Supplementary Materials: Efficient Training for Multilingual Visual Speech Recognition: Pre-training with Discretized Visual Speech Representation

Anonymous Authors

1 VISUALIZATION OF SPEECH UNITS

Fig. 1 shows the visualization of all 1,000 units of both the audio speech units and the visual speech units. In visual speech units, more units are classified as vowels while audio speech units have more distinct phonemes. As we discussed before, the ambiguity of lip movements is reflected in the figure.

2 DATASET STATISTICS FOR EACH LANGUAGE

The dataset statistics for each language are shown in Table 1. Please note that there are only 181,034 non-English human-labeled videos (*i.e.*, mTEDx). Therefore, we increase the quantity of labeled data by utilizing the automatic labels proposed by [1, 2]. With this, we can construct 2,014,212 multilingual video-text paired data. The number of these automatic labels can be found in the 'Auto-labeled # of Video' column in the table.

Language	Dataset	Human-labeled # of Video	Auto-labeled # of Video	Hours	Total Hours
En	LRS2	142,157	-	223	3,481
	LRS3	150,498	-	433	
	VoxCeleb2	-	628,418	1,326	
	AVSpeech	-	837,044	1,499	
Es	mTEDx	44,532	-	72	384
	VoxCeleb2	-	22,682	42	
	AVSpeech	-	151,173	270	
It	mTEDx	26,018	-	46	152
	VoxCeleb2	-	19,261	38	
	AVSpeech	-	38,227	68	
Fr	mTEDx	58,426	-	85	331
	VoxCeleb2	-	66,943	124	
	AVSpeech	-	69,020	122	
Pt	mTEDx	52,058	-	82	420
	VoxCeleb2	-	4,843	9	
	AVSpeech	-	176,601	329	
De	VoxCeleb2	-	-	190	333
	AVSpeech	-	-	143	
Ru	VoxCeleb2	-	-	2	288
	AVSpeech	-	-	286	
Ar	VoxCeleb2	-	-	7	114
	AVSpeech	-	-	107	
El	VoxCeleb2	-	-	1	9
	AVSpeech	-	-	8	

Table 1: Data statistics of each language used in this work including automatic labels.

3 DETAILED TRAINING SETUP

We provide the detailed training setup used for experiments in Table 2. For pre-training mAV-HuBERT, we use a polynomial decay Learning Rate (LR) scheduler, batch size of 1,000 frames for each GPU, and train steps of 350k. For pre-training with the visual speech unit, we use 3,000 frames per GPU even though we can increase it to 6,000 frames. During finetuning, the pre-trained encoder is frozen for 10k steps and 7.2k steps for multilingual finetuning.

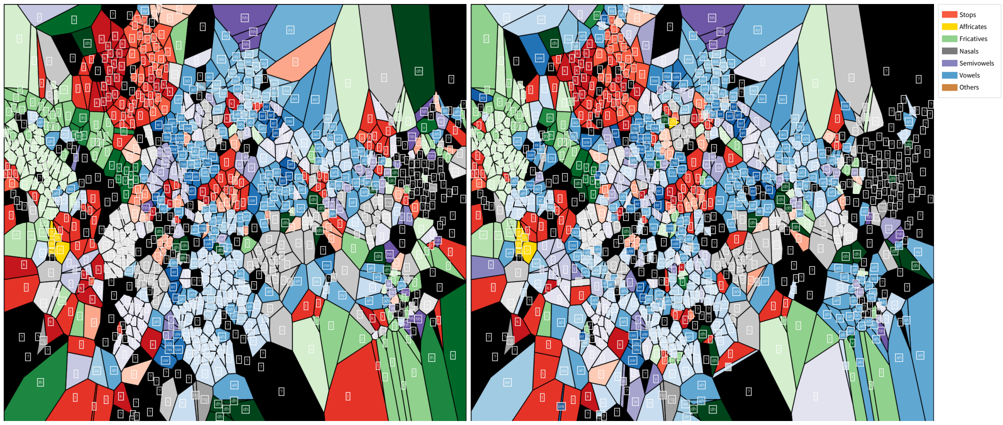
4 EXAMPLES OF PREDICTED SENTENCES

We show some examples of predicted transcriptions by the proposed multilingual VSR model and ground-truth transcriptions in Fig. 2. For each language, we show two examples. The red-colored words indicate the deletion error and the blue-colored words indicate the wrong prediction (*i.e.*, insertion or substitution)

REFERENCES

[1] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. 2023. Auto-AVSR: Audio-visual speech recognition with automatic labels. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[2] Jeong Hun Yeo, Minsu Kim, Shinji Watanabe, and Yong Man Ro. 2023. Visual Speech Recognition for Low-resource Languages with Automatic Labels From Whisper Model. *arXiv preprint arXiv:2309.08535* (2023).



(a) Audio Speech Unit

(b) Visual Speech Unit

Figure 1: Visualization of speech units. Each boundary represents a single unit and the same color represents the same phoneme or phoneme family. (a) Audio speech unit. (b) Visual speech unit.

	Pre-training (mAV-HuBERT)	Pre-training (Visual speech unit to text translation)	Fine-tuning (Multilingual)	Fine-tuning (Monolingual)
# of epochs	40	11	8	-
# of steps	350,000	60,000	120,000	60,000
# of frozen steps	-	-	10,000	7,200
# of GPUs	64	32	32	8
Max frames / batch	1000	3000	1000	1000
LR scheduler	polynomial decay	tri-stage	tri-stage	tri-stage
warmup updates	48,000	15,000	15,000	15,000
peak learning rate	2e-3	1e-3	4e-4	4e-4
Adam (β_1, β_2)	(0.9, 0.98)	(0.9, 0.98)	(0.9, 0.98)	(0.9, 0.98)

Table 2: Details of hyperparameters used in training.

English (En)	Ground Truth:	the choices don't make sense because it's the wrong question
	Prediction:	choices don't make sense because it's the wrong question
	Ground Truth:	this is not a statement on malnutrition or anything else
	Prediction:	this is not a statement on malnutrition or anything
Spanish (Es)	Ground Truth:	si os digo la verdad hasta hace poco no me había hecho esa pregunta
	Prediction:	yo sigo la verdad que hasta hace pocos me había hecho esa pregunta
	Ground Truth:	los papelitos oficiales están a la venta desde ahora corran que se acaban
	Prediction:	los preparatos oficiales están en la venta desde ahora corren que sacaban
Italian (It)	Ground Truth:	ed è molto diversa dalle precedenti per almeno cinque motivi
	Prediction:	era molto diversa dalle precedenti perché erano cinque motivi
	Ground Truth:	ora andiamo nella parrocchia di quartiere a corso francia
	Prediction:	ora andiamo nella persona di quel diritto su francia
French (Fr)	Ground Truth:	et je vais vous en citer trois mais il y en a énormément
	Prediction:	et je vais vous enregistrer trois milieux énorméments
	Ground Truth:	ça cest le maîtremot de lévolution et des espèces qui veulent survivre
	Prediction:	ça cest de mettre le bonheur de lévolution des espèces qui veulent survivre
Portuguese (Pt)	Ground Truth:	e na puberdade a gente usa muito o método científico
	Prediction:	e nesse momento a gente usa muito mais um científico
	Ground Truth:	se um dia tu pudesse conhecer alguma pessoa quem seria essa pessoa
	Prediction:	se o jeito pudesse conhecer alguma pessoa crescer nessa pessoa

Figure 2: Example sentences predicted from the single proposed multilingual VSR model on LRS3 and mTEDx test set. The Red and Blue indicate deletion and wrong predicted words, respectively.