., 2021) takes the masked audio-visual stream as input and predicts the hidden units initialized with MFCC clusters and gradually refined with layerwise features, which extends HuBERT (Hsu et al., 2021) to audio-visual setting. The framework has shown to be effective for multiple downstream tasks, including lip reading (Shi et al., 2021), audio-visual speech recognition and translation (Shi et al., 2022; Anwar et al., 2023). In (Choi et al., 2023), it has been extended to a multilingual setting. RAVEN () leverage modality-specific EMA teachers to produce targets for masked prediction to avoid iterative refinement process. Similarly under the hood of single-iteration pertaining, AV-data2vec (Lian et al., 2023) is based on data2vec (Baevski et al., 2022) and regresses multimodal feature with an audio-visual EMA teacher. AV2vec (Zhang et al., 2023a) further combines av-data2vec with the masked prediction objective in AV-HuBERT.

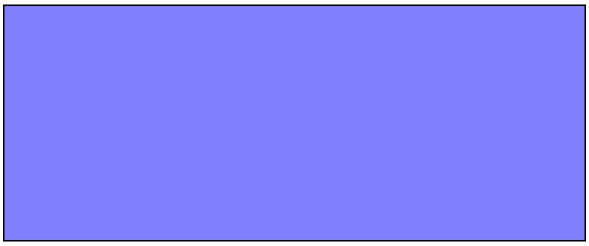

Figure 1: Overview of XLAVS-R. XLAVS-R is based on an variant of AV-HuBERT (AV-HuBERT v2), which can be trained in only one round and has better performance. We leverage an audio-only self-supervised speech model (XLS-R) to incorporate large-scale audio-only data and improve training efficiency.

Typically audio-visual SSL models are built upon a synchronized audio-visual stream. Prior works also explored using unpaired unimodal data to enhance the multimodal representation. u-HuBERT (Hsu and Shi, 2022) augments AV-HuBERT with audio-only speech in pre-training, which has shown to boost audio-visual speech recognition. VATLM (Zhu et al., 2023b) further added text-only data and trained the model with masked prediction on an arbitrary modality stream.

Audio-visual adaptation of audio-only speech models Recent work starts to explore adaptation of audio-only speech models into audio-visual models. MixSpeech (Cheng et al., 2023) builds a visual speech translation model based on a pre-trained audio-only speech translation model by minimizing the discrepancy between the probability from audio-only and multi-modal stream. AV-Former (Seo et al., 2023) adds lightweight modules to an audio-only speech recognizer to adapt it into visually grounded speech recognition through two-stage fine-tuning. FAVA (May et al., 2023) directly finetunes a pre-trained BEST-RQ () encoder for audio-visual speech recognition with a randomly initialized visual encoder. [BS: References to be added.]

3 Methods

Figure 1 provides an overview of XLAVS-R, which is based on a variant of AV-HuBERT (AV-HuBERT v2). We leverage an audio-only self-supervised speech model (XLS-R (Babu et al., 2022)) to incorporate large-scale audio-only training data and improve training efficiency.

3.1 AV-HuBERT v2

3.1.1 Single-Round Training With Self-Supervised Audio-Only Targets

AV-HuBERT (Shi et al., 2021) requires multiple training rounds, with training targets switching from quantized audio-only local features to quantized audio-visual contextualized representation that is obtained from earlier rounds of training. In each round, model is re-trained from scratch. Hence, model quality gains across training rounds all come from the updated training targets, which encode more and more contextual and bi-modal information in the later stage.

To improve training efficiency, we propose to create first-round training targets from a contextualized representation instead of local features that have noisier, lower-level information. This accelerates the masked prediction learning of high-level semantic information in the first round and reduces the necessity of additional training rounds. We obtain audio-only contextualized representation from a self-supervised multilingual speech model, XLS-R (Babu et al., 2022).

3.1.2 Learned Audio Feature Extractor

AV-HuBERT v2 has the same model architecture as AV-HuBERT except the audio feature extractor. Instead of using 26-dimensional filterbank features as audio inputs to the Transformer encoder, we jointly train a convolutional audio feature extractor (AFE) as that in wav2vec 2.0 (Baevski et al., 2020) for 512-dimensional audio inputs. This provides more capacity for multilingual models where there are cross-lingual inferences, and helps capture cleaner phonetic information for simple fusion with visual features through a linear projection.

3.2 XLAVS-R

3.3 Audio-Only Speech Representation Learning

Audio-only speech data usually has high availability than the audio-visual one, especially for low-resource languages. Moreover, audio-visual speech models usually run slower than the audio-only one of similar architecture and size because of the addition of visual feature extraction and bi-modal feature fusion. Instead of training an audio-visual model in one step with the mix of audio-only data and audio-visual data, we first train an audio-only speech model and adapt it to an audio-visual one via self-supervised parameter-efficient fine-tuning

Model	Mode	OOD	In-Domain									
		Avg	En	Ar	De	El	Es	Fr	It	Pt	Ru	Avg
						Clean environment, Test WER ↓						
Whisper V2 Large†	A	12.4	31.3	81.3	33.2	25.3	21.6	23.6	23.5	23.3	35.6	33.2
AV-HuBERT‡	A	41.6	3.2	88.1	53.6	46.7	18.0	20.4	21.2	22.5	44.8	35.4
(MuAViC-En + VC2-En)	AV	-	1.8	87.8	51.6	45.7	17.1	19.8	20.8	21.7	42.8	34.3
AV-HuBERT v2	A	34.2	3.7	83.3	48.7	26.5	13.0	15.8	14.8	16.0	35.2	28.6
(MuAViC)	AV	-	3.0	81.5	46.6	25.0	12.1	15.3	14.3	15.1	32.8	27.3
XLAVS-R 300M	A	29.2	3.8	81.5	40.0	21.2	11.2	13.6	12.6	14.0	30.7	25.4
(MuAViC)	AV	-	2.4	81.0	39.1	20.9	10.8	13.6	12.4	13.8	29.3	24.8
XLAVS-R 2B	A	22.4	6.9	82.9	33.4	16.5	9.6	11.5	10.7	11.8	25.8	23.2
(MuAViC)	AV	-	2.1	79.3	32.7	15.9	9.2	11.3	10.4	11.2	24.8	21.9
						Noisy environment, Test WER ↓						
Whisper V2 Large†	A	59.8	47.4	105.2	66.6	65.0	62.2	52.0	64.6	73.6	57.3	66.0
AV-HuBERT‡	A	91.0	63.9	105.9	87.7	84.4	65.1	59.5	70.5	72.1	75.8	76.1
(MuAViC-En + VC2-En)	AV	-	6.4	100.4	73.0	68.9	41.7	42.1	47.4	46.5	65.6	54.7
AV-HuBERT v2	A	90.8	113.6	104.8	77.0	68.7	49.7	42.5	52.7	56.7	60.8	69.6
(MuAViC)	AV	-	7.7	96.2	64.7	49.3	32.9	30.6	34.0	35.6	50.4	44.6
XLAVS-R 300M	A	92.0	114.8	103.7	73.8	61.7	48.8	39.4	50.2	55.4	57.8	67.3
(MuAViC)	AV	-	8.6	97.9	61.9	47.6	34.5	30.9	36.2	37.5	51.5	45.2
XLAVS-R 2B	A	79.0	84.6	109.2	70.1	57.0	43.7	36.5	46.9	51.8	55.8	61.7
(MuAViC)	AV	-	7.6	100.4	54.8	40.3	29.4	27.1	31.5	32.7	45.8	41.1

Table 2: In-domain (on MuAViC) and out-of-domain (OOD, on FLEURS) evaluation for multilingual audio-visual speech recognition models (A: audio, AV: audio+video). †Radford et al. (2022). Trained with 567 times of labeled audio-only data than MuAViC. ‡Shi et al. (2022).

with audio-visual data. We dedicate most of the training computation budget to the first stage since audio-only data is of much larger scale and the audio modality usually contains richer semantic information than the visual modality. We adopt wav2vec 2.0 (Baevski et al., 2020) as the audio-only model architecture, which can be viewed as a sub-model of AV-HuBERT v2.

3.4 Visual Modality Injection and Modality Dropout

After audio-only self-supervised learning with wav2vec 2.0, we add ResNet-based visual feature extractor (VFE) and a linear projection based feature fusion module to build up an AV-HuBERT v2 model. As in AV-HuBERT, we apply modality dropout to encourage the fusion of audio and visual representation space. With the masked prediction objective on audio-only targets, the second phase of model training aligns audio-visual and visual-only representations to corresponding audio-only representation, which is established in the first training phase.

3.5 Noise Injection

Besides the visual-to-audio representation alignment, audio-only and audio-visual representation for noisy speech is also aligned in the second training phase to the audio-only representation for its corresponding clean speech. This is implemented by randomly adding noises sampled from a noise dataset to clean speech audio and using the potentially noised audio as model inputs.

3.6 Parameter-Efficient Fine-Tuning

In the second training phase, visual inputs and audio noises are injected into model, leading to the need for adaptation of the contextualized encoder for new forms of inputs. We insert lightweight bottleneck adaptors (Bapna et al., 2019) before every Transformer layers in the contextualized encoder, and we fine-tune only the following modules: adaptors, visual feature extractor and feature fusion layer. The amount of trainable model parameters in the second training phase is 5% and 1% for XLAVS-R 300M and 2B models, respectively. Besides the cross-modal and de-noising alignments, the adaptors also provide additional capacity for further learning of semantic information in the au-

Model	Mode	OOD	In-Domain						
		Avg	El	Es	Fr	It	Pt	Ru	Avg
		Clean environment, Test BLEU ↑							
Whisper V2 Large†	A	xx	24.2	28.9	34.5	29.2	32.6	16.1	29.9
AV-HuBERT‡	A	12.7	13.9	22.3	28.1	23.5	26.1	10.7	20.8
(MuAViC-En + VC2-En)	AV	-	14.3	22.9	28.3	23.9	26.5	11.2	21.2
AV-HuBERT v2	A	13.0	17.6	23.0	28.6	23.5	27.6	11.7	22.0
(MuAViC)	AV	-	17.4	23.3	28.9	23.8	28.1	12.2	22.3
XLAVS-R 300M	A	14.5	19.6	24.1	29.7	24.3	29.0	12.3	23.2
(MuAViC)	AV	-	19.7	24.3	29.6	24.7	29.2	12.6	23.3
XLAVS-R 2B	A	16.0	21.7	25.0	30.6	26.5	30.2	13.9	24.7
(MuAViC)	AV	-	21.6	25.1	30.6	26.6	29.9	13.9	24.6
		Noisy environment, Test BLEU ↑							
Whisper V2 Large†	A	xx	0.1	0.4	0.7	0.1	0.1	0.2	0.3
AV-HuBERT‡	A	3.2	4.4	9.1	13.1	8.3	8.8	4.8	8.1
(MuAViC-En + VC2-En)	AV	-	8.8	15.6	19.2	15.0	17.6	7.2	13.9
AV-HuBERT v2	A	4.4	9.3	12.9	18.5	13.6	14.1	7.6	12.7
(MuAViC)	AV	-	12.6	17.8	22.9	18.5	20.9	9.1	17.0
XLAVS-R 300M	A	5.6	10.4	14.0	19.8	14.9	15.1	7.7	13.6
(MuAViC)	AV	-	13.5	17.2	22.5	17.8	20.0	8.5	16.6
XLAVS-R 2B	A	6.4	11.0	15.1	20.9	16.2	16.0	9.0	14.7
(MuAViC)	AV	-	15.7	19.2	24.6	20.1	22.3	10.4	18.7

Table 3: In-domain (on MuAViC) and out-of-domain (OOD, on FLEURS) evaluation for multilingual audio-visual speech-to-text translation models (A: audio, AV: audio+video). †Radford et al. (2022). ‡Shi et al. (2022).

dio, similar to that in the first training phase (via the case of audio-only clean speech inputs).

4 Data

We combine MuAViC (Anwar et al., 2023) and VoxCeleb2 (Chung et al., 2018) for a total of 3.5K hours of audio-visual pre-training data in more than 100 languages[1]. We adopt XLS-R (Babu et al., 2022) for audio-only pre-training on 128 languages with the combination of VoxPopuli (Wang et al., 2021), MLS (Pratap et al., 2020), Common Voice (Ardila et al., 2020), Babel (Gales et al., 2014) and VoxLingua107 (Valk and Alumäe, 2020).

5 Experiments

We perform multilingual fine-tuning on XLAVS-R with MuAViC labeled data for audio-visual speech recognition (AVSR) and audio-visual speech-to-text translation (AVS2TT). We use MuAViC for in-domain evaluation and FLEURS (Conneau et al., 2023) for audio-only out-of-domain evaluation.

[1]Estimated by a speech language identification model

5.1 Experimental Setup

We build XLAVS-R models at two model sizes: 300M and 2B. The former has 24 encoder layers with a model dimension of 1024. The latter has 48 encoder layers with a model dimension of 1920. For AV-HuBERT v2 training targets, we extract audio-only speech representation from the 36th layer of XLS-R 2B (Babu et al., 2022) and quantize it with 2000 k-means clusters. We add bottleneck adaptors with an inner dimension of 32.

For fine-tuning to AVSR and AVS2TT models, we follow Anwar et al. (2023) to add a Transformer decoder that has 6 layers and a dimension of 512. We randomly augment 25% of the input samples with multiple types of additive noises with a SNR (signal-to-noise ratio) of 0. The noise audio clips in the categories of "natural", "music" and "babble" are sampled from MUSAN dataset (Snyder et al., 2015), while the overlapping "speech" noise samples are drawn from MuAViC English. In creating "speech" and "babble" noise sets, we ensure there are no speaker overlap among different partitions. We remove extremely short utterances (less than 0.2 seconds) and long utterances (more than 20 seconds) for better training stability.

For inference of AVSR and AVS2TT models,

Model	Mode	OOD	In-Domain									
		Avg	En	Ar	De	El	Es	Fr	It	Pt	Ru	Avg
Clean environment, Test WER ↓												
AV-HuBERT (MuAViC-En)[‡]	A	45.5	4.3	92.3	56.4	48.9	19.4	22.2	23.5	25.0	48.3	37.8
	AV	-	2.2	91.1	54.7	47.7	18.6	21.6	22.6	23.8	46.5	36.5
+ Single-round w/ SSL units	A	44.5	3.8	89.9	55.4	48.5	20.2	22.3	24.1	26.0	48.1	37.6
	AV	-	2.4	89.4	53.4	47.4	18.7	22.0	23.2	24.3	45.8	36.3
+ 8 non-English languages	A	45.2	6.1	88.6	55.8	47.8	18.9	23.6	23.2	24.7	47.0	37.3
	AV	-	3.5	87.3	53.7	46.2	17.6	23.4	22.4	23.4	44.9	35.8
+ Learned AFE	A	34.2	3.7	83.3	48.7	26.5	13.0	15.8	14.8	16.0	35.2	28.6
	AV	-	3.0	81.5	46.6	25.0	12.1	15.3	14.3	15.1	32.8	27.3
Noisy environment, Test WER ↓												
AV-HuBERT (MuAViC-En)[‡]	A	100.2	85.3	113.9	95.0	89.9	74.8	67.1	78.5	79.2	82.3	85.1
	AV	-	9.5	107.3	79.5	74.3	49.4	48.0	54.1	52.5	71.5	60.7
+ Single-round w/ SSL units	A	94.4	79.5	112.6	92.0	87.3	70.7	63.8	75.2	77.9	80.1	82.1
	AV	-	8.5	105.8	76.1	70.7	45.8	46.1	51.8	50.6	69.1	58.3
+ 8 non-English languages	A	84.3	90.7	109.3	94.0	89.2	72.2	67.1	76.3	78.4	81.2	84.3
	AV	-	13.5	102.8	76.5	71.3	46.1	46.8	52.0	51.1	69.1	58.8
+ Learned AFE	A	90.8	113.6	104.8	77.0	68.7	49.7	42.5	52.7	56.7	60.8	69.6
	AV	-	7.7	96.2	64.7	49.3	32.9	30.6	34.0	35.6	50.4	44.6

Table 4: Analysis for the effectiveness of changes in AV-HuBERT v2 on multilingual audio-visual speech recognition (A: audio, AV: audio+video). [†]Radford et al. (2022). [‡]Shi et al. (2022).

we use the best checkpoint by validation accuracy. We use a beam size of 5 and default values for the other beam search decoding hyperparameters. For AVSR, we apply Whisper text normalizer (Radford et al., 2022) before calculating WER (word error rate). For AVS2TT, we use SacreBLEU (Post, 2018) with default options, where texts are processed by its built-in *13a* tokenizer before BLEU (Papineni et al., 2002) calculation.

5.2 Multilingual Speech Recognition

Clean setup. We evaluate AVSR models in both audio-only ("A") and audio-visual ("AV") modes, where the former leverages only audio modality in fine-tuning and inference while the latter leverages both audio and visual modalities. As shown in the upper section of Table 2, XLAVS-R 300M outperforms the original AV-HuBERT (Shi et al., 2022), an English-only pre-trained model of similar size, by average 10% WER and average 9.5% WER respectively for audio-only and audio-visual modes. It outperforms the baseline by large margins on every non-English language. Our best model, XLAVS-R 2B outperforms the English-only baseline by average 12.2% WER and average 12.4% WER respectively for the two modes. Audio-only self-supervised speech pre-training ("XLAVS-R 300M" compared to "AV-HuBERT v2") brings an average 12.6% and 9.2% WER reduction for the two modes, respectively.

Noisy setup. The lower section of Table 2 shows the test WER of our AVSR models in a noisy setup, where we simulate noisy environments by adding multilingual babble noises to clean speech inputs with a SNR (signal-to-noise ratio) of 0. We observe that Whisper, a SOTA multilingual ASR model trained on 680K hours of labeled data and has 1.6B model size, suffers from performance degradation in this challenging setup, with an average WER of 66.0 over the 9 languages. In the audio-visual mode, the average WER for XLAVS-R 2B drop significantly by 33.4%, suggesting their efficient use of visual information to alleviate the distraction of noisy environments.

5.3 Multilingual X-En Speech-To-Text Translation

Clean setup. We report test BLEU for X-En AVS2TT models in Table 3. We see that XLAVS-R 300M outperforms the English-only AV-HuBERT by average 11.5% and 9.9% BLEU for audio-only and audio-visual modes. It outperforms the baseline by large margins on all directions. Our best model, XLAVS-R 2B outperforms the English-only baseline by average 18.8% and 16.0% BLEU respectively for the two modes. Audio-only self-supervised speech pre-training ("XLAVS-R 300M" compared to "AV-HuBERT v2") brings an average 5.4% and 4.4% BLEU gain for the two modes, respectively.

Noisy setup. We evaluate our X-En AVS2TT models in a noisy setup, whose test BLEU are shown in the lower section of Table 3. We simulate noisy environments in the same approach as that for AVSR models, where multilingual babble noises are added to clean speech inputs with a SNR of 0. We observe that Whisper, a SOTA multilingual X-En speech-to-text translation model, has a catastrophic performance under this setup, with only 0.3 average BLEU over the 6 directions. XLAVS-R 300M outperforms the English-only baseline of similar size ("AV-HuBERT") largely with 5.5 and 2.7 average BLEU improvement, respectively. In the audio-visual mode, the average BLEU for XLAVS-R 300M and XLAVS-R 2B increase significantly by 3.0 and 4.0 BLEU compared to the audio-only mode, showing their efficiently capturing the semantic information in the visual inputs.

5.4 Effectiveness of AV-HuBERT v2

In Table 4, we validate the effectiveness of the two main changes in AV-HuBERT v2 by ablation studies on the training setting of MuAViC English. We observe that single-round AV-HuBERT training with self-supervised contextualized audio-only units from XLS-R is slightly better than the original AV-HuBERT that has 5 training rounds. The introduction of learnable audio feature extractor greatly improves the model performance especially on the low-resource languages.

6 Conclusion

In this paper, we present XLAVS-R, a cross-lingual audio-visual speech representation for noise-robust speech perception in over 100 languages. On the MuAViC benchmark, its largest model outperforms the previous state of the art by average 24.9% on speech recognition over 9 languages and average 34.5% on speech-to-text translation over 6 directions into English. XLAVS-R is based on a variant of AV-HuBERT, which requires only one training round and has better performance. Besides audio-visual speech data, we also leverage nearly half a million hours of audio-only speech data and scale model size up to 2B parameters.

References

Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. 2023. Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR.

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.

Ankur Bapna, N. Arivazhagan, and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*.

Xize Cheng, Linjun Li, Tao Jin, Rongjie Huang, Wang Lin, Zehan Wang, Huangdai Liu, Ye Wang, Aoxiong Yin, and Zhou Zhao. 2023. Mixspeech: Cross-modality self-learning with audio-visual stream mixup for visual speech translation and recognition. *arXiv preprint arXiv:2303.05309*.

Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. 2022. Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning*, pages 3915–3924. PMLR.

Jeongsoo Choi, Se Jin Park, Minsu Kim, and Yong Man Ro. 2023. Av2av: Direct audio-visual speech to audio-visual speech translation with unified audio-visual speech representation.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. *Proc. Interspeech 2018*, pages 1086–1090.

Seamless Communication. 2023a. Seamless: Multilingual expressive and streaming speech translation.

Seamless Communication. 2023b. Seamlessm4t: Massively multilingual & multimodal machine translation.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pages 16–23. International Speech Communication Association (ISCA).

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.

Wei-Ning Hsu and Bowen Shi. 2022. u-hubert: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality. In *NeurIPS*.

Jiachen Lian, Alexei Baevski, Wei-Ning Hsu, and Michael Auli. 2023. Av-data2vec: Self-supervised learning of audio-visual speech representations with contextualized target representations. *arXiv preprint arXiv:2302.06419*.

Avner May, Dmitriy Serdyuk, Ankit Parag Shah, Otavio Braga, and Olivier Siohan. 2023. Audio-visual fine-tuning of audio-only asr models.

Dianwen Ng, Ruixiong Zhang, Jia Qi Yip, Zhao Yang, Jinjie Ni, Chong Zhang, Yukun Ma, Chongjia Ni, Eng Siong Chng, and Binchao Ma. 2023. De'hubert: Disentangling noise in a self-supervised model for robust speech recognition. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. Interspeech 2020*, pages 2757–2761.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. 2023. Avformer: Injecting vision into frozen speech models for zero-shot av-asr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22922–22931.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2021. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*.

Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. 2022. Robust Self-Supervised Audio-Visual Speech Recognition. In *Proc. Interspeech 2022*, pages 2118–2122.

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

Jörgen Valk and Tanel Alumäe. 2020. Voxlingua107: A dataset for spoken language recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Jing-Xuan Zhang, Genshun Wan, Zhen-Hua Ling, Jia Pan, Jianqing Gao, and Cong Liu. 2023a. Self-supervised audio-visual speech representations learning by multimodal self-distillation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023b. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

Qiu-Shi Zhu, Long Zhou, Jie Zhang, Shu-Jie Liu, Yu-Chen Hu, and Li-Rong Dai. 2023a. Robust data2vec: Noise-robust speech representation learning for asr by combining regression and improved contrastive learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. 2023b. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Transactions on Multimedia*.