XLAVS-R: Cross-Lingual Audio-Visual Speech Representation From Efficient Modality Injection

Anonymous ACL submission

Abstract

In this paper, we present an efficient multilingual noise-robust speech representation built in an efficient approach. We create XLAVS-R from XLS-R. We improve noise-robustness of Whisper model, a state-of-the-art model for speech recognition and speech-to-text translation 100 languages. On the MuAViC benchmark, it outperforms previous English-only pre-trained model by XXX%. After adaptation, the out-of-domain performance is kept on the FLEURS benchmark. We open source this model at XXX.

1 Introduction

AV-HuBERT (Shi et al., 2021, 2022).

AVFormer (Seo et al., 2023). (60K-hour English-only BEST-RQ, fine-tuned on LibriSpeech.)

MuAViC (Anwar et al., 2023). (LRS3 and VoxCeleb2-English pre-training.)

XLS-R (Babu et al., 2022).

Multilingual audio-visual speech pre-training. Leverage audio-only data efficiently via audio-only models.

Challenges:

- audio-visual speech data scarcity
- Computational cost. Prevents scaling.

Differene to AVFormer: finer-grained information that global semantic info.

[CW: Add an overview figure.]

2 Related Work

Audio-only speech representation learning. wav2vec and wav2vec 2 (Baevski et al., 2020). HuBERT. XLSR-53 (Conneau et al., 2021). XLS-R (Babu et al., 2022). MMS (Pratap et al., 2023).

Multimodal speech representation learning. AV-HuBERT (Shi et al., 2021). u-HuBERT (Hsu and Shi, 2022). VATLM (Zhu et al., 2023). AV-data2vec (Lian et al., 2023). AV2vec (Zhang et al., 2023).

Less related: VATT (Akbari et al., 2021). TriBERT (Rahman et al., 2021). CAV-MAE (Gong et al., 2022). XDC (Alwassel et al., 2020).

Visual modality injection into audio-only speech models. AVFormer (Seo et al., 2023) for visual grounding setting, which focused on the downstream task instead of pre-training. MixSpeech (Cheng et al., 2023) uses supervised speech-to-text translation tasks.

Audio-visual cross-modal speech alignment. Lip2Vec (Djilali et al., 2023). ADC-SSL (Sheng et al., 2021).

3 Methods

3.1 Audio-Only Speech Representation

Base model given the amount of data and the amount of information in audio-only speech.

Encoder-only model where there is an local feature extractor and a Transformer-based/Conformer-based trunk for contextualized representations.

Feature extractor can either be filterbank with lightweight downsampling module or convolutional feature extractor.

Unsupervised approach: wav2vec 2.0.

Supervised approach.

3.2 Visual-to-Audio Feature Alignment

Aligning visual feature space to audio feature space.

3.3 Noise-Reduced Audio Feature Learning

XLS-R features to replace MFCC features for unit extraction.

3.4 Visual Modality Injection Into Audio-Only Speech Representations

XLS-R model weights for pre-training.

2nd stage pre-training.

2-stage supervised fine-tuning.

4 Data

5 Experiments

We evaluate models on the following tasks: Audio-Visual Speech Recognition (AVSR), Audio-Visual Speech-to-Text Translation (AVS2TT), Audio-Visual Emotion Recognition (AVER).

5.1 Experimental Setup

AV-HuBERT Large. Unit targets from XLS-R.

Data: LRS3 (Afouras et al., 2018), Vox-Celeb2 (Chung et al., 2018), MuAViC (Anwar et al., 2023), AVSpeech (Ephrat et al., 2018). Totaling **XXX** hours for **XX** languages.

Evaluation: FLEURS (Conneau et al., 2023), MuAViC, CMLR (Zhao et al., 2019), CREMAD-D (Cao et al., 2014)

Adaptor location.

Modality dropout to encourage picking up visual info.

We show advantages in the noisy setting. VSR WER as a metric for the degree of visual injection.

5.2 Audio-Visual Speech Self-Supervised Learning

Results on speech recognition.

Results on speech-to-text translation.

Results on language identification.

Results on emotion recognition.

5.3 Cross-Lingual Transfer

5.4 Visual Modality Injection Into Speech-to-Text Models

6 Conclusion

References

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

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221.

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. 2020. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770.

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.

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, 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.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

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*.

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

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.

Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Haithem Boussaid, Ebtessam Almazrouei, and Merouane Debbah. 2023. Lip2vec: Efficient and robust visual speech recognition via latent-to-latent visual to audio representation mapping. *arXiv preprint arXiv:2308.06112*.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*.

Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R Glass. 2022. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*.

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*.

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*.

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*.

Tanzila Rahman, Mengyu Yang, and Leonid Sigal. 2021. Tribert: Human-centric audio-visual representation learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

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.

Changchong Sheng, Matti Pietikäinen, Qi Tian, and Li Liu. 2021. Cross-modal self-supervised learning for lip reading: When contrastive learning meets adversarial training. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2456–2464.

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.

Jing-Xuan Zhang, Genshun Wan, Zhen-Hua Ling, Jia Pan, Jianqing Gao, and Cong Liu. 2023. 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.

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

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

Model	Mode	FLEURS-9	MuAViC									
		Avg	En	Ar	De	El	Es	Fr	It	Pt	Ru	Avg
Clean environment, Test WER ↓												
mAV-HuBERT	A	39.2	5.3	85.1	48.2	45.1	17.1	19.5	21.0	22.4	44.4	34.2
	V	-	59.4	122.7	98.3	97.4	91.2	96.2	90.0	93.9	98.9	94.2
Modality Dropout=0.0 (800K upd)	AV	-	2.4	84.4	46.1	43.9	16.3	19.4	20.6	21.3	42.4	33.0
mAV-HuBERT	A	39.8	3.9	85.4	48.3	45.7	17.9	19.9	21.4	22.6	44.0	34.3
	V	-	56.9	107.2	98.5	126.4	87.7	91.5	86.6	88.3	98.1	93.5
Modality Dropout=0.5 (64k bsz, 800K upd)	AV	-	2.4	84.3	47.0	44.7	16.8	19.4	20.7	21.6	42.6	33.3
mAV-HuBERT	A	39.3	4.7	84.5	49.1	45.1	17.3	20.5	21.5	22.4	43.0	34.2
	V	-	57.6	107.8	104.6	97.0	87.3	93.1	86.4	89.1	98.1	91.2
+AVS, MD=0.5, (64k bsz, 800K upd)	AV	-	2.6	83.8	47.5	44.2	16.5	20.0	21.0	21.6	41.8	33.2
XLAVS-R	A	37.8	7.9	86.1	45.7	42.6	16.2	18.5	19.3	21.2	41.3	33.2
Modality Dropout=0.0	V	-	74.4	112.3	99.9	97.4	96.0	97.7	95.1	95.3	99.7	96.4
DN XLS-R, w/ adpt (400K+400K)	AV	-	2.5	84.9	43.6	41.7	15.4	18.1	18.7	20.3	40.0	31.7
Whisper V2 Large[†]	A	**xx.x**	3.1	91.5	24.8	25.4	12.0	12.7	13.0	15.5	31.1	25.5
	A	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**
V-injected Whisper V2 Large	V	-	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**
	AV	-	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**
XLS-R CTC	A	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**
V-injected XLS-R CTC	A	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**
	AV	-	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**
Noisy environment, Test WER ↓												
mAV-HuBERT	A	90.3	80.5	105.4	83.9	83.7	63.6	56.7	68.9	70.8	73.9	76.4
Modality Dropout=0.0 (800K upd)	AV	-	8.6	98.0	65.1	65.7	39.2	38.9	44.0	44.3	63.0	51.9
mAV-HuBERT	A	90.0	73.7	104.8	82.6	82.6	63.1	57.6	68.4	70.5	74.2	75.3
MD=0.5 (64k bsz, 800K upd)	AV	-	8.2	98.3	66.6	67.2	40.3	40.1	45.7	45.3	63.8	52.8
mAV-HuBERT	A	88.1	73.2	103.7	83.0	81.3	61.3	57.2	67.5	68.5	70.9	74.1
+AVS, MD=0.5 (64k bsz, 800K upd)	AV	-	8.5	97.9	67.3	66.5	39.9	40.7	45.7	44.9	61.9	52.6
V-injected DN XLS-R	A	91.0	86.2	105.7	83.1	82.0	63.2	56.5	68.0	70.8	73.8	76.6
w/ adpt 4200K+400K), MD=0.0	AV	-	9.1	99.0	66.1	65.5	40.3	39.7	44.7	45.3	62.6	52.5
Whisper V2 Large[†]	A	**xx.x**	202.4	197.9	244.4	113.3	116.3	172.3	172.4	223.6	126.2	174.3
V-injected Whisper V2 Large	A	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**
	AV	-	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**
XLS-R CTC	A	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**
V-injected XLS-R CTC	A	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**
	AV	-	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**	**xx.x**

Table 1: Results for multilingual speech recognition (A: audio, AV: audio+video). [†]Radford et al. (2022).

Table 2: Results for X-En speech-to-text translation (SNR=0. A: audio, AV: audio+video).

Model	Mode	MuAViC						Avg
		El	Es	Fr	It	Pt	Ru	
Clean environment, Test BLEU ↑								
Whisper V2 Large (Radford et al., 2022)	A	24.2	28.9	34.5	29.2	32.6	16.1	29.9
AV-HuBERT (Shi et al., 2022) [MA: It should be mAV-HuBERT & cite MuAViC]	A	9.3	21.0	26.3	21.2	24.3	9.3	18.6
	AV	7.6	20.5	25.2	20.0	24.0	8.1	17.6
mAV-HuBERT from XLS-R w/ adpt (400K+100K upd, 24K bsz)	A	12.8	20.2	25.7	21.0	24.6	9.0	18.9
	AV	12.8	20.5	25.8	21.1	24.7	9.3	19.0
Noisy environment, Test BLEU ↑								
Whisper V2 Large (Radford et al., 2022)	A	0.1	0.4	0.7	0.1	0.1	0.2	0.3
AV-HuBERT (Shi et al., 2022) [MA: mAV-HuBERT & cite MuAViC]	A	2.9	8.4	12.4	8.1	8.6	0.9	6.9
	AV	4.2	12.8	15.0	12.5	14.8	4.6	10.7

Table 3: Abalation of adaptation approaches on speech recognition. (A: audio, AV: audio+video)

mAV-HuBERT (with 32K total batch size)	Mode	En	Ar	De	El	Es	Fr	It	Pt	Ru	Avg
		Source/Target									
Clean environment, Test WER ↓											
English-only, From scratch, MD=0.5 (600K upd)	A	3.2	88.1	53.6	46.7	18.0	20.4	21.2	22.5	44.8	35.4
	AV	1.8	87.8	51.6	45.7	17.1	19.8	20.8	21.7	42.8	34.4
From scratch, MD=0.5 (400K upd)	A	3.9	86.3	50.2	46.4	18.5	21.6	23.0	24.4	45.6	35.5
	AV	2.7	85.5	48.6	45.1	17.6	21.0	22.0	23.2	43.7	34.4
From XLS-R, MD=0.5 (400K upd)	A	4.9	86.6	51.2	46.7	18.5	21.4	23.4	24.3	45.5	35.8
	AV	2.7	85.3	49.4	45.7	17.6	21.1	22.3	23.4	44.0	34.6
From XLS-R w/ adaptor(400K+400K frz+unfrz upd), MD=0	A	4.5	86.8	47.0	42.7	16.6	18.9	19.9	21.5	41.9	33.3
	AV	2.7	84.8	44.2	41.3	15.5	18.1	19.1	20.3	40.0	31.8
From DN (40k) XLS-R w/ adaptor (400K+400K upd), MD=0	A	7.9	86.1	45.7	42.6	16.2	18.5	19.3	21.2	41.3	33.2
	AV	2.5	84.9	43.6	41.7	15.4	18.1	18.7	20.3	40.0	31.7
From DN (40k upd,2k unit) XLS-R w/ adpt (400K+400K upd), MD=0	A	4.2	85.7	45.2	42.7	16.4	18.2	19.7	21.1	40.8	32.7
	AV	2.4	84.8	43.3	41.4	15.4	17.8	18.9	20.2	39.6	31.5
From DN (40k upd,2k) XLS-R w/ adpt (400K+400K upd), + AVS, MD=0	A	5.2	85.2	45.4	42.5	15.9	18.8	19.4	20.9	40.0	32.6
	AV	2.6	84.1	44.0	41.7	15.4	18.4	18.9	20.2	38.8	31.6
Noisy environment, Test WER ↓											
English-only, From scratch, MD=0.5 (600K upd)	A	63.9	105.9	87.7	84.4	65.1	59.5	70.5	72.1	75.8	76.1
	AV	6.4	100.4	73.0	68.9	41.7	42.1	47.4	46.5	65.6	54.7
From scratch, MD=0.5 (400K upd)	A	76.3	104.6	84.8	83.0	65.2	59.9	70.1	71.6	75.1	76.7
	AV	9.3	98.4	68.4	67.7	41.8	42.2	47.8	46.5	65.2	54.2
From XLS-R, MD=0.5 (400K upd)	A	77.9	105.8	85.8	83.9	65.8	60.1	70.7	73.4	77.1	77.8
	AV	9.0	99.5	69.7	68.2	41.5	42.0	48.2	47.5	66.3	54.6
From XLS-R w/ adaptor (400K+400K frz+unfrz upd), MD=0	A	90.4	111.5	89.9	85.2	67.9	59.2	70.3	73.7	76.5	80.5
	AV	9.3	100.7	67.8	65.0	40.8	40.4	45.5	45.1	63.1	53.1
From DN (40k) XLS-R w/ adaptor (400K+400K upd), MD=0	A	86.2	105.7	83.1	82.0	63.2	56.5	68.0	70.8	73.8	76.6
	AV	9.1	99.0	66.1	65.5	40.3	39.7	44.7	45.3	62.6	52.5
From DN (40k upd,2k unit) XLS-R w/ adpt (400K+400K upd), MD=0	A	80.3	108.0	81.2	81.3	62.5	55.3	68.1	69.0	72.4	75.4
	AV	9.2	100.8	66.3	65.3	40.2	39.8	44.7	44.7	62.6	52.6
From DN (40k upd,2k) XLS-R w/ adpt (400K+400K upd), + AVS, MD=0	A	89.2	109.0	87.0	84.7	68.6	60.9	70.8	75.7	75.3	80.1
	AV	9.8	99.8	68.6	66.5	41.4	41.9	45.3	45.6	61.6	53.4

Table 4: (Deprecated, **ToBeUpdated**)Abalation of AVSpeech data on speech recognition. (A: audio, AV: audio+video)

mAV-HuBERT (with 32K total batch size)	Mode	Source/Target									Avg
		En	Ar	De	El	Es	Fr	It	Pt	Ru	
Clean environment, Test WER ↓											
From scratch (400K upd)	A	4.1	70.2	49.9	47.4	21.2	24.3	26.0	27.2	46.5	35.2
	AV	3.7	70.0	48.2	46.1	20.2	23.6	25.2	26.0	45.3	34.2
From XLS-R w/ adaptor (400K frz upd)	A	3.8	69.2	48.3	41.9	17.5	20.0	19.9	20.2	41.8	31.4
	AV	3.3	69.0	47.3	41.1	16.8	19.7	19.6	19.2	41.4	30.9
From denoising (avs) XLS-R w/ adaptor (400K frz upd)	A	3.2	64.9	42.7	38.8	15.3	18.3	17.9	18.1	37.9	28.6
	AV	3.2	64.3	41.5	38.2	14.8	18.0	17.5	17.6	37.2	28.0
From denoising (avs) XLS-R w/ adaptor (400K+400k frz,unfrz upd)	A	3.1	67.8	43.5	42.4	17.0	19.5	19.7	20.8	39.4	30.6
	AV	3.0	67.3	42.2	41.6	16.1	19.1	19.4	20.0	38.6	29.7
From denoising (avs) XLS-R w/ adaptor (400K frz upd) (avs)	A										
	AV										
Noisy environment, Test WER ↓											
From scratch (400K upd)	A	52.6	96.3	82.5	83.3	70.4	63.8	73.1	75.3	75.7	74.8
	AV	13.8	85.7	69.5	70.0	51.1	49.5	56.1	54.0	67.8	57.5
From XLS-R w/ adaptor (400K frz upd)	A	50.3	93.1	83.0	81.3	69.0	60.4	69.8	71.9	71.4	72.2
	AV	14.4	84.7	72.4	70.4	52.7	48.1	56.4	52.7	65.7	57.5
From denoising (avs) XLS-R w/ adaptor (400K frz upd)	A	45.4	89.6	76.7	78.2	64.3	56.6	66.7	67.9	67.0	68.0
	AV	12.0	81.5	66.0	65.6	45.9	44.0	51.3	47.6	60.8	52.7
From denoising (avs) XLS-R w/ adaptor (400K+400k frz,unfrz upd)	A	43.4	89.1	78.5	77.7	62.9	55.1	65.5	66.7	67.9	67.4
	AV	10.1	80.7	64.8	65.6	43.9	43.2	48.7	46.3	60.0	51.5
From denoising (avs) XLS-R w/ adaptor (400K frz upd) (avs)	A										
	AV										

Table 5: (Deprecated) Abalation of batch size on speech recognition. (A: audio, AV: audio+video)

mAV-HuBERT	Mode	Source/Target									Avg
		En	Ar	De	El	Es	Fr	It	Pt	Ru	
Clean environment, Test WER ↓											
From XLS-R w/ adaptor (400K frz upd, 32K bsz)	A	3.9	68.9	48.5	40.9	16.8	19.7	19.7	20.1	41.4	31.1
	AV	3.6	68.5	47.0	39.9	16.3	19.2	19.5	19.4	40.9	30.5
From XLS-R w/ adaptor (400K+400K frz+unfrz upd, 32K bsz)	A	3.0	67.2	43.7	38.9	15.2	17.9	18.0	18.4	39.1	29.0
	AV	2.5	66.6	42.3	38.0	14.3	17.8	17.8	17.6	38.1	28.3
From XLS-R w/ adaptor (400K frz upd, 64K bsz)	A	3.6	69.9	48.5	41.2	16.8	19.9	20.2	19.9	41.7	31.3
	AV	3.1	68.7	46.9	40.3	16.2	19.5	19.5	19.4	41.2	30.5
From XLS-R w/ adaptor (600K frz upd, 64K bsz)	A	4.0	68.8	48.6	41.3	17.0	20.0	20.1	20.4	41.6	31.3
	AV	3.8	68.7	47.0	40.4	16.2	20.1	19.7	19.2	41.1	30.7
From XLS-R w/ adaptor (400K+400K frz+unfrz upd, 64K bsz)	A	3.3	66.3	44.0	38.7	15.2	17.7	18.1	17.9	39.3	29.0
	AV	3.1	65.5	42.3	37.8	14.2	17.4	17.6	17.0	38.3	28.1
Noisy environment, Test WER ↓											
From XLS-R w/ adaptor (400K frz upd)	A	46.8	92.5	83.7	82.0	69.1	60.7	70.6	70.5	71.1	71.9
	AV	14.2	84.5	73.1	69.3	51.5	48.3	55.5	51.5	65.3	57.0
From XLS-R w/ adaptor (400K+400K frz+unfrz, 32K bsz)	A	49.0	91.2	78.2	78.3	64.7	57.0	67.0	68.4	69.1	69.2
	AV	11.2	81.7	66.7	64.4	44.0	43.0	49.1	46.0	61.6	52.0
From XLS-R w/ adaptor (400K frz upd, 64K bsz)	A	50.5	99.0	83.1	80.4	68.7	60.2	71.3	69.9	70.1	72.6
	AV	13.9	89.0	73.1	69.3	50.6	47.9	56.8	51.3	65.6	57.5
From XLS-R w/ adaptor (600K frz upd, 64K bsz)	A	50.5	91.8	82.1	80.0	67.3	60.2	69.8	70.0	70.6	71.4
	AV	14.2	83.6	72.9	69.1	50.8	48.6	55.7	51.7	65.1	56.9
From XLS-R w/ adaptor (400K+400K frz+unfrz, 64K bsz)	A	45.2	90.8	78.4	78.4	64.2	56.9	66.1	67.2	68.2	68.4
	AV	10.5	81.9	65.8	63.6	44.2	42.5	48.5	45.6	61.6	51.6

Model	Mode	ASR			AVSR	
		FLEURS-82	MLS	VP	MuAViC	CMLR
Clean environment, Test WER ↓						
Whisper V2 Large†	A	xx.x	xx.x	xx.x	25.5	xx.x
V-injected Whisper V2 Large	A	xx.x	xx.x	xx.x	xx.x	xx.x
	V	-	-	-	xx.x	xx.x
	AV	-	-	-	xx.x	xx.x
SeamlessM4T Large	A	xx.x	xx.x	xx.x	xx.x	xx.x
V-injected SeamlessM4T Large	A	xx.x	xx.x	xx.x	xx.x	xx.x
	AV	-	-	-	xx.x	xx.x
Noisy environment, Test WER ↓						
Whisper V2 Large†	A	xx.x	xx.x	xx.x	174.3	xx.x
V-injected Whisper V2 Large	A	xx.x	xx.x	xx.x	xx.x	xx.x
	AV	-	-	-	xx.x	xx.x
SeamlessM4T Large	A	xx.x	xx.x	xx.x	xx.x	xx.x
V-injected SeamlessM4T Large	A	xx.x	xx.x	xx.x	xx.x	xx.x
	AV	-	-	-	xx.x	xx.x

Table 6: Results for out-of-domain multilingual speech recognition (A: audio, AV: audio+video). †Radford et al. (2022).