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

Minsu Kim* Jeong Hun Yeo* Jeongsoo Choi Yong Man Ro†
School of Electrical Engineering, KAIST, South Korea
{ms.k, sedne246, jeongsoo.choi, ymro}@kaist.ac.kr

Abstract

This paper proposes a novel lip reading framework, especially for low-resource languages, which has not been well addressed in the previous literature. Since low-resource languages do not have enough video-text paired data to train the model to have sufficient power to model lip movements and language, it is regarded as challenging to develop lip reading models for low-resource languages. In order to mitigate the challenge, we try to learn general speech knowledge, the ability to model lip movements, from a high-resource language through the prediction of speech units. It is known that different languages partially share common phonemes, thus general speech knowledge learned from one language can be extended to other languages. Then, we try to learn language-specific knowledge, the ability to model language, by proposing Language-specific Memory-augmented Decoder (LMDecoder). LMDecoder saves language-specific audio features into memory banks and can be trained on audio-text paired data which is more easily accessible than video-text paired data. Therefore, with LMDecoder, we can transform the input speech units into language-specific audio features and translate them into texts by utilizing the learned rich language knowledge. Finally, by combining general speech knowledge and language-specific knowledge, we can efficiently develop lip reading models even for low-resource languages. Through extensive experiments using five languages, English, Spanish, French, Italian, and Portuguese, the effectiveness of the proposed method is evaluated.

1. Introduction

It is a fascinating ability to understand the conversation by only looking at the speaker's lip movements without listening [1]. If this were possible, we could easily hold conversations in crowded places, such as at concerts, and even with people who have trouble speaking up. With the great advance of deep learning, a technology called lip reading has made it possible to accurately infer what a speaker is saying without having to approach the speaker's voice. In recent years, the performance of lip reading has significantly improved from 60.1% Word Error Rate (WER) to 26.9% WER [2, 3] in LRS3 [4], a popular English benchmark database.

Such rapid progress could be made with large-scale audio-visual datasets [4–8], improved neural network architecture [9–15], enhanced multi-modal learning strategies [3, 16–22], and carefully designed training methods [23–25]. Among these progresses, self-supervised learning methods using audio-visual data show remarkable advancement in both audio-visual speech recognition and lip reading. Recently, AV-HuBERT [3] which pre-trains the transformer encoder with multi-modal inputs (*i.e.*, audio and video) through masked prediction in a self-supervised manner, outperforms other previous lip reading methods once it is finetuned on lip reading tasks. Despite these advances, lip reading technologies have been developed primarily in English rather than in other languages. One main reason for this is the lack of enough labeled video-text paired data in other languages. For example, the popular lip reading dataset in English, LRS3 [4], consists of about 443 hours of video, while the available video-text paired dataset in Italian [26] is only about 47 hours, which is not enough for the model to learn the characteristics of both lip movements and language. Therefore, to build lip reading models for other languages rather than English, a new method considering the insufficient training data should be developed.

In this paper, we focus on developing a novel lip reading method for low-resource languages which has not been well explored in previous literature. Specifically, we propose a novel training method for low-resource language lip reading that learns 1) general speech knowledge and 2) language-specific knowledge, and combines the two learned knowledge. First, general speech knowledge refers to the knowledge of modelling short-term speech that can be regarded as speech units (*i.e.*, phonemes or visemes). Since different languages partially share common phonemes [27–29],

*Both authors have contributed equally to this work.
†Corresponding author

learning to model accurate speech units from high-resource language can be beneficial in modelling speech representations for low-resource language. To this end, we train the visual encoder to predict speech units from input lip movements through masked prediction using a high-resource language, English. Second, language-specific knowledge refers to the knowledge of translating learned speech representations into text, which can be regarded as the language modelling ability of a model. Since learning a language requires large-scale data [30–32], it might be insufficient to only utilize the video-text paired data of low-resource language. To mitigate the problem, we propose a Language-specific Memory-augmented Decoder (LMDecoder) which can be trained from audio-text paired data in the target language and be applied for lip reading. The input of LMDecoder is set to speech units derived from audio, and Language-specific Memory (LM) saves language-specific audio features into memory banks, which are for transforming speech units into language-specific speech representations. Finally, after learning the two knowledge, we cascade the two modules (*i.e.*, visual encoder and LMDecoder) and we can employ both the accurate lip movements modelling ability (*i.e.*, general speech knowledge) of the visual encoder and the rich language modelling ability (*i.e.*, language-specific knowledge) of LMDecoder, for low-resource language lip reading.

The effectiveness of the proposed method is evaluated with five languages, English (EN), Spanish (ES), French (FR), Italian (IT), and Portuguese (PT). Especially, English is utilized as a high-resource language so employed to learn general speech knowledge, and other languages are utilized as low-resource languages thus LMDecoder is trained on each low-resource language data. Through comprehensive experiments, we show the proposed method is effective in developing lip reading models not only for low-resource languages but also effective for the high-resource language, by achieving state-of-the-art performance on English data. The contributions of this paper can be summarized as:

- To the best of our knowledge, this is the first attempt to analyze the effectiveness of different pre-training methods, self-supervised pre-training of encoders, supervised pre-training in a high-resource language, and pre-training of decoders with audio-text data, in building low-resource lip reading model.
- We propose a novel method of learning and combining general speech knowledge and language-specific knowledge to effectively develop lip reading models for low-resource languages.
- We conduct comprehensive experiments with five languages, English, Spanish, French, Italian, and Portuguese, and we show the effectiveness of the proposed method in developing lip reading models for different nationalities, even with a small-scale dataset.

2. Related Work

2.1. Lip reading

Lip reading [33–38] aims to predict the speech content by watching talking face videos only. Along with the advancement of Deep Learning and speech processing technology, lip reading technology achieves significant development. Early work [5] proposed a lip reading model consisting of CNN to predict word from word-level English data. [10, 11] proposed an architecture composed of ResNet [39] and RNN [40,41] to improve the word-level lip reading performances. [13,42] proposed to use optical flow information with RGB information by encoding them with two-stream networks. [43] changed the RNN-based back-end architecture with temporal convolutions and achieved significant performance improvement in word-level lip reading. Besides the word-level lip reading, [9] proposed an end-to-end sentence-level lip reading model that utilizes Connectionist Temporal Classification (CTC) [44]. Sentence-level large-scale audio-visual datasets, LRS2 [7] and LRS3 [4], are proposed to boost lip reading research. By adopting transformer [45], powerful architecture for modelling sequence data, [12] significantly improved the sentence-level lip reading performances. Recently, transformer-variants architectures [15, 46–48] are shown promising lip reading and audio-visual speech recognition performances. There are other works that tried to enhance lip reading performances by focusing on developing training strategies. [17,21,24,49] employed knowledge distillation [50] to bring knowledge of the superior model into the student model. [19,20,22,51,52] proposed to use memory networks to use the auditory knowledge in lip reading without audio inputs. [53–55] handled the speaker-dependency issue and proposed speaker-adaptive lip reading models. Recently, pre-training neural networks using self-supervised training methods showed remarkable lip reading performances [3, 16, 18].

However, most of the previous research is focused on developing lip reading models in principal languages, such as English and Mandarin. Lip reading for different languages, especially low-resource languages, has not been well explored [25]. In this paper, we propose a new method for low-resource languages that contain a small-scale visual-text paired dataset. By learning and combining general speech knowledge and language-specific knowledge, the proposed method can effectively learn how to model the lips and the target language, even for the low-resource language.

2.2. Pre-training strategies

Recently, in diverse areas, the pre-trained model shows remarkable performances when they are applied to different downstream tasks [30, 56–62]. It is also shown remark-

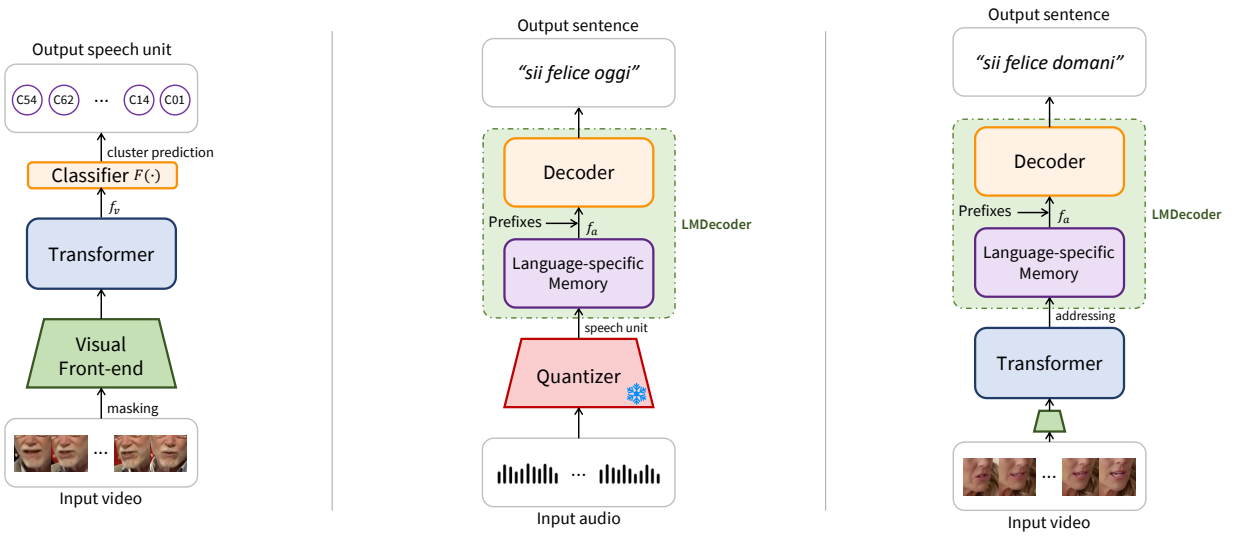

Figure 1. Overview of the proposed method for low-resource language lip reading. (a) Learning general speech representation by using masked prediction of speech units in a high-resource language. (b) The proposed Language-specific Memory-augmented Decoder (LMDecoder) learns language-specific knowledge from audio-text paired data by quantizing the input into speech units. (c) Lip reading models for low-resource languages can be built by combining general speech knowledge and language-specific knowledge.

able performances in the speech recognition area. In audio speech modelling, wav2vec2.0 [63] and HuBERT [64], proposed to learn the speech representations by predicting speech units obtained by clustering the acoustic features (*e.g.*, MFCC). They achieved state-of-the-art speech recognition performances by pre-training the model on large-scale unlabeled data. In visual speech modelling, [3, 18, 65] proposed self-supervised pretraining methods using audio-visual correspondences or masked prediction similar to audio pre-training methods. By finetuning the pre-trained model to the lip reading task, they achieved better lip reading performances than the trained model from the scratch.

In this paper, we also try to pre-train the visual encoder to learn general speech knowledge by predicting speech units from lip video using high-resource languages. Moreover, to learn language-specific knowledge which will be utilized to translate the captured speech units into words, we propose Language-specific Memory-augmented Decoder (LMDecoder) which can be pre-trained on audio-text paired data.

2.3. Vector quantization

Since discrete representations are natural to express many modalities, using discrete representations in Deep Learning shows great progresses in diverse areas, such as image generation [66–68] and speech processing [63, 64, 69–73]. Especially, by discretizing audio using vector quantization, we can obtain discriminative hidden units which are highly correlated with the acoustic units (*i.e.*, phoneme) [64]. We try to use the speech units obtained through vector quantization of input video and audio in learning general speech knowledge and language-specific knowledge.

3. Method

Our objective in this paper is to develop lip reading models for low-resource languages. Different from English, other languages (*e.g.*, Italian, French, Korean, Japanese, etc.) have smaller video-text paired data for developing lip reading networks. Therefore, lip reading research has been mainly focused on English. To mitigate the insufficient visual-text paired data of the low-resource language in building a lip reading model, we propose to learn 1) general speech knowledge from a high-resource language and 2) language-specific knowledge from audio-text paired data.

3.1. Learning general speech knowledge

It is known that different languages share some common phonemes [27–29, 73], which means that knowledge learned to model speech units from lip movements in one language can be extended to other languages. Therefore, to effectively learn to model the lip movements of low-resource language, we try to bring the knowledge of a pre-trained model that is trained to model the speech units from a high-resource language, English. Motivated by the recent success of learning speech representations by predicting speech units in a self-supervised manner [3, 63, 64, 74], we train the visual encoder with masked prediction to learn general speech knowledge.

Specifically, the contiguous α frames of input video x_v with T frames are masked out. Then the masked video $\tilde{x}_v$ is encoded through a visual front-end and a transformer to

produce visual features f_v. Then, the visual encoder (*i.e.*, visual front-end and transformer) is guided to predict the speech units of the masked region indicated by an indicator $M_t \in \{0, 1\}$, where the value 1 indicates t-th frame is masked while the value 0 indicates not. The target speech units $z_t \in \{1, \dots, C\}$ with C classes are obtained by quantizing Mel-frequency Cepstral Coefficient (MFCC) of the audio corresponding to the input video using a discrete latent variable model (*e.g.*, K-means), which will be iteratively improved by using the learned features instead of MFCC similar to [3,63,64]. The process of learning general speech knowledge can be written as follows,

$$\mathcal{L}_{GSK} = -\sum_{\{t|M_t=1\}} z_t \log(\hat{z}_t), \tag{1}$$

where $\hat{z}_t = \text{Softmax}(F(f_v^t))$ is the probability of the predicted speech unit using a classifier $F(\cdot)$. For the implementation, we follow the recent observation of [3] that using both audio and video inputs to learn the speech representations is better than utilizing the video inputs only, and we use the multi-modal inputs. By training the visual encoder with the masked prediction of speech units on the large-scale dataset, the visual encoder can embed lip video into discriminative speech representations, which will be extended to other languages. The process for learning general speech knowledge is illustrated in Fig.1a.

3.2. Learning language-specific knowledge

The final goal of lip reading is translating the captured lip movements into words, which implies that the ability of language modeling can largely affect the final performance. However, for the low-resource language, video-text paired data might be insufficient for the model to learn to construct language. To handle this, as shown in Fig.1b, we propose a Language-specific Memory-augmented Decoder (LMDecoder) which can learn language-specific knowledge from audio-text paired data usually richer than video-text data.

Specifically, LMDecoder includes Language-specific Memory (LM) that can save speech representations of the target language according to speech units. Therefore, after training, we can extract language-specific speech representations from LM by examining the input speech units. When input audio is given, it is quantized to speech units x_a having C classes, similar to the obtaining of z_t in Sec.3.1. By quantizing the input audio into speech units, the learned general speech knowledge can be naturally fit to be utilized in LMDecoder when the visual encoder and LMDecoder are combined. Then, LM converts the input speech units into language-specific audio features f_a by accessing memory banks $B \in \mathbb{R}^{C \times d}$ corresponding to speech units as follows,

$$f_a^t = B_i \quad \text{such that} \quad x_a^t = i, \tag{2}$$

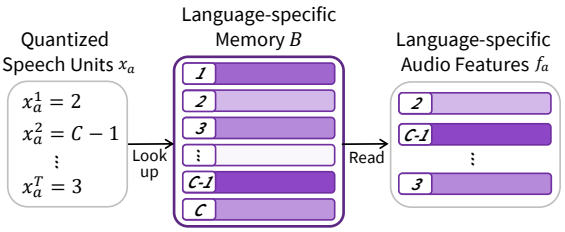

Figure 2. Illustration of Language-specific Memory (LM) and the memory banks B of LM. When a quantized speech unit is given, LM transforms it into a language-specific audio feature by reading the memory value. Therefore, the mapping of speech units to language-specific audio features can be constructed.

where d is the dimension of audio features. This procedure is illustrated in Fig.2 and it is similar to accessing the codebook in [67] and also to using auditory features in lip reading using the memory network of [19,20,22,52]. Then, a decoder translates the audio features into words in an autoregressive manner [75]. Let y be the ground-truth text tokens, then the process of learning language-specific knowledge of LMDecoder can be written as follows,

$$\mathcal{L}_{LSK} = -\log p(y|x_a) \tag{3}$$

where $p(y|x_a) = \Pi_{j=1}^{J} p(y_j|y_{<j}, x_a)$ and J represents the length of text tokens. As the learning of language-specific knowledge is purely available with audio-text paired data, the LMDecoder can learn to model the target language from large-scale data, even if the video-text paired dataset is small for the target language. Moreover, since the saved language-specific representations in LM are auditory features, we can bring the rich speech information of audio into lip reading, similar to [19,20,22,52].

3.3. Lip reading for low-resource language

After training the visual encoder to have general speech knowledge and the LMDecoder to have language-specific knowledge, we combine the two modules to compose the lip reading pipeline for low-resource language (Fig.1c). To access the saved language-specific audio features in LM, we employ scaled dot-product attention of [45] using the encoded visual features f_v. Through attention, the potential mismatches between speech units predicted from video and predicted from audio can be minimized when accessing the memory banks. Specifically, when visual features f_v are encoded by the visual encoder, language-specific audio features saved in LM (*i.e.*, B) are retrieved through an attention mechanism as follows,

$$\begin{aligned} Q^t = f_v^t W_q, \quad K = BW_k, \quad V = BW_v, \\ f_a^t = \text{Softmax}(\frac{Q^t K^\top}{\sqrt{d}})V, \end{aligned} \tag{4}$$

where W_q, W_k, and W_v are embedding matrices for query, key, and value, respectively. By using visual features as a query, we can access the memory banks B of LM to find and extract language-specific audio features related to the input lip movements. This also can be viewed as a soft attention [76] version of Eq. (2). Finally, with the extracted language-specific audio features, LMDecoder predicts text tokens of the target language, $\hat{y}$, in an auto-regressive manner by utilizing the learned language-specific knowledge.

4. Experimental Setup

4.1. Network architecture

Basically, the visual encoder has the same architecture as that of the AV-HuBERT Base [3] except for the LRS2 experiment which utilizes AV-HuBERT Large configuration. It is composed of a visual front-end and transformer encoders. Specifically, the visual front-end is comprised of ResNet18 [39] whose first stem layer is modified with 3D convolution [11]. The transformer has 12 encoder layers where each encoder has a 768 embedding dimension (*i.e.*, $d = 768$), a 3,072 feed-forward dimension, and 12 attention heads. The LMDecoder consists of Language-specific Memory (LM), transformer encoders, and transformer decoders. LM has memory banks with an embedding matrix size of $C \times d$, where C is set to 1,000. The transformer encoder has 4 layers to model the context from the extracted audio features f_a, with the same embedding size as the transformer in the visual encoder. The transformer decoder has 6 layers with a 768 embedding dimension, a 3,072 feed-forward dimension, and 4 attention heads, to predict the text tokens. To bridge the visual encoder and LMDecoder, we utilize scaled dot-product attention of [45] and the size of each embedding layer is set to the dimension of audio features (*i.e.*, $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$).

For obtaining the target speech units z, we use that of [3]; they are obtained by clustering MFCC at the first iteration, and then changed to the cluster of learned audio-visual representations through the iteration. For obtaining the speech units x_a from the input audio, we use the features encoded at the 11-th layer of a pre-trained HuBERT [64] that trained on VoxPopuli [77] and perform K-means clustering.

4.2. Implementation details

Our experiments are implemented using an open-source toolkit, fairseq [78]. For the video input, the lip region is detected through face detection [79] and landmark detection [80] and we crop the region with a size of 96×96. Every input frame is converted into grayscale. For data augmentation purposes, horizontal flipping and random cropping into a size of 88×88 are applied to the visual inputs during training. For masked prediction of speech units, α is set to 5 at the last iteration and masking is performed by substituting with random contiguous frames of the same video, following [3]. For training LMDecoder, the audio input is resampled to 16kHz and quantized through the aforementioned pre-trained HuBERT [64]. For the text tokenizer, we use a subword-level tokenizer, sentencepiece [81], and set the dictionary size to 1,000 for all languages.

The visual encoder and LMDecoder are pre-trained separately. The visual encoder is trained on audio-visual data of a high-resource language, English, and we directly utilize the pre-trained model of [3] for the visual encoder. LMDecoder is trained on audio-text paired data in the target language. For each language (*i.e.*, ES, FR, IT, and PT), we use the data corresponding to the target language from Multilingual LibriSpeech (MLS) [82], the audio-text paired dataset, to train LMDecoder. After training the visual encoder and LMDecoder, we combine them with an attention layer and finetune the entire model on the lip reading data (*i.e.*, video-text paired data) of the target language. For the objective function to train LMDecoder and finetuning the entire lip reading models on the target language video-text paired data, we use hybrid CTC/attention loss [83]. For decoding, we do not use an external language model and the joint CTC/attention decoding, and only utilize the output of the decoder for all experiments.

For pre-training LMDecoder and finetuning the entire lip reading model in the target languages, we employ Adam [84] optimizer and tri-stage learning rate schedules for all experiments. The peak learning rate is set to 0.001, and the warmup stages for pre-training and finetuning are set to 15,000 steps and 10,000 steps, respectively. LMdecoder is trained for 60,000 steps on audio-text paired data of the target language. For finetuning the lip reading model, we train the model for 50,000 steps by using video-text paired data in the target language, except for English. We train the English lip reading model for 30,000 steps by freezing the visual encoder for 20,000 steps. Further details can be found in the supplementary material.

4.3. Dataset

Multilingual TEDx (mTEDx) [26] is a multilingual TEDx corpus for speech recognition and translation. The dataset is composed of speech audio and transcriptions, for 8 languages. In order to use the dataset in lip reading, we download the video from Youtube by using the links provided by the dataset. Based on the data splits of the dataset, we follow [25] to remove the video not containing a speaker and unavailable video online. We utilize Spanish (ES), French (FR), Italian (IT), and Portuguese (PT) to evaluate the proposed method. The dataset size of each language is represented in Table 1.

Multilingual LibriSpeech (MLS) [82] dataset is a large multilingual audio-text paired dataset for Audio-based Speech Recognition (ASR). The dataset is derived from au-

Modality	Dataset	Train	Validation	Test
A-T	MLS-ES	918	10	10
A-T	MLS-FR	1,077	10	10
A-T	MLS-IT	247	5	5
A-T	MLS-PT	161	4	4
V-T	mTEDx-ES	74	0.7	0.5
V-T	mTEDx-FR	86	0.4	0.3
V-T	mTEDx-IT	47	0.4	0.4
V-T	mTEDx-PT	93	0.7	0.7

Table 1. Data length (Hours) of each dataset. A-T represents audio-text paired data and V-T represents video-text paired data.

diobooks and consists of 8 languages. We utilize ES, FR, IT, and PT languages to train the proposed LMDecoder to learn language-specific knowledge. The dataset size of each language is represented in Table 1. Please note that the available audio-text paired data is much larger than video-text paired data (*i.e.*, mTEDx).

LRS3 [4] is a large-scale English sentence-level audio-visual dataset. It consists of about 439 hours of videos. We use 433 hours of training data to pre-train the visual encoder with masked predictions of speech units, for learning general speech knowledge.

VoxCeleb2 [85] is a large-scale unlabeled audio-visual dataset. It consists of about 2,442 hours of videos. We use 1,326 hours of training data following [3] to pre-train the visual encoder along with the LRS3 dataset.

LRS2 [7] is another large-scale English sentence-level audio-visual dataset derived from television shows. It has about 224 hours of data. We use the dataset to evaluate the effectiveness of the proposed lip reading framework in the high-resource language, English.

4.4. Baselines

In order to analyze the effectiveness of the proposed method in developing lip reading models for low-resource languages, we set five baselines to be compared. All the methods are implemented with the same settings.

Supervised pre-training. This baseline is to evaluate whether a well-trained lip reading model in a high-resource language, English, can be employed for other languages. To this end, we pre-train a state-of-the-art lip reading model, **CM-seq2seq** [15], on large-scale labeled datasets in English, a total amount of 814 hours composed of LRW [5], LRS2 [7], and LRS3 [4]. Then, the entire pre-trained model is directly finetuned on each target language.

Self-supervised pre-training of encoders. This baseline is to evaluate whether the learned general speech knowledge, the ability to model speech units from lip movements, is beneficial when it is applied to other languages. To this end, we only utilize the visual encoder pre-trained in a high-resource language (*i.e.*, 1,759 hours of English data) with the objective of masked prediction of speech units. Then, a decoder, to be trained from scratch, is attached to the visual encoder to construct the lip reading pipeline and trained on the target lip reading dataset. The model is trained on a total of 50K iterations and the pre-trained visual encoder is frozen until 20K iterations. Since this method can be viewed as the application of AV-HuBERT [3] in different languages, we denote this method as **AV-HuBERT**.

Method	WER (%)
Afouras *et al.* [49]	58.5
Zhang *et al.* [2]	51.7
TM-seq2seq [12]	48.3
CroMM-VSR [19]	46.2
MVM [22]	44.5
CM-seq2seq [15]	37.9
Prajwal *et al.* [47]	28.9
Auxiliary Task [25]	28.7
AV-HuBERT [3]	25.5
VATLM [86]	24.3
Proposed Method	**23.8**

Table 2. Comparisons with state-of-the-art methods on LRS2.

Pre-training of decoders. This baseline is to evaluate the effectiveness of pre-training of decoders on audio-text paired data. To this end, we pre-train a decoder through ASR task on each target language data of MLS [82] by attaching it to a pre-trained audio encoder of [3]. After training, the decoder is attached to the pre-trained visual encoder and the entire model is finetuned on the target lip reading dataset. We denote this method as **ASR Pre-train**. This method can be viewed as the absence of LM and quantized speech units in the proposed method.

Distillation of pre-trained knowledge. This baseline is to evaluate the effectiveness of the knowledge distillation-based method of [25] in low-resource language lip readings. To this end, we first pre-train both lip reading and ASR models initialized from AV-HuBERT [3] on the target lip reading dataset. Then, by utilizing the pre-trained lip reading and ASR models as teachers, a new lip reading model initialized from AV-HuBERT is trained by distilling the knowledge of the two teachers. We follow other training configurations of [25] to train the model, and denote this method as **Auxiliary Task**.

Proposed Method. The final method is the proposed method that utilizes the pre-trained general speech knowledge and language-specific knowledge. To evaluate the effectiveness of the proposed method, we pre-train the visual encoder in a high-resource language and LMDecoder on the target language data from MLS [82]. Then, the two modules are attached by using an attention layer and the entire model is finetuned on the target lip reading dataset.

5. Experimental Results

5.1. Comparison with the state-of-the-art methods

Before evaluating the lip reading performances for the low-resource languages, we first evaluate the effectiveness of the proposed framework on a high-resource language dataset, LRS2 [7]. To this end, we train our LMDecoder with the combination of training datasets of LRS2 and LRS3. Then, the LMDecoder is attached to the pre-trained visual encoder and finetuned on LRS2 dataset. The evaluation results on LRS2, are shown in Table 2. We compare the performances obtained by using 'video-text data' of LRS2 only, if some works utilize extra video-text data, we report the performance obtained by using minimum extra video-text data. The proposed method outperforms the previous state-of-the-art methods and sets a new state-of-the-art performance, by achieving 23.8% WER. In particular, the proposed method outperforms AV-HuBERT [3] that shares the same visual encoder by 1.7% WER, which means that the proposed LMDecoder can contribute to even high-resource language lip-reading by enriching language modeling ability.

5.2. Effectiveness in low-resource languages

To evaluate the effectiveness of the different methods on low-resource lip reading, we compare the performances of five different lip reading methods described in Sec. 4.4 on four low-resource languages, ES, FR, IT, and PT. Table 3 shows the comparison results on mTEDx-IT, Table 4 shows results on mTEDx-FR, Table 5 shows results on mTEDx-ES, and Table 6 shows results on mTEDx-PT.

Effectiveness of learning general speech knowledge. Firstly, we compare *CM-seq2seq* and *AV-HuBERT* to confirm whether learning lip reading in a high-resource language using large-scale labeled video-text paired data is better or learning general speech knowledge from a high-resource language is better for low-resource languages lip reading. *CM-seq2seq* is pre-trained on the lip reading task using 814 hours of English video-text data and then finetuned on each target language, while *AV-HuBERT* is pre-trained on the speech units prediction task using 1,759 hours of English audio-visual data and then finetuned on each target language. *CM-seq2seq* achieves 88.41% WER and *AV-HuBERT* achieves 77.36% WER, on French shown in Table 4. The results indicate that even if *CM-seq2seq* is utilized large-scale English labeled data, the knowledge cannot be fully transferred for French lip reading. On the other hand, *AV-HuBERT*, which only utilizes labeled data of French, achieves better results by expanding the general speech knowledge learned from English data into French. Similar tendencies can be found in other languages, IT, ES, and PT in Tables 3, 5, and 6. The results indicate that it would be beneficial to learn how to encode general speech units instead of learning to translate the lips into text in a high-resource language, for the purpose of adapting the pre-trained model to low-resource languages.

Method	Unlabeled V-A Data	Labeled A-T Data	Labeled V-T Data	WER
CM-seq2seq [15]	-	-	47h (+814h)	78.31%
AV-HuBERT [3]	1759h	-	47h	73.24%
ASR Pre-train	1759h	294h	47h	71.28%
Auxiliary Task [25]	1759h	47h	47h	71.99%
Proposed Method	1759h	294h	47h	**68.04%**

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

Method	Unlabeled V-A Data	Labeled A-T Data	Labeled V-T Data	WER
CM-seq2seq [15]	-	-	86h (+814h)	88.41%
AV-HuBERT [3]	1759h	-	86h	77.36%
ASR Pre-train	1759h	1163h	86h	75.67%
Auxiliary Task [25]	1759h	86h	86h	76.79%
Proposed Method	1759h	1163h	86h	**74.74%**

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

Effectiveness of pre-training the decoder. In order to evaluate the effectiveness of pre-training the decoder using audio-text data, we compare *AV-HuBERT* and *ASR Pre-train*. The decoder of *AV-HuBERT* is trained from scratch by using the video-text paired data of the target language, while the decoder of *ASR Pre-train* is trained after initializing with the pre-trained model on audio-text paired data of the target language. Since the visual encoders of the two models are the same, we can focus on the effects of pre-training the decoder using audio-text paired data. As shown in Table 5, the performance of *AV-HuBERT* on Spanish is 71.68% WER while *ASR Pre-train* achieves 70.80% WER. The results confirm that even if we learn general speech knowledge from a large-scale English dataset, the ability to model language might be insufficient to be learned from the small-scale target language dataset (*i.e.*, video-text paired dataset). By adapting the language knowledge learned from an audio-text paired dataset which is larger than the video-text paired dataset, we can improve the lip reading performances for the low-resource languages. Results for other languages, IT, FR, and PT, are shown in Tables 3, 4, and 6.

Effectiveness of learning language-specific knowledge through LMDecoder. In order to evaluate the effectiveness of the proposed method of learning language-specific knowledge through LMDecoder, we compare *ASR Pre-train* and *Proposed Method*. Different from *ASR Pre-train*, the *Proposed Method* is additionally trained to save

Method	Unlabeled V-A Data	Labeled A-T Data	Labeled V-T Data	WER
CM-seq2seq [15]	-	-	74h (+814h)	81.75%
AV-HuBERT [3]	1759h	-	74h	71.68%
ASR Pre-train	1759h	992h	74h	70.80%
Auxiliary Task [25]	1759h	74h	74h	70.91%
Proposed Method	1759h	992h	74h	**70.16%**

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

Method	Unlabeled V-A Data	Labeled A-T Data	Labeled V-T Data	WER
CM-seq2seq [15]	-	-	93h (+814h)	79.17%
AV-HuBERT [3]	1759h	-	93h	71.87%
ASR Pre-train	1759h	254h	93h	70.39%
Auxiliary Task [25]	1759h	93h	93h	70.39%
Proposed Method	1759h	254h	93h	**69.33%**

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

language-specific audio features in LM and to construct the mapping between speech units and language-specific audio features by using quantized speech units. Therefore, by comparing the two methods, we can validate the effectiveness of the proposed lip reading framework for low-resource languages. *ASR Pre-train* achieves 71.28% WER on Italian while the *Proposed Method* achieves 68.04% WER which outperforms *ASR Pre-train* by about 3.24% WER, shown in Table 3. Since the proposed LMDecoder can transform the encoded visual features (*i.e.*, speech units) into language-specific audio features, it can fully utilize the learned general speech knowledge of the pre-trained visual encoder, when the two pre-trained modules (*i.e.*, visual encoder and LMDecoder) are combined. Moreover, as the LM can provide rich speech information of audio by reading the memory banks, we can also employ the complementary effects of multi-modality as proven to be effective for lip reading in previous works [19, 20, 22]. Similar tendencies can be found in tables 4, 5, and 6.

Comparison with distillation-based method. Finally, we compare the lip reading performance with a distillation-based method that utilizes knowledge distillation as an auxiliary task. In Tables 3, 4, 5, and 6, compared to *Auxiliary Task*, the *Proposed Method* outperforms the method in all languages. Even if *Auxiliary Task* tried to learn from using the knowledge of superior models (*i.e.*, pre-trained lip reading and ASR models), the *Proposed Method* can achieve better performance by employing language-specific knowledge learned from a larger audio-text paired dataset.

Comparing the performances of *Proposed Method* with other baselines, we can confirm the effectiveness of the proposed lip reading framework for low-resource languages.

5.3. Ablation study

Different audio-text paired datasets. We perform ablation studies to examine the effectiveness of the proposed lip reading framework. Firstly, we examine the effect of different audio-text paired datasets in learning language-specific knowledge of LMDecoder. We pre-trained three variants of LMDecoder by using MLS, mTEDx, and both datasets. Then, each model is attached to the pre-trained visual encoder, and the entire model is finetuned on the target language lip reading dataset. For the ablation study, we use Italian datasets (*i.e.*, MLS-IT and mTEDx-IT). The ablation results are shown in Table 7. MLS-IT dataset has 247 hours of training data and mTEDx-IT dataset has 47 hours of training data. Using MLS-IT only to train LMDecoder achieves 70.45% WER. By using an extra audio-text paired dataset, MLS-IT, to train language-specific knowledge for LMDecoder, we can improve the performance from the baseline that uses the scratch decoder (*i.e.*, 73.24% WER) by 2.79% WER. Moreover, by using audio-text paired data of mTEDx-IT only, we can still improve the performance and achieve 71.47% WER. This shows the effectiveness of the LM in providing the saved language-specific audio features corresponding to speech units. By using both datasets, the performance improves to 68.04% WER, which shows the effectiveness of learning language-specific knowledge using audio-text paired data in building low-resource language lip reading models.

Train data for LMDecoder	Labeled A-T Data	Labeled V-T Data	WER
Baseline Decoder	0h	47h	73.24%
MLS-IT	247h	47h	70.45%
mTEDx-IT	47h	47h	71.47%
MLS-IT+mTEDx-IT	294h	47h	**68.04%**

Table 7. Ablation study using different audio-text paired data.

Method	Unlabeled V-A Data	Labeled A-T Data	Labeled V-T Data	WER
Without LM	1759h	293h	46h	71.01%
With LM	1759h	293h	46h	**68.04%**

Table 8. Ablation study with and without LM on mTEDx-IT.

Effectiveness of Language-specific Memory (LM). To evaluate the effectiveness of Langauge-specific Memory (LM) in LMDecoder, we experiment by eliminating the LM from the proposed method. Therefore, the decoder of *Without LM* model is trained with quantized speech units but the LM is not included. The performance of *Without LM* on Italian is shown in Table 8. By eliminating the proposed

LM, the lip reading performance is degraded by about 3% WER. The result clearly indicates that the saved language-specific audio features in LM can provide beneficial information when it is combined with general speech knowledge, with the following two roles; 1) constructing mapping between speech units and language-specific audio features, and 2) providing rich auditory information which can complement the lip reading model.

Different amounts of video-text data. In order to investigate the effectiveness of the proposed method under different amounts of video-text data situations, we experiment with 1/3 (15.7h), 2/3 (31.3h), and all (47h) of the video-text data of mTEDx-IT. This experiment is to confirm how much the low resources the model can handle. The results are shown in Table 9. When only 15.7 hours of labeled video-text data are used, it achieves 75.62% WER, which shows the model cannot correctly learn from only 1/3 of the data. When 31.3 hours of data are utilized, the WER performance is 69.63%. This performance is better than that of the other methods obtained using the full data in Table 3. The results indicate that the proposed method can perform well even with 2/3 of the data on mTEDx-IT by outperforming the previous methods trained on full data. By using 100% of data (47h), the performance is improved to 68.04% WER.

Different amounts of audio-text data. We also experiment with different amounts of audio-text data including the cases where the audio-text data is even smaller than the video-text data (47h). The results are shown in Table 10. The results indicate too small audio data (12h) leads to even worse performance than the randomly initialized decoder (*i.e.*, 73.24% WER). We found that when we use about 75% amount (35h) of the video-text data, it starts to improve the performance. By using more audio-text data, we can improve the performance more. When using 147h and 294h of the audio-text paired data, we achieve 70.3% and 68.0% WERs, respectively, on mTEDx-IT dataset.

Performances of pre-trained ASR models. We provide the performances of the pre-trained ASR models on each audio-text paired dataset, before being applied to the lip reading tasks. We also provide the performances of the pre-trained LMDecoders on each audio-text paired dataset, before being applied to the lip reading tasks. Different from the ASR models, LMDecoders are trained from quantized audio units with LM while the ASR models are trained from continuous audio. Please note the objective of pre-training the LMDecoder is for applying it to lip reading, not for performing ASR (*i.e.*, Audio-based Speech Recognition). The WER(%) results are shown in Table 11. As the results indicate, the ASR performances of LMDecoder do not perform better than the *ASR Pre-train*. However, the lip reading performances for low-resource languages of the proposed LMDecoder outperform the *ASR Pre-train* as shown in Tables 3, 4, 5, and 6. From the results, we can confirm that the proposed pre-training strategies are more suitable for low-resource language lip reading than just pre-training a decoder through ASR.

V-T Data Amount	15.7h	31.3h	47h
WER	75.62%	69.63%	68.04%

Table 9. Ablation study using different amounts of video-text paired data on mTEDx-IT.

A-T Data Amount	12h	35h	47h	147h	297h
WER	86.7%	72.2%	71.5%	70.3%	68.0%

Table 10. Ablation study using different amounts of audio-text paired data on mTEDx-IT.

Dataset	ASR Pre-train	LMDecoder
mTEDx-IT	24.65%	29.21%
mTEDx-FR	22.96%	27.48%
mTEDx-ES	25.01%	24.65%
mTEDx-PT	28.35%	36.01%

Table 11. Performances of pre-trained models (ASR) on mTEDx.

6. Conclusion

This paper proposed a novel lip reading framework for low-resource languages. To address the challenge of insufficient video-text paired data of low-resource languages, we proposed to learn and combine general speech knowledge and language-specific knowledge. Specifically, the visual encoder is trained with masked predictions of speech units to learn general speech knowledge, and Language-specific Memory-augmented Decoder (LMDecoder) is proposed to learn language-specific knowledge from audio-text paired data. By combining the learned general speech knowledge and language-specific knowledge, we can efficiently develop lip reading models for low-resource languages. Through comprehensive experiments on a total of five languages (English, Italian, French, Spanish, and Portuguese), we verified the effectiveness of the proposed lip reading framework in low-resource languages.

7. Acknowledgment

This work was partly supported by two funds: the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2C2005529) and IITP grant funded by the Korea government(MSIT) (No.2020-0-00004, Development of Previsional Intelligence based on Long-Term Visual Memory Network)

References

[1] Barbara Ed Dodd and Ruth Ed Campbell. *Hearing by eye: The psychology of lip-reading*. Lawrence Erlbaum Associates, Inc, 1987.

[2] 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.

[3] 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.

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

[5] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 87–103. Springer, 2017.

[6] 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.

[7] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3444–3453. IEEE, 2017.

[8] 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.

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

[10] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*, 2017.

[11] Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. End-to-end audiovisual speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6548–6552. IEEE, 2018.

[12] 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*, 44(12):8717–8727, 2018.

[13] Jingyun Xiao, Shuang Yang, Yuanhang Zhang, Shiguang Shan, and Xilin Chen. Deformation flow based two-stream network for lip reading. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 364–370. IEEE, 2020.

[14] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip to speech synthesis with visual context attentional gan. *Advances in Neural Information Processing Systems*, 34:2758–2770, 2021.

[15] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE, 2021.

[16] 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.

[17] 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.

[18] 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.

[19] Minsu Kim, Joanna Hong, Se Jin Park, and Yong Man Ro. Cromm-vsr: Cross-modal memory augmented visual speech recognition. *IEEE Transactions on Multimedia*, 24:4342–4355, 2021.

[20] Minsu Kim, Joanna Hong, Se Jin Park, and Yong Man Ro. Multi-modality associative bridging through memory: Speech sound recollected from face video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 296–306, 2021.

[21] 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.

[22] Minsu Kim, Jeong Hun Yeo, and Yong Man Ro. Distinguishing homophenes using multi-head visual-audio memory for lip reading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1174–1182, 2022.

[23] Dalu Feng, Shuang Yang, Shiguang Shan, and Xilin Chen. Learn an effective lip reading model without pains. *arXiv preprint arXiv:2011.07557*, 2020.

[24] Pingchuan Ma, Brais Martinez, Stavros Petridis, and Maja Pantic. Towards practical lipreading with distilled and efficient models. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7608–7612. IEEE, 2021.

[25] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, pages 1–10, 2022.

[26] Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. Multilingual tedx corpus for speech recognition and translation. In *Proceedings of Interspeech*, 2021.

[27] Tanja Schultz and Alex Waibel. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35(1-2):31–51, 2001.

[28] Ngoc Thang Vu, David Imseng, Daniel Povey, Petr Motlicek, Tanja Schultz, and Hervé Bourlard. Multilingual deep neural

network based acoustic modeling for rapid language adaptation. In *2014 IEEE international Conference on acoustics, speech and signal processing (ICASSP)*, pages 7639–7643. IEEE, 2014.

[29] Mingshuang Luo, Shuang Yang, Xilin Chen, Zitao Liu, and Shiguang Shan. Synchronous bidirectional learning for multilingual lip reading. *arXiv preprint arXiv:2005.03846*, 2020.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[31] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[33] Xing Zhao, Shuang Yang, Shiguang Shan, and Xilin Chen. Mutual information maximization for effective lip reading. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 420–427. IEEE, 2020.

[34] Yuanhang Zhang, Shuang Yang, Jingyun Xiao, Shiguang Shan, and Xilin Chen. Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 356–363. IEEE, 2020.

[35] Joanna Hong, Minsu Kim, Se Jin Park, and Yong Man Ro. Speech reconstruction with reminiscent sound via visual voice memory. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3654–3667, 2021.

[36] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip-to-speech synthesis in the wild with multi-task learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[37] Jeongsoo Choi, Minsu Kim, and Yong Man Ro. Intelligible lip-to-speech synthesis with speech units. *arXiv preprint arXiv:2305.19603*, 2023.

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

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[40] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[41] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[42] Xinshuo Weng and Kris Kitani. Learning spatio-temporal features with two-stream deep 3d cnns for lipreading. *arXiv preprint arXiv:1905.02540*, 2019.

[43] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323. IEEE, 2020.

[44] Alex Graves and Alex Graves. Connectionist temporal classification. *Supervised sequence labelling with recurrent neural networks*, pages 61–93, 2012.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[46] Joanna Hong, Minsu Kim, Daehun Yoo, and Yong Man Ro. Visual context-driven audio feature enhancement for robust end-to-end audio-visual speech recognition. *arXiv preprint arXiv:2207.06020*, 2022.

[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.

[48] Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18783–18794, 2023.

[49] 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.

[50] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[51] Jeong Hun Yeo, Minsu Kim, and Yong Man Ro. Multi-temporal lip-audio memory for visual speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[52] Jeong Hun Yeo, Minsu Kim, Jeongsoo Choi, Dae Hoe Kim, and Yong Man Ro. Akvsr: Audio knowledge empowered visual speech recognition by compressing audio knowledge of a pretrained model, 2023.

[53] Ibrahim Almajai, Stephen Cox, Richard Harvey, and Yuxuan Lan. Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2722–2726. IEEE, 2016.

[54] Minsu Kim, Hyunjun Kim, and Yong Man Ro. Speaker-adaptive lip reading with user-dependent padding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 576–593. Springer, 2022.

[55] Minsu Kim, Hyung-Il Kim, and Yong Man Ro. Prompt tuning of deep neural networks for speaker-adaptive visual speech recognition. *arXiv preprint arXiv:2302.08102*, 2023.

[56] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[57] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020.

[58] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049, 2020.

[59] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382, 2020.

[60] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.

[61] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[67] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[68] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022.

[69] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021.

[70] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019.

[71] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*, 2021.

[72] Xuankai Chang, Brian Yan, Yuya Fujita, Takashi Maekaku, and Shinji Watanabe. Exploration of efficient end-to-end asr using discretized input from self-supervised learning. *arXiv preprint arXiv:2305.18108*, 2023.

[73] Minsu Kim, Jeongsoo Choi, Dahun Kim, and Yong Man Ro. Many-to-many spoken language translation via unified speech and text representation learning with unit-to-unit translation. *arXiv preprint arXiv:2308.01831*, 2023.

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

[75] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[76] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[77] 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.

[78] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

[79] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: Dual shot face detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[80] Hong Joo Lee, Seong Tae Kim, Hakmin Lee, and Yong Man Ro. Lightweight and effective facial landmark detection using adversarial learning with face geometric map generative network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(3):771–780, 2019.

[81] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

[82] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411, 2020.

[83] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253, 2017.

[84] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[85] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

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