# Supplemental Materials: 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

We report more detailed results of unsupervised and semi-supervised VSR in Section 1 and Section 2, respectively. Finally, we provide some detailed discussions in Section 3.

## 1 Unsupervised VSR

In this section, we first provide more implementation details in Section 1.1, which are applied for all the experiments of unsupervised VSR including the ones in the main submission, and then report our results of zero-shot unsupervised VSR in Section 1.2.

### 1.1 Implementation Details

For the generator $\mathcal{G}$, the scaling factor of BatchNorm is set to 30 as [3]. The feature dimension of the whole multi-head self-attention module is the same as the dimension of the input visual features, i.e. 1024 for the Large model and 768 for the Base model, while the dimension of each attention head is 64. In our discriminator $\mathcal{D}$, its final causal convolutional layer's output is mean averaged along the temporal dimension and squeezed into the range of $[0,1]$ by a softmax function to compute the binary CE loss. The (input channel, output channel) of the three causal convolution layers are $(V,384)$, $(384,384)$, and $(384,1)$, respectively, where $V$ denotes the vocabulary size of phonemes. We use a set of 44 phonemes in our case.

All of our models in different settings are trained on four GeForce RTX 3090 GPUs. The stride of $\mathcal{G}$'s convolution is set to 2 for training and 1 for inference to encourage $\mathcal{G}$ to produce longer outputs[3]. We leverage the popular HLG decoding to transform $\mathcal{G}$'s phoneme-level outputs into word-level sentences, which involves $\mathcal{G}$'s outputs (H), a pronunciation lexicon (L), and word-level n-gram language models (G). We create the pronunciation lexicon with the G2P tool [4] on the text and train n-gram language models separately with the uni-modal text corpus introduced during adversarial training. $\mathcal{D}$ is not used in inference.

### 1.2 Zero-Shot Unsupervised VSR

**Experimental Settings.** In this section, we present zero-shot unsupervised results where both training videos and texts come from other dataset which is different from the test data. Since the training videos of LRS3 have been utilized during the feature extractor's pre-training[9], we couldn't create a "zero-shot" setting for LRS3 and only provide the zero-shot results on LRS2. For a detailed comparison, we also provide results obtained with the videos from LRS2 but trained with the uni-modal texts from other datasets, i.e. TEDLIUM and Cantab.

**Conclusions.** As shown in Table 1, UniLip shows strong performances even in the harsh condition where both training videos and texts are different from the test set. In the four experiments, most of them obtain similar performances with only a slight degradation compared to the counterpart where training videos are from LRS2. Surprisingly, when training texts are from TEDLIUM, the result using Vox2-en-433h as training videos even outperforms using LRS2 by 1.6%. These results show that UniLip is capable of learning the correct visual-textual mapping even using both uni-modal videos and uni-modal texts for lip reading in a zero-shot setting.

Table 1: Results of Zero-shot Unsupervised VSR on LRS2

| Training Video | Training Text | Test WER/%(↓) |
|---|---|---|
| LRS2 | TEDLIUM | 57.3 |
| LRS3 | | 56.6(0.7↓) |
| Vox2-en-433h | | 58.9(1.6↑) |
| LRS2 | Cantab | 58.9 |
| LRS3 | | 59.9(1.0↑) |
| Vox2-en-433h | | 59.7(0.8↑) |

## 2 Semi-Supervised VSR

In this section, we report our detailed experimental settings, together with more evaluation of UniLip's ability to co-work with existing supervised frameworks.

### 2.1 Detailed Experimental Settings

For each lip reading dataset, i.e. LRS2 or LRS3, we compare the performances of two types of models: the supervised baseline model and the semi-supervised models. The supervised baselines in this setting are marked with *, which are reproduced following the fine-tuning pipeline in [9] with a smaller and shallower 6-layer Transformer decoder due to the constraint of computation budgets. The decoder has a hidden dimension of 256 and feedforward dimentsion of 2048 for both the Base and Large models. Compared with the supervised baseline, our semi-supervised version firstly trains unsupervised models $\mathcal{G}$ and $\mathcal{D}$ with uni-modal data, and then performs supervised fine-tuning with video-text pairs involving an extra loss $L_{GAN}$ from $\mathcal{D}$, as shown in equation (6) of the main submission. When computing $L_{GAN}$, we first feed the encoder's outputs to $\mathcal{G}$ to generate phonemes, and then evaluate the realness of the phonemes with $\mathcal{D}$.

## 2.2 Incorporation with Other Supervised Approaches

**Motivation.** Besides the evaluation of UniLip to co-work with Seq2Seq[7], we evaluate UniLip's ability to be incorporated into other supervised training frameworks. We choose three popular supervised frameworks in lip reading for evaluation: Seq2Seq, CTC, and the hybrid of both[1, 8].

**Detailed Experimental Settings.** We report the performances of the Base model under both supervised and semi-supervised settings on LRS2. The results are shown in Table 2. The GAN model used during semi-supervised training is obtained with the uni-modal videos of Vox2-en-433h and uni-modal texts of TEDLIUM. The first column indicates the corresponding supervised training approach, where the hybrid of Seq2Seq and CTC are shortened as "Seq2Seq&CTC". The second and third columns are the test WERs obtained without and with the incorporation of UniLip, respectively. With the incorporation of UniLip, the model receives an extra loss $L_{GAN}$ besides the supervised loss, e.g. CE loss or CTC loss. The four column shows the relative Word Error Rate Reduction (WERR) introduced by UniLip. The down arrows mean that lower is better.

Table 2: Evaluation of Supervised Approaches on LRS2

| **Supervised Approach** | **Baseline/%(↓)** | **Ours/%(↓)** | **Relative WERR/%(↓)** |
|---|---|---|---|
| Seq2Seq | 32.0 | 31.0 | 3.1 |
| CTC | 36.2 | 35.8 | 1.1 |
| Seq2Seq&CTC | 29.9 | 29.4 | 1.7 |

**Conclusions.** As shown in Table 2, the introduction of UniLip steadily improves the performance with all three approaches, showing our method's ability to co-work with different supervised frameworks.

# 3 Discussions

**Training&Inference Overheads.** Besides the performance potential, UniLip also shows advantages on training and inference overheads, including model size, training strategies, training duration, and inference speed. Typically, sentence-level supervised training in lip reading often requires heavyweight models, complex training strategies, and a long training time. The inference speed of attention-based Seq2Seq models is further limited due to the multi-pass nature of beam search. While our model is super lightweight, which only involves a half number of parameters of ResNet18[2], it could also be directly trained in an end-to-end way and doesn't require any complex training strategies such as curriculum learning[5]. With four GeForce RTX 3090 GPUs, it only takes five hours to fully converge. In addition, our model is super fast to infer. During inference, inferring all the samples on the test set of LRS3 only takes about 1.5 minutes, which is much faster than attention-based beam search and CTC prefix beam search. The whole lightweight design makes it suitable for real-world applications where the computation budget is limited.

**Loss Weights.** Besides the classical adversarial objectives, we introduce four task-oriented auxiliary objective targets: gradient penalty $\mathcal{L}_{gp}$, smoothness penalty $\mathcal{L}_{sp}$, phoneme diversity loss $\mathcal{L}_{pd}$, and auxiliary prediction loss $\mathcal{L}_{aux}$. With four task-oriented auxiliary losses, a natural concern is how to adjust their weights to make the whole objective effectively guide the training process. Our rule of thumb is using fixed weights of $\mathcal{G}$-related losses($\mathcal{L}_{sp}$, $\mathcal{L}_{pd}$, and $\mathcal{L}_{aux}$) and only adjusting the weights of $\mathcal{L}_{gp}$ with respect to different text datasets. In our experiments, we use weights 0.1, 3.0, 0.5 for $\mathcal{G}$-related losses ($\mathcal{L}_{sp}$, $\mathcal{L}_{pd}$, and $\mathcal{L}_{aux}$) in all experiments and only adjust the weights of $\mathcal{L}_{gp}$ when using different text datasets. $\mathcal{L}_{gp}$ slows down $\mathcal{D}$'s pace of fitting real samples by constraining $\mathcal{D}$'s gradient norm on mixtures of real and fake samples, which leads to its close influence on $\mathcal{D}$'s ability to fit different text corpus. Generally speaking, the smaller $\mathcal{L}_{gp}$ is, the easier $\mathcal{D}$ fits the uni-modal texts. For example, the huge domain gap between audiobooks and TED talks makes the texts of LibriSpeech hard to be fitted into our task, so we should use a relatively smaller value of $\mathcal{L}_{gp}$ compared with its value on other text datasets.

**Length Choice of N-gram Clips.** The length range of n-gram clips is set to $[20,25)$, instead of a single fixed-length, in our experiments. In this case, every n-gram clip contains the phonemes of about 5 words, so learning linguistic priors from uni-modal texts could be viewed as 5-gram word-level language modeling. In this range, every n-gram clip contains roughly the phonemes of 5 or 6 words, so learning linguistic priors from uni-modal texts could be viewed as 5-gram or 6-gram word-level language modeling, which we found appropriate for our task. On the other hand, the lengths of n-gram clips vary slightly with each other with a maximum of 4. This helps ease the difficulty of $\mathcal{D}$'s learning because $\mathcal{D}$ doesn't have to fit the inputs beyond this range. The reason why we didn't use a fixed value for the lengths of n-gram clips is that picking a sub-sequence of consecutive phonemes with a fixed length can not always be satisfied and often requires excessive clamping on the original sequence.

**Choice of Tokenization Unit.** In sequence prediction tasks, such as lip reading and Automatic Speech Recognition (ASR), the most popular choices of tokenization unit include chars, sub-word invariants, e.g. Byte-Pair Encoding(BPE)[6], and phonemes. Among them, chars are combinations of characters, and sub-words are obtained by statistically decomposing words into smaller units. Neither of them has a relation to pronunciation patterns. On the other hand, each phoneme could represent a distinct auditory vocal pattern, making it a candidate for our lip reading task. Even though the mapping between phonemes and visemes is not one-to-one due to the existence of homophones, we empirically found that it is not a severe problem in our task and adopt phoneme as the tokenization unit in all unsupervised experiments. We've also experimented with char or BPE, but neither of them successfully converge in our preliminary experiments.

## 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] 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.

[3] 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.

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

[5] 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.

[6] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[7] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27:3104–3112, 2014.

[8] 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.

[9] 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.