Supplementary: Lip Reading for Low-resource Languages by Learning and Combining General Speech Knowledge and Language-specific Knowledge

1. Training Details

1.1. LRS2

Our proposed method outperforms the previous state-of-the-art methods and sets a new state-of-the-art performance in Table 2 (Manuscript). In this section, we provide further details for pre-training the proposed LMDecoder. In addition, we also give details for finetuning the entire lip reading model with the LMDecoder on LRS2 dataset.

1.1.1 Pre-training

We pre-train the LMDecoder to learn English-specific knowledge from 656 hours of audio-text paired data of LRS2 and LRS3. The LMDecoder consists of LM, transformer encoders, and transformer decoders. We set the memory bank size as 1,000 and use 4 layers for transformer encoders with a 1,024 embedding dimension, a 4,096 feed-forward dimension, and 8 attention heads. The configuration of the transformer decoders is the same as transformer encoders except for having 9 layers. All components of the LMDecoder are trained in an end-to-end manner with 60,000 steps. We use warmup steps of 15,000. The other training options such as learning rate are shown in Table 1. The tri-learning rate schedule in the table indicates (warmup, hold, decay) percentage for the total steps.

1.1.2 Finetuning

After the pre-training stage, we compose the lip reading pipeline by concatenating the pre-trained visual encoder and the LMDecoder. We employ a pre-trained AV-HuBERT Large model for the visual encoder. The entire lip reading model is finetuned for 30,000 steps. During finetuning, we freeze the visual encoder until 20,000 steps. Adam optimizer with a peak learning rate of 0.0005 and warmup steps of 10,000 is utilized for finetuning. Details are provided in the last column of Table 1.

	Pre-training	Fine-tuning
# steps	60,000	30,000
# frozen steps	-	20,000
tri-stage LR schedule	(25%, 0%, 75%)	(33%, 0%, 67%)
peak learning rate	1e-3	5e-4
# GPUs	8	8
Adam (β_1, β_2)	(0.9, 0.98)	(0.9, 0.98)

Table 1. Training details on LRS2 (EN).

	Pre-training	Fine-tuning
# steps	60,000	50,000
# frozen steps	-	-
tri-stage LR schedule	(25%, 0%, 75%)	(20%, 0%, 80%)
peak learning rate	1e-3	1e-3
# GPUs	8	4
Adam (β_1, β_2)	(0.9, 0.98)	(0.9, 0.98)

Table 2. Training details on mTEDx (IT, FR, ES, and PT).

1.2. mTEDx

For the low-resource languages, our goal is to learn language-specific knowledge on each target language by using audio-text paired data to supplement insufficient video-text paired data. Therefore, we jointly utilize mTEDx and MLS datasets to pre-train LMDecoder on the target language data. We note that the MLS dataset has more audio-text paired data than the mTEDx.

1.2.1 Pre-training

We pre-train the LMDecoder to learn language-specific knowledge from audio-text paired data of each target language (IT: 294h, FR: 1,163h, ES: 992h, and PT: 254h). The LMDecoder consists of LM, transformer encoders, and transformer decoders. We set the memory bank size as 1,000 and use 4 layers for transformer encoders with a 768 embedding dimension, a 3,072 feed-forward dimension, and 12 attention heads. The configuration of the transformer decoders is the same as transformer encoders except

for having 6 layers. All components of the LMDecoder are trained in an end-to-end manner with 60,000 steps. We use warmup steps of 15,000. The other training options such as learning rate are shown in Table 2.

1.2.2 Finetuning

After the pre-training stage, we compose the lip reading pipeline by concatenating the pre-trained visual encoder and the LMDecoder for each target language. We employ a pre-trained AV-HuBERT Base model for the visual encoder. The entire lip reading model is finetuned for 50,000 steps. In contrast to the experiment on LRS2, we do not freeze the visual encoder. Adam optimizer with a peak learning rate of 0.001 and warmup steps of 10,000 is utilized for finetuning. Details are provided in the last column of Table 2.

2. Utilizing Speech Knowledge of Large-scale Pre-trained English Lip Reading Model

Recently, large-scale pre-trained lip reading models using ASR-labeled English data have been proposed [38]. They utilized a pre-trained ASR model to label unlabeled English datasets and obtained 3,448 hours of visual-text data. In this section, we explore whether we can utilize these large-scale pre-trained English lip reading models' speech knowledge for low-resource lip reading. However, as the visual encoder of their pre-trained model is not trained with the speech unit prediction task, it is not matched well with the LMDecoder that is trained using speech unit inputs. Therefore, we find the performance degradation when directly cascading the pre-trained visual encoder of [38] and the LMDecoder. To handle this, we add a residual connection between the output of the visual encoder and the input of the decoder so that the imperfect memory addressing in Language-specific Memory (LM) can be complemented through the residual connection. With this simple modification, we applied the proposed method to combine the speech knowledge learned from large-scale English data and the language-specific knowledge learned from language-specific audio-text data. The results on mTEDx are shown in Table 3, 4, 5, and 6. Compared to using 814 hours of English data, by employing the knowledge learned from 3,448 hours of English data, we can largely improve the lip reading performances for low-resource languages (*i.e.*, IT, FR, ES, and PT). These results confirm that the speech knowledge learned from one language can be transferred to different languages. By combining speech knowledge with language-specific knowledge through the proposed method, we can further improve lip reading performances on low-resource language datasets. For example, we can improve about 3% WER more from that of [38] on the mTEDx-ES dataset.

Method	Labeled A-T Data	Labeled V-T Data	WER
CM-seq2seq [15]	-	47h (+814h)	78.31%
CM-seq2seq [38]	-	47h (+3448h)	60.40%
Proposed Method	294h	47h (+3448h)	**59.74%**

Table 3. Lip reading performance comparisons on mTEDx-IT. (+α) represents the amount of labeled English data.

Method	Labeled A-T Data	Labeled V-T Data	WER
CM-seq2seq [15]	-	86h (+814h)	88.41%
CM-seq2seq [38]	-	86h (+3448h)	65.25%
Proposed Method	1163h	86h (+3448h)	**64.92%**

Table 4. Lip reading performance comparisons on mTEDx-FR. (+α) represents the amount of labeled English data.

Method	Labeled A-T Data	Labeled V-T Data	WER
CM-seq2seq [15]	-	74h (+814h)	81.75%
CM-seq2seq [38]	-	74h (+3448h)	59.90%
Proposed Method	992h	74h (+3448h)	**56.96%**

Table 5. Lip reading performance comparisons on mTEDx-ES. (+α) represents the amount of labeled English data.

Method	Labeled A-T Data	Labeled V-T Data	WER
CM-seq2seq [15]	-	93h (+814h)	79.17%
CM-seq2seq [38]	-	93h (+3448h)	59.45%
Proposed Method	254h	93h (+3448h)	**58.57%**

Table 6. Lip reading performance comparisons on mTEDx-PT. (+α) represents the amount of labeled English data.