UniLip: Learning Visual-Textual Mapping with Uni-Modal Data for Lip Reading

Bingquan Xia[1,2]
xiabingquan21s@ict.ac.cn
Shuang Yang[1,2]
shuang.yang@ict.ac.cn
Shiguang Shan[1,2]
sgshan@ict.ac.cn
Xilin Chen[1,2]
xlchen@ict.ac.cn

[1] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China

[2] University of Chinese Academy of Sciences, Beijing, China

Abstract

In this paper, we propose a novel way, namely **UniLip**, to utilize uni-modal texts and uni-modal talking face videos for lip reading. With only uni-modal data, we achieve totally unsupervised lip reading for the first time. We reformulate the lip reading task with uni-modal data into two sub-tasks: learning linguistic priors from uni-modal texts and mapping uni-modal videos to texts under the constraint of the learned linguistic priors. We formulate the two sub-tasks as language modeling and conditional generation tasks, respectively, and introduce a multi-grained adversarial learning strategy to embed these two sub-tasks into a unified framework. Specifically, for the language modeling sub-task, we introduce a novel multi-grained discrimination strategy based on local n-gram sub-utterances, considering the linguistic features often related to local language patterns such as word spelling and grammar correctness. For the conditional generation sub-task, we first leverage self-supervised models to extract base visual features, and then construct a generator by adapting them to our task with a multi-grained feature fusion module that aggregates both local and global information. With only uni-modal data, we yield a best unsupervised Word Error Rate of 51.2% and 57.3% on LRS3 and LRS2, respectively. The result on LRS3 is comparable with mainstream supervised models trained on it. With both uni-modal and labeled data, we show that UniLip could co-work with traditional supervised frameworks. In our case, it improves supervised Seq2Seq methods by relatively 4.2% and 1.4% on LRS3 and LRS2, respectively. Our UniLip's ability to work with uni-modal data under the settings of both with and without supervision shows its great potential for real-world applications.

1 Introduction

Lip reading, also known as Visual Speech Recognition (VSR), is the task of recognizing speech by analyzing talking face videos. Its core is to build the correct mapping from visual features to textual modality, which is still a challenging task at present due to the large variations in both visual appearance and spoken words. With the availability of large-scale lip

reading datasets [2, 4, 9, 30, 35], several works[27, 31, 38] have made amazing achievements by taking advantage of labeled transcriptions. In this work, we investigate how to use easily accessible unlabeled uni-modal videos and texts for lip reading.

The main idea behind our approach is that uni-modal texts can provide valuable prior information about the target language to be recognized, including word spelling, phrase composition, and grammatical structure. Although we have no access to the exact speech content of an unlabeled uni-modal talking face video, we can use the learned linguistic knowledge of the same language to constrain the mapping of the visual features to the textual modality. For example, in English, "From the core of the earth" is more likely to be a correct output than "Frum the kare ov the erth".

Based on this idea, we take advantage of the uni-modal texts to learn the linguistic priors of the target language. The learned priors are then used to supervise the mapping process of the uni-modal videos to plausible texts. In this process, we can not require the model to map each visual input to its exact textual transcription because we have no annotated labels. But we are able to narrow the output space to a correct space by restraining the model's text outputs to adhere to the priors and characteristics of the target language. Specifically, we subdivide the lip reading task with uni-modal data into two sub-tasks: (S1) learning linguistic priors from uni-modal texts and (S2) learning to map uni-modal videos to texts consistent with such priors. For (S1), we formulate it as a language modeling task based on various uni-modal texts. For (S2), we formulate it as a text generation task conditioned on visual features, constrained by the learned priors in (S1). The overview of our method is shown in Figure 1.

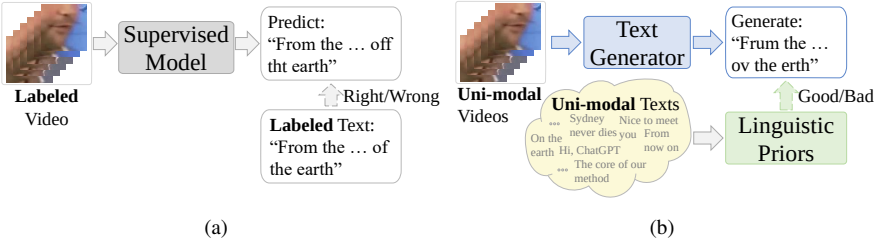

Figure 1: (a) The Traditional Supervised Approach (b) Our Approach

We face two obstacles in the learning process. Firstly, the uni-modal texts are often collected from various sources and have source-specific characteristics. Linguistic priors learned from these texts may not accurately reflect the properties of the target language because the training uni-modal texts are easily biased to specific domains and sources. Secondly, with only uni-modal data, we lack strong and explicit supervision signals to effectively capture visual speech cues from talking face videos, which poses challenges to providing solid visual conditions for mapping lip movements to spoken words.

To overcome the above two issues, we propose a multi-grained learning strategy. Firstly, even though texts collected from various sources differ in some ways, they still share common linguistic characteristics in local spelling grammars owning to the specific type of target language. The common linguistic characteristics are mainly reflected in short n-gram clips and these n-grams are capable of encoding general linguistic features and alleviating bias from text sources. We introduce a novel multi-grained discrimination strategy based on n-gram sub-utterances. It breaks down the whole text and only focuses on the realness of its n-gram clips. Secondly, we first leverage on-the-shelf self-supervised models to extract base visual features and then introduce a multi-grained feature fusion module to adapt these fea-

tures for our task by integrating both local information of individual frames and the global context of the whole sequence.

In conclusion, our contributions can be summarized as below:

- We propose UniLip, a novel way to utilize uni-modal videos and texts for lip reading.
- We subdivide lip reading with uni-modal data into two sub-tasks of language modeling and conditional text generation, and embed them into a unified training framework by a multi-grained adversarial learning strategy.
- We demonstrate UniLip's advantages to utilize uni-modal data both with and without labeled data and provide further analyses of its capacity to work with data of different scales and sources.

2 Related Work

2.1 Lip Reading

Existing lip reading methods can be categorized into two categories: fully-supervised methods and self-supervised methods. In the fully-supervised category, the model typically consists of a visual front-end, an encoder, and a decoder. These methods mainly rely on large-scale labeled data to learn how to encode speech-related visual features from talking face videos and decode them into corresponding textual transcriptions. Several works attempt to tackle lip reading by proposing effective network architectures to achieve better feature extraction, feature integration, or decoding, including attention-enhanced visual front-end[27], MS-TCN encoder[24] and transducer-based decoder[23], respectively. Recently, some works have explored multi-modal self-supervised pre-training with unlabeled data and finetuning on labeled data for the final lip reading task[14, 28, 39]. They utilize unlabeled audio-video pairs to perform pre-training with a multi-modal masked prediction task and then finetune on labeled video-text pairs. In this work, we propose a way to both perform unsupervised training with unlabeled data and semi-supervised training with labeled and unlabeled data.

2.2 Self-supervised and Unsupervised Automatic Speech Recognition

In recent years, numerous self-supervised automatic speech recognition(ASR) methods have emerged. Most methods are based on a masked prediction strategy, such as wav2vec 2.0[5], HuBERT[16], W2V-BERT[11] and BEST-RQ[8]. With the availability of high-quality auditory pre-trained models, researchers have attempted to completely eliminate the reliance on labeled textual transcriptions in ASR. One notable example is wav2vec-U[6] and its successor wav2vec-U 2.0[19]. Prior to these advancements, early works have also made valuable contributions to unsupervised ASR[7, 18, 20]. Their ideas of adversarial training have inspired later works. However, even though lip reading and ASR share a lot in task formulation, there have never been efforts made to achieve entirely unsupervised lip reading. This is mainly because current unsupervised ASR methods severely rely on high-quality speech representations, which are still hard to obtain in lip reading.

3 UniLip

3.1 Overview

Given an unlabeled talking face video $\mathbf{x} = (x_1, x_2, \cdots, x_T)$ with T frames, the goal of our method is to recognize its speech contents. In this work, we aim to learn the semantic mapping from the talking face videos to their textual transcriptions by taking advantage of

uni-modal videos and texts. We decompose this target into two steps: (S1) learning linguistic priors from uni-modal texts of the target language and (S2) mapping the uni-modal videos to texts with constraints of the learned linguistic priors. Then we unify the two tasks via an adversarial training framework.

For (S1), we formulate it as a language modeling task and introduce a multi-grained discriminator $\mathcal{D}$ to finish it by judging inputs from uni-modal texts as real. We use causal temporal convolutions to construct $\mathcal{D}$. This causal design requires each judgment only made based on its history inputs, which is similar to causal language modeling. Formally, given a sentence $\mathbf{t} = (t_1, t_2, \cdots, t_T)$ from the text corpus $\mathcal{T}$, it is tokenized into a sequence of phonemes. Then each phoneme in the sequence is encoded as a one-hot vector. The one-hot encoded phoneme sequence is denoted as $\mathbf{p} = (p_1, p_2, \cdot, p_{T'})$, where each p_i is a $|V|$-dimensional one-hot vector and $|V|$ is the size of the phoneme set. Finally, with the input $\mathbf{p}$, $\mathcal{D}$ is trained to output a probability as close to 1 as possible for the input uni-modal texts, maximizing the confidence of $\mathbf{p}$ being real. We choose phoneme for tokenization here because each phoneme could represent a visual vocal pattern, making it suitable for our task.

For (S2), we formulate it as a text generation task conditioned on visual features of lip movements, which is constrained by the linguistic priors learned in (S1). We introduce a generator $\mathcal{G}$ to finish this task. $\mathcal{G}$ receives visual inputs $\mathbf{x} = (x_1, x_2, \cdots, x_T)$, and outputs a sequence of phoneme posterior distributions $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_{T''})$ where $\hat{p}_i, \forall i \in 1, \cdots, T''$ is a distribution over the phoneme set V. Then $\mathcal{G}$ sends $\hat{\mathbf{p}}$ to $\mathcal{D}$ and tries to deceive $\mathcal{D}$ into outputting a probability close to 1.

During training, $\mathcal{G}$ receives the extracted raw features of uni-modal videos and generates phonemes, and is optimized by the loss that $\mathcal{D}$ judges the realness of the generated phonemes. $\mathcal{D}$ receives both real phonemes from uni-modal texts and fake phonemes from $\mathcal{G}$, and is optimized by the losses to judge real phonemes as real and fake phonemes as fake. In this adversarial game, if $\mathcal{G}$ succeeds in producing outputs that are indistinguishable by $\mathcal{D}$, it is believed to be able to produce reasonable textual transcriptions based on uni-modal talking face videos. During inference, $\mathcal{G}$'s outputs are decoded into word-level sentences via HLG decoding built with k2[1]. The whole framework is shown in Figure 2.

3.2 Multi-grained Learning of Linguistic Priors

In this section, we elaborate on our detailed design of $\mathcal{D}$. When learning linguistic priors from uni-modal texts, we do not impose any restrictions on the source of texts so that our method could be universally applied to texts collected in various ways.

Generally, texts collected from different sources often share some local linguistic properties, such as grammar and spelling patterns, but they may also differ in some holistic aspects, such as sentence average length and topic. Based on this observation, we aim to build a multi-grained criterion to learn the consistent and valid priors of the target language from different texts considering removing biases introduced by different text sources.

Specifically, we propose GramNet, short for ngram-based discrimination network, to formulate $\mathcal{D}$. GramNet focuses on capturing multi-grained linguistic characteristics from local n-grams so as to reduce source-specific biases among different texts. Because these short n-gram clips mainly reflect the rule of basic linguistic characteristics, they allow us to capture underlying language properties that are less influenced by text sources.

Formally, GramNet clips the real or fake input sequence into n n-gram clips along the temporal dimension and feeds each n-gram into 3 causal convolutional layers, then calculates

[1] https://github.com/k2-fsa

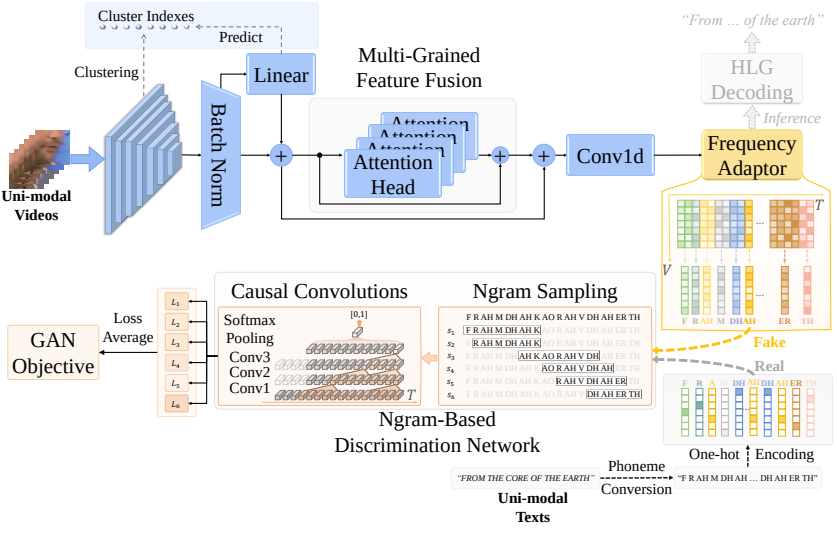

Figure 2: Model Architecture

the loss of the input by mean-averaging the losses of these n n-gram clips, as shown in Figure 2. We denote the minimum and maximum lengths of n-grams as $l_{\min}$ and $l_{\max}$ respectively. Given an input phoneme sequence $\mathbf{p} = (p_1, p_2, \cdots, p_n)$, the start position of its i-th n-gram st_i is set to $i * l_{\min}$, and its length is sampled from the uniform distribution $U(l_{\min}, l_{\max})$. This sampling strategy ensures that all the n-grams uniformly cover the whole sequence, which promises the input to be fully utilized.

3.3 Multi-grained Visual-Textual Mapping

In this section, we elucidate our design of $\mathcal{G}$. In our framework, $\mathcal{G}$ is responsible for mapping videos to texts with constraints of the linguistic priors. It mainly includes a BatchNorm layer, a linear layer, and a 1D temporal convolution layer. In the forward pass, $\mathcal{G}$ receives a sequence of talking face images and outputs a sequence of textual posterior distributions. The output sequence is then down-sampled along the temporal dimension by merging adjacent time steps that share the same phoneme prediction[19], denoted by "Frequency Adaptor" in Figure 2. However, due to the small receptive field of the 1D temporal convolution, $\mathcal{G}$'s output at each time step is just a local aggregation of the input features from adjacent visual frames. This assumes a strong local correspondence between input visual features and output texts, which is hard to stand due to homophones and the difficulty of visual feature extraction. Even though we leverage pre-trained models to extract base visual features, we empirically prove that these features are not capable of directly building a mapping with texts in our task, as Section 5.1 suggests.

To alleviate this issue, we ease the above localness assumption by providing the generator with global context information to obtain multi-grained visual feature integration. Formally, we add a temporal self-attention module[32] after the addition block of the BatchNorm and linear layers to extract the global temporal context and fuse its output with the addition block's output through summation.

3.4 Optimization Objective

In our setup, the overall optimization objective includes the classical GAN objective and four task-oriented auxiliary losses:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathop{\mathbb{E}}_{\mathbf{p}\sim\mathcal{P}} [\log(\mathcal{D}(\mathbf{p}))] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{X}} [\log(\mathcal{D}(\mathcal{G}(\mathbf{x}))] - \lambda\mathcal{L}_{gp} + \gamma\mathcal{L}_{sp} + \varepsilon\mathcal{L}_{pd} + \mu\mathcal{L}_{aux}, \quad (1)$$

where $\mathbf{p} \sim \mathcal{P}$ is a sequence of one-hot encoded phonemes from uni-modal texts and $\mathbf{x} \sim \mathcal{X}$ is a sequence of lip images of uni-modal videos. The first two terms are the classical GAN objective[12] on real and fake samples, respectively. $\mathcal{L}_{gp}$ is a gradient penalty[13] to aid $\mathcal{D}$'s training. $\mathcal{L}_{sp}$, $\mathcal{L}_{pd}$ and $\mathcal{L}_{aux}$ are specialized for $\mathcal{G}$ to generate more plausible texts, which stand for smoothness penalty, phoneme diversity, and auxiliary prediction loss, respectively.

Gradient Penalty. We use the gradient penalty technique to constrain the gradient norm of $\mathcal{D}$ with respect to mixtures of real and fake text samples. This approach has been shown to be beneficial for convergence[13]. Let $\mathbf{x}'$ denote a mixture of real and fake samples and the gradient penalty is as below:

$$\mathcal{L}_{gp} = \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{X}'} \left[(||\nabla\mathcal{D}(\mathbf{x}')|| - 1)^2 \right]. \quad (2)$$

Smoothness Penalty. Via phoneme-level CTC decoding, we find that adjacent phoneme distributions in $\mathcal{G}$'s outputs often correlate with the same phoneme segment. This indicates that $\mathcal{G}$'s outputs should be temporally consistent within each phoneme segment and different at phoneme segment boundaries. We introduce $\mathcal{L}_{sp}$ to achieve the two requirements:

$$\mathcal{L}_{sp} = \sum_{(x_t, x_{t+1})\in\mathcal{G}(\mathbf{x})} ||x_{t+1} - x_t||^2 - \xi. \quad (3)$$

The first term tries to force the phoneme outputs at adjacent time steps consistent with each other, and the threshold ξ ensures tolerance for segment boundaries.

Phoneme Diversity. We incorporate a phoneme diversity loss $\mathcal{L}_{pd}$ [6] to encourage $\mathcal{G}$ to produce diverse contents by maximizing the entropy of $\mathcal{G}$'s outputs:

$$\mathcal{L}_{pd} = -\frac{1}{|\mathcal{T}|} \sum_{\mathbf{x}\in\mathcal{X}} -H(\mathcal{G}(\mathbf{x})). \quad (4)$$

Auxiliary Prediction. Different from ordinary generation tasks, our $\mathcal{G}$ receives visual features instead of Gaussian noises. In this formulation, $\mathcal{G}$ may overlook the visual inputs and produces irrelevant textual outputs. We regularize $\mathcal{G}$'s outputs related to visual inputs by auxiliary prediction loss $\mathcal{L}_{aux}$ as below:

$$\mathcal{L}_{aux} = -\sum_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} \log P(y_t|x_t). \quad (5)$$

Before training, we cluster the frame-level visual features of the whole dataset and dump cluster indexes of each frame. Within a sequence of image features, the cluster index of the i-th feature x_i is denoted by y_i. $P(y_i|x_i)$ is the probability of correctly predicting the cluster index y_i by x_i, obtained by applying a linear transformation and a softmax function to $\mathcal{G}$'s BatchNorm layer output.

4 Experiments

4.1 Datasets

We choose our uni-modal videos and texts from three video and five text datasets separately in different settings, and report performances on lip reading datasets LRS3[2] and LRS2[30].

Uni-modal Videos. LRS3[2], LRS2[30], and Vox2-en-433h. LRS3 is currently the largest English lip reading dataset, with a total of 433 hours. LRS2 is a 224-hour lip reading dataset. Vox2-en-433h is a 433-hour subset of VoxCeleb2[10], containing uni-modal English talking face videos collected from YouTube videos.

Uni-modal Texts. LRS3[2], LRS2[30], TEDLIUM-v3[15], Cantab-TEDLIUM[33] and LibriSpeech[25]. LRS3 contains 0.19 million sentences, and LRS2 contains 0.14 million sentences[2]. TEDLIUM-v3 and LibriSpeech are both ASR datasets, containing 452 and 960 hours of transcribed audio, respectively. We use their textual transcriptions as uni-modal texts for experiments here, where TEDLIUM-v3 has 0.27 million sentences and LibriSpeech has 0.29 million sentences. Cantab-TEDLIUM is a text corpus containing over 7 million English sentences. Among them, LRS3, TEDLIUM-v3 and Cantab-TEDLIUM are all collected from TED talks, while LRS2 is from BBC programs and LibriSpeech is from audiobooks. In TEDLIUM-v3 and Cantab-TEDLIUM, sentences overlapped with the test split of LRS3 have been removed. To simplify notation, TEDLIUM-v3 and Cantab-TEDLIUM are denoted by TEDLIUM and Cantab, respectively.

Pre-processing. We use the recently released VatLM[39] to extract the raw visual features. When tokenizing texts, we first perform phonemization[26], and then add <SIL> token before the first word, after the last word, and between every two adjacent words with a probability of 0.5 following [6].

4.2 Experimental Settings

The 1D convolution in $\mathcal{G}$ has a kernel size of 6, a stride of 2, and a padding of 3. All of the convolution layers in $\mathcal{D}$ are of the configuration of kernel size 6, stride 1, and padding 5. The minimum and maximum lengths of n-gram clips are set to 20 and 25, respectively. $\mathcal{G}$ and $\mathcal{D}$ are trained alternately for a total of 100 epochs with a batch size of 20, optimized by Adam[17] with a fixed learning rate of 1.5e-5 and 1.5e-4, respectively. The threshold ξ in $\mathcal{L}_{sp}$ is set to 0.5. The weights γ, ε, and μ are set to 0.1, 3.0, and 0.5, respectively, and kept constant in all experiments, while different λ are adopted for different text datasets. It is typically chosen from $\{0.35, 0.40, 0.45, 0.50\}$, except LibriSpeech where it is 0.2.

4.3 Unsupervised VSR

In this section, we present the results of unsupervised VSR. The uni-modal videos are from all the training videos of LRS3, LRS2, or Vox2-en-433h. The uni-modal texts are from LRS3, LRS2, TEDLIUM, Cantab, or LibriSpeech. The models are tested on the test sets of LRS3 and LRS2.

Text Settings. We evaluate our method with two settings of the uni-modal texts: constrained and unconstrained. In the constrained setting, only a subset of texts is used to make the number of text samples equal to video samples, i.e. 0.19 million in LRS3 and 0.14 million in LRS2. In the unconstrained setting, all available texts in each text dataset are used. In Table 1, the down arrows in the last column indicate the WER reduction when enlarging texts from constrained to unconstrained, and N/A indicates failure of convergence.

Quantitative Results. On LRS3, We achieve a test WER of 51.2%, rivaling mainstream lip reading models trained with labeled data[29, 34]. On LRS2, even though LRS2 isn't utilized in VatLM's pre-training and the training videos are totally "unseen" for the feature extractor, we still yield a test WER of 57.3%.

[2]We split samples in LRS3 and LRS2 with a maximum length of 15s.

Main Conclusions. (a) UniLip's performance scales with the size of texts; (b) UniLip can effectively accommodate videos and texts from different sources. On LRS3, when the amount of text increases, UniLip shows better performance and easier convergence on Cantab and LibriSpeech. On LRS2, UniLip's ability to scale performance with the size of texts is further demonstrated, achieving better WER by average 1.6%. Among all the experiments in Table 1, 13 out of 16 successfully converge, demonstrating that UniLip could work with data from different sources. Surprisingly, constrained TEDLIUM's WER is 1.9% lower than the unconstrained one. We assume it's because the usage of TEDLIUM's texts during VatLM's[39] pre-training makes it somehow insensitive to the text scale in our experiments.

Table 1: Results of Unsupervised VSR on LRS3 and LRS2

Training Video	Training Text	Test WER/%(↓) (Constrained)	Test WER/%(↓) (Unconstrained)
LRS3	LRS3	-	51.9(-)
	TEDLIUM	**51.2**	53.1(1.9↑)
	Cantab	61.8	60.8(1.0↓)
	LibriSpeech	N/A	64.9(∞↓)
LRS2	LRS2	-	**57.2**(-)
	LRS3	59.7	57.8(1.9↓)
	TEDLIUM	58.3	57.3(1.0↓)
	Cantab	60.7	58.9(1.8↓)
	LibriSpeech	N/A	N/A

In Table 2, we compare our method with other mainstream works in lip reading. Among the supervised results, except for [23] which uses 31000 hours of labeled data, UniLip's performance is comparable with all other results, even surpassing four of them. This indicates UniLip's potential to perform lip reading even with only uni-modal data.

Table 2: Comparison with Other Works on LRS3

Method	Backbone	Criterion	Labeled Data/h	Unlabeled AV Data/h	Test WER/%()
Supervised					
Zhang *et al.* [36]	CNN	CE	855	-	60.1
Afouras *et al.* [1]	Transformer	CE	1519	-	58.9
Xu *et al.* [34]	RNN	CE	590	-	57.8
Shillingford *et al.* [29]	RNN	CTC	3886	-	55.1
Ma *et al.* [22]	Conformer	CTC+CE	433	-	46.9
Ma *et al.* [22]	Conformer	CTC+CE	590	-	43.3
Makino *et al.* [23]	RNN	Transducer	31000	-	33.6
Semi-supervised&Self-supervised					
Afouras *et al.* [3]	CNN	CTC	590	334	59.8
Zhang *et al.* [37]	Transformer	CTC	30	433	67.8
Ma et al[21]	Transformer	CE	433	1759	49.6
AV-HuBERT[28]	Transformer-Base	CE	433	1759	34.8
	Transformer-Large	CE	433	1759	28.6
VATLM[39]	Transformer-Base	CE	433	1759	34.2
	Transformer-Large	CE	433	1759	28.4
RAVEn[14]	Transformer-Base	CTC+CE	433	1759	33.1
	Transformer-Large	CTC+CE	433	1759	28.2
Unsupervised					
UniLip(Ours)	GAN	GAN	-	1759	51.2

4.4 Semi-supervised VSR

In the presence of labeled data, UniLip could also be incorporated into traditional supervised frameworks. We mainly consider the popular Seq2Seq models in this work. We show that UniLip could help Seq2Seq models bridge the gap between visual and textual modality, by evaluating the realness of phonemes generated based on the Seq2Seq model encoder's output. Specifically, we first train the unsupervised model and then freeze and save the best model. We use $\mathcal{G}$ to generate phoneme distributions based on the encoder's output and evaluate their realness with $\mathcal{D}$. The total loss is formulated as below, where $L_{seq2seq}$ is the conventional cross-entropy loss and L_{GAN} is the loss provided by $\mathcal{D}$:

$$L = L_{seq2seq} + \alpha L_{GAN}. \tag{6}$$

Settings. We report the results in VatLM[39] and our reproduced baselines(marked with *). We use a smaller and shallower 6-layer Transformer decoder due to the constraint of computation budgets and compare two settings of uni-modal data: with and without extra uni-modal data, as shown in Table 3. If without extra uni-modal data, we use videos and texts in LRS2 or LRS3 in an unpaired way to train unsupervised parts. Otherwise, we use uni-modal data from two different datasets, e.g. videos of Vox2-en-433h and texts of TEDLIUM-v3. Meanwhile, we leverage labeled data for Seq2Seq training. We set α to 0.001 or 0.01 in all experiments and N/A denotes that the corresponding unsupervised model fails to converge.

Main Conclusions. UniLip could achieve better efficacy of annotated data and also effectively incorporate extra uni-modal data into the Seq2Seq framework. When no extra uni-modal data is involved, UniLip boosts performances by at most relatively 2.7% as shown in the third row of the full set of LRS3. When extra uni-modal data is involved, a maximum relative improvement of 4.2% is observed in the last row of the 30h subset of LRS3.

Table 3: Results of Semi-supervised VSR on LRS3 and LRS2

Labeled Hours/h	Uni-modal Videos	Uni-modal Texts	Test WER/%(↓) (Base)	Test WER/%(↓) (Large)
		LRS2		
224	-		30.6[39]	24.3[39]
			32.0*	28.1*
	LRS2	LRS2	31.2(0.8↓)	27.8(0.3↓)
	Vox2-en-433h	TEDLIUM	31.0(1.0↓)	27.7(0.4↓)
		LRS3		
30	-		42.6[39]	31.6[39]
			42.0*	35.5*
	LRS2	TEDLIUM	41.1(0.9↓)	34.0(1.5↓)
433	-		34.2[39]	28.4[39]
			36.6*	32.6*
	LRS3	LRS3	35.4(1.2↓)	31.7(0.9↓)
	Vox2-en-433h	TEDLIUM	N/A	31.5(1.1↓)
	LRS2	TEDLIUM	36.2(0.4↓)	N/A

5 More Analysis

5.1 Model Ablation

In Table 4, With only either the local or the global feature, the model fails to converge, demonstrating the necessity of our multi-grained feature fusion module. For the discrimina-

tor, the ngram-based discrimination network(GramNet) outperforms the naive one by 0.8%. Combined with the conclusions in Section 4.3, it is shown that GramNet could not only incorporate texts from various sources but also bring better performances.

Table 4: Ablation of Model Architectures

Generator	Discriminator	Test WER
Local Feature	Naive	N/A
Global Feature		N/A
Feature Fusion	Naive	52.7%
	GramNet	**51.9%**

5.2 Visualization

We perform phoneme-level decoding on the LRS3 test set and retrieve the corresponding input lip images of each phoneme according to the decoding results. The results are shown in Figure 3. We found that UniLip successfully maps different phonemes to different lip shapes, such as "CH" and "M". Surprisingly, we also notice that UniLip could associate phonemes of similar pronunciations with similar lip shapes, even though they're not homophones. For example, "AE" and "AY" both require an open mouth, but "AY" requires a wider stretch of the mouth horizontally, while "AE" requires a greater stretch of the mouth vertically. These findings demonstrate UniLip's capacity for learning fine-grained semantic mappings between textual and visual modalities in lip reading.

Figure 3: Visualization of Phonemes

5.3 Unsupervised Results on Combined Datasets

Since UniLip does not require any video-text pairs, we combine videos of LRS3 and Vox2-en-433h and texts of TEDLIUM and Cantab for unsupervised training, yielding a test WER of 57.4% on LRS3 and 60.8% on LRS2. This is likely due to the difficulty of handling multiple datasets with such a simple model described in Section 3. Better techniques will be explored to extend UniLip to combined datasets.

6 Conclusion

We propose a novel way to perform lip reading by incorporating uni-modal data, namely UniLip. With uni-modal data, we achieve unsupervised lip reading for the first time. When also provided labeled data, UniLip is capable of incorporating extra uni-modal data and further improves the efficacy of labeled data. The results show its great potential for real-world applications.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (No. 62276247, 62076250).

References

[1] 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.

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

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2143–2147. IEEE, 2020.

[4] Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. *arXiv preprint arXiv:2303.00628*, 2023.

[5] Alexei Baevski, Henry 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.

[6] Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Unsupervised speech recognition. *Advances in Neural Information Processing Systems*, 34:27826–27839, 2021.

[7] Kuan-Yu Chen, Che-Ping Tsai, Da-Rong Liu, Hung-Yi Lee, and Lin-shan Lee. Completely unsupervised phoneme recognition by a generative adversarial network harmonized with iteratively refined hidden markov models. In *Interspeech*, pages 1856–1860, 2019.

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

[9] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103, 2017.

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

[11] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250, 2021.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, 2014.

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in Neural Information Processing Systems*, 30:5767–5777, 2017.

[14] Alexandros Haliassos, Pingchuan Ma, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Jointly learning visual and auditory speech representations from raw data. *arXiv preprint arXiv:2212.06246*, 2022.

[15] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Y. Estève. Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *International Conference on Speech and Computer*, pages 198–208, 2018.

[16] 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(TASLP)*, 29:3451–3460, 2021.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015.

[18] Alexander H Liu, Hung-yi Lee, and Lin-shan Lee. Adversarial training of end-to-end speech recognition using a criticizing language model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6176–6180, 2019.

[19] Alexander H Liu, Wei-Ning Hsu, Michael Auli, and Alexei Baevski. Towards end-to-end unsupervised speech recognition. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 221–228, 2023.

[20] Da-Rong Liu, Kuan-Yu Chen, Hung-yi Lee, and Lin-shan Lee. Completely unsupervised phoneme recognition by adversarially learning mapping relationships from audio embeddings. *arXiv preprint arXiv:1804.00316*, 2018.

[21] 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.

[22] 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.

[23] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 905–912, 2019.

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

[25] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210, 2015.

[26] Jongseok Park, Kyubyong & Kim. g2pe. https://github.com/Kyubyong/g2p, 2019.

[27] 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.

[28] 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.

[29] Brendan Shillingford, Yannis Assael, Matthew W Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, et al. Large-scale visual speech recognition. *Interspeech*, pages 4135–4139, 2019.

[30] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6447–6456, 2017.

[31] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with lstms for lipreading. *Interspeech*, pages 3652–3656, 2017.

[32] 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*, 30:5998–6008, 2017.

[33] Will Williams, Niranjani Prasad, David Mrva, Tom Ash, and Tony Robinson. Scaling recurrent neural network language models. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5391–5395, 2015.

[34] Bo Xu, Cheng Lu, Yandong Guo, and Jacob Wang. Discriminative multi-modality speech recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14433–14442, 2020.

[35] 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 *IEEE international conference on automatic face & gesture recognition (FG)*, pages 1–8, 2019.

[36] 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.

[37] Zi-Qiang Zhang, Jie Zhang, Jian-Shu Zhang, Ming-Hui Wu, Xin Fang, and Li-Rong Dai. Learning contextually fused audio-visual representations for audio-visual speech recognition. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1346–1350. IEEE, 2022.

[38] Ya Zhao, Rui Xu, and Mingli Song. A cascade sequence-to-sequence model for chinese mandarin lip reading. *Proceedings of the ACM Multimedia Asia*, pages 32:1–32:6, 2019.

[39] 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*, pages 1–11, 2023. doi: 10.1109/TMM.2023.3275873.