

Audio-Visual Representation Learning via Knowledge Distillation from Speech Foundation Models

Jing-Xuan Zhang^{a,*}, Genshun Wan^{b,c}, Jianqing Gao^c, Zhen-Hua Ling^b

^a*School of Computer Science, Shaanxi Normal University, No. 620, West Chang'an Avenue, Xi'an, 710119, Shaanxi, P.R.China*

^b*University of Science and Technology of China, No.96, JinZhai Road, Hefei, 230026, Anhui, P.R.China*

^c*iFLYTEK Research, iFLYTEK Co. Ltd., No. 666, Wangjiang West Road, Hefei, 230088, Anhui, P.R.China*

Abstract

Audio-visual representation learning is crucial for advancing multimodal speech processing tasks, such as lipreading and audio-visual speech recognition. Recently, speech foundation models (SFMs) have shown remarkable generalization capabilities across various speech-related tasks. Building on this progress, we propose an audio-visual representation learning model that leverages cross-modal knowledge distillation from SFMs. In our method, SFMs serve as teachers, from which multi-layer hidden representations are extracted using clean audio inputs. We also introduce a multi-teacher ensemble method to distill the student, which receives audio-visual data as inputs. A novel representational knowledge distillation loss is employed to train the student during pretraining, which is also applied during finetuning to further enhance the performance on downstream tasks. Our experiments utilized both a self-supervised SFM, WavLM, and a supervised SFM, iFLYTEK-speech. The results demonstrated that our proposed method achieved superior or at least comparable performance to previous state-of-the-art baselines across automatic speech recognition, visual speech recognition, and audio-visual speech recognition tasks. Additionally, comprehensive ablation studies and the visualization of learned representations were conducted to evaluate the effectiveness of our proposed method.

Keywords: representation learning, knowledge distillation, speech foundation model,

*Corresponding author

Email addresses: jxzhanggg@snnu.edu.cn (Jing-Xuan Zhang), gswan@iflytek.com (Genshun Wan), jqgao@iflytek.com (Jianqing Gao), zhling@ustc.edu.cn (Zhen-Hua Ling)

1. Introduction

Speech processing techniques have advanced significantly in recent years, with speech foundation models (SFMs) emerging as a prominent development. Trained on vast speech datasets, these large-scale models such as WavLM [1] and Whisper [2], excel in numerous tasks. However, environmental noise remains a critical challenge, potentially compromising the performance of audio-based models. This limitation has spurred interest in audio-visual speech processing. Visual speech recognition (VSR), as known as lipreading [3], focuses on decoding speech content from visual cues alone, while audio-visual speech recognition (AVSR) [4] integrates both auditory and visual inputs. These multi-modal approaches contribute to more robust speech recognition systems, showing promise for enhancing communication in noisy environments and improving hearing aids.

Audio-visual representation learning underpins contemporary advancements in audio-visual speech processing. In contrast to early work using handcrafted features, deep neural networks now extract useful representations by leveraging large-scale datasets. The training objectives can be broadly categorized into supervised learning and self-supervised learning. While supervised learning requires labeled data, self-supervised learning utilizes unlabeled data to extract general representations during pretraining phase [5]. Subsequently, the model can be adapted to a wide-range of downstream tasks with minimal labeled data. Various pretext tasks have been introduced for audio-visual self-supervised learning (AV-SSL), including mask prediction [6, 7], contrastive learning [8, 9], and cross-modal distillation [10, 11]. Despite their success, AV-SSL models face several limitations. First, collecting parallel audio and corresponding lip video data is more challenging than acquiring audio data alone. Consequently, current AV-SSL models are optimized with audio-visual datasets that are substantially smaller than those used to train SFMs. Second, AV-SSL methods often rely on iterative training [6, 12] or an online teacher module [10, 13] to generate the learning targets for pretext tasks, which can introduce additional computational overhead or potentially

lead to unstable training.

To address the aforementioned limitations and inspired by the success of speech foundation models, we propose transferring the knowledge from SFMs to a joint audio-visual representation learning model through cross-modal knowledge distillation. In our method, the SFM serves as the teacher, with its multi-layer hidden representations extracted and aggregated to distill knowledge into the student model. An ensemble of multiple teachers can further enhance the student’s generalization capability. The student is exposed to noisy audio-visual data to improve its robustness to noise. We introduce a novel representational knowledge distillation loss that combines both feature regression and KL divergence loss using soft cluster labels. Furthermore, the pretrained model is finetuned with an auxiliary knowledge distillation loss to enhance performance on downstream tasks, including Automatic Speech Recognition (ASR), VSR, and AVSR. Our work is closely related to LiRA [14], which learns visual speech representations through self-supervision of audio representation extracted by PASE+ [15]. The key distinction between our approach and LiRA lies in our focus on exploiting knowledge learned from large-scale audio data. Consequently, we utilize SFMs trained on substantially larger audio datasets compared to PASE+. Moreover, our method jointly learns audio-visual representations within a single model.

In our experiments, we evaluated two speech foundation models: a self-supervised model, WavLM [1], and a supervised model, iFLYTEK-speech. Our method demonstrated superior or at least comparable performance to baselines across ASR, VSR, and AVSR tasks, particularly in scenarios with low-resource labeled data. Additionally, we compared our method to baseline models using noisy audio-visual test sets in the AVSR task. Comprehensive ablation studies were conducted to validate the effectiveness of our approach, focusing on key components such as multi-teacher ensemble, our proposed knowledge distillation loss, and the joint modeling of audio-visual speech in a single student model. Moreover, we performed a visualization analysis of the representations derived from both audio and visual inputs. Our findings revealed a notable alignment between the audio and visual representations, providing further support for the recently observed trend of representational convergence across modalities [12].

2. Related works

2.1. Knowledge distillation

Knowledge distillation (KD) has emerged as a crucial technique in model compression and transfer learning. Cross-modal knowledge distillation (CMKD), an extension of KD, focuses on transferring knowledge between different modalities [16]. Many studies have explored cross-modal distillation for audio-visual speech data, leveraging the inherent synergy and correspondence between audio signals and lip movements. Afouras et al. [17] train a lipreading model by distillation from an ASR model, combining connectionist temporal classification with frame-wise cross entropy loss. LIBS [18] enhances lipreading by distilling multi-scale knowledge from speech recognizers. Another line of research employs CMKD within a self-supervised learning framework. RAVEn [10] utilizes an asymmetric distillation scheme for audio-visual representation learning with modality-specific student. Both AV-data2vec [11] and our previously proposed AV2vec [13] adopt a multimodal teacher for distilling a student that accepts audio, visual, and audio-visual inputs.

Our work is also related to research exploring ensemble-based approaches in knowledge distillation. For instance, Huang et al. [19] propose an ensemble knowledge distillation framework to transfer knowledge from large self-supervised speech models. Similarly, recent studies [20, 21] develop a dynamic ensemble teacher-student architecture for fake audio detection. Building upon these works, our approach leverages aggregated representations from multiple teacher networks across different layers to enhance knowledge distillation.

2.2. Visual and audio-visual speech recognition

Nowadays, end-to-end frameworks have become the predominant approach for visual and audio-visual speech recognition [4, 22], where a single network consumes raw lip video pixels and audio features to generate text outputs. The end-to-end paradigm relies on sequence-to-sequence model architectures and their corresponding loss function, including encoder-decoder with attention (ED) [4], connectionist temporal classification (CTC) [23], and neural transducer [24]. Under this framework, researchers

elaborate on designing neural network architecture for extracting visual representations, such as temporal convolutional network (TCN) [25], Transformer [26], visual Transformer pooling [27], and Conformer [22]. Concurrently, another line of research explores effective audio-visual fusion strategy for AVSR, such as V-CAFE [28], and unified cross-modal attention [29].

Recently, there has been a growing trend in leveraging large-scale ASR models for visual and audio-visual speech recognition. Djilali et al. [30] pioneered the re-programming of ASR models for VSR by introducing a prior network that transforms visual representations into audio representations. Building on a similar concept, Prajwal et al. [31] proposed a VSR approach with a simplified training objective. In the AVSR domain, Simic et al. [32] enhanced the Whisper model, originally designed for ASR, by incorporating a fusion module for audio and visual inputs. Similarly, Whisper-flamingo [33] extended the Whisper model’s capability through additional cross-attention components that process visual representations for AVSR tasks. More recently, the integration of large language models with pretrained audio and visual encoders has achieved state-of-the-art performance in AVSR tasks [34]. Our work distinguishes itself by learning general audio-visual representations applicable beyond conventional lipreading and AVSR tasks. Moreover, our approach can be seamlessly integrated with these existing techniques to further enhance visual and audio-visual speech recognition performance in future work.

2.3. *Speech foundation models*

Foundation models, a term first coined by Bommasani et al. [35], are defined as large models trained on broad data that adapts to a wide range of downstream tasks. Various speech foundation models have emerged, broadly categorized into two types, i.e. supervised learning-based and self-supervised learning (SSL)-based models. Among SSL-based speech foundation models, HuBERT [36] utilizes k-means clustering to generate pseudo-class labels for model training using a mask prediction loss, with iterative refinement through subsequent clustering and mask prediction steps. WavLM [36] improves HuBERT by using a more diverse pretraining dataset and performs speech denoising modeling during pretraining. On the other hand, Whisper

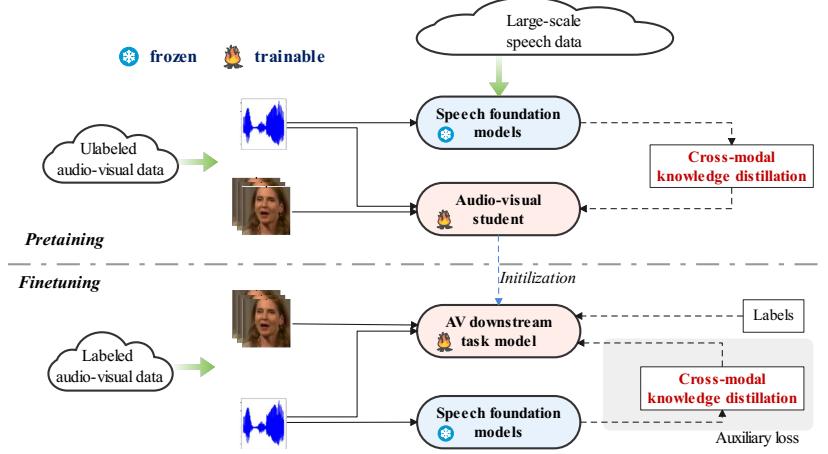


Figure 1: Illustration of overall scheme of our proposed method during the pretraining and finetuning phases.

family [2] exemplifies supervised learning-based speech foundation model. Trained on large-scale multilingual and multitask labeled data, it demonstrates robust performance in ASR and speech translation tasks. In our experiments, we evaluated both a self-supervised speech foundation model, WavLM, and a supervised one, iFLYTEK-speech.

3. Proposed method

In our method, speech foundation models (SFMs) serve as teachers, whose hidden representations are extracted as learning targets for the student model. During pre-training phase, the audio-visual student model is optimized using our proposed knowledge distillation loss function on unlabeled data. During finetuning phase, KD loss is retained in conjunction with the primary training objective for downstream tasks. Figure 1 illustrates the overall scheme of our proposed method.

3.1. Student

The structure of the student is depicted in the left portion of Figure 2. This model processes both audio and visual inputs concurrently. For the audio stream, Mel filterbank features are used. The visual stream incorporates the entire face images, departing

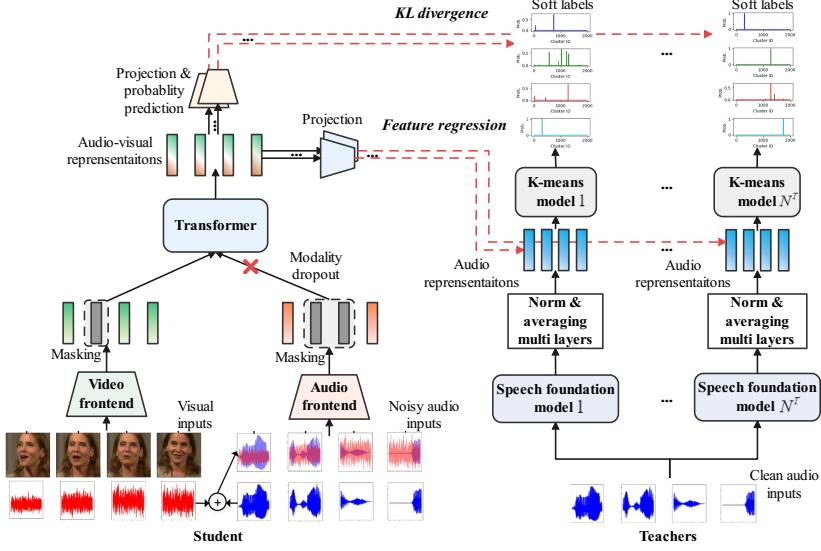


Figure 2: Illustration of our proposed method during the pretraining phase.

from the conventional approach of using only the lip region-of-interest (ROI). Previous research [37], as well as our own studies [38, 13] have demonstrated that entire faces contain more detailed information than lips alone, thereby enhancing visual representation learning. To enhance noise robustness, the audio input of the student model, A , is derived by adding random noise to the clean audio with a probability p_{noise} . Both audio and visual inputs are processed through modality-specific frontend to generate audio local features $\mathbf{F}^a \in \mathbb{R}^{T \times D_f}$ and visual local features $\mathbf{F}^v \in \mathbb{R}^{T \times D_f}$ as

$$\mathbf{F}^a = \text{AF}(A), \mathbf{F}^v = \text{VF}(V), \quad (1)$$

where $\text{AF}(\cdot)$ and $\text{VF}(\cdot)$ represent audio and visual frontend respectively. Subsequently, features from each modality undergo independent corruption using a span-based mask. Let M_a denote the set of masked indices for audio. The corrupted audio feature is derived as $\tilde{\mathbf{F}}^a = r(\mathbf{F}^a, M_a)$, where \mathbf{F}_i^a is replaced with an embedding e^a if $i \in M_a$, with i represents frame index. An analogous strategy is applied to corrupt visual features, utilizing different mask indices M_v and embedding e^v . Then modality dropout [6] is implemented to train the student in the absence of either audio or video stream randomly. Specifically, with a probability p_m , both modalities are used. The probability of select-

ing audio is p_a when only one modality is used, and features of the omitted modality are set to zero. The audio-visual features are then concatenated along the channel dimension as $\mathbf{F}^{av} = \text{concat}(\tilde{\mathbf{F}}^a, \tilde{\mathbf{F}}^v) \in \mathbb{R}^{T \times 2D_f}$. \mathbf{F}^{av} is subsequently processed through a stack of Transformer blocks to generate contextualized audio-visual representations as

$$\mathbf{O} = \text{Transformer}(\mathbf{F}^{av}), \quad (2)$$

where $\mathbf{O} = [\mathbf{o}_1, \dots, \mathbf{o}_T] \in \mathbb{R}^{T \times D_o}$ and D_o represents the channel dimension.

3.2. Teacher

Speech foundation models (SFMs) serve as the teachers in our proposed method. For distilling student, latent representations from the last k layers of the SFM are extracted from clean audio as

$$\{\mathbf{H}^{L-k}, \dots, \mathbf{H}^L\} = \text{SFM}(\mathbf{A}^c), \quad (3)$$

where L denotes the total number of layers in the SFM, $\mathbf{H}^i \in \mathbb{R}^{T \times D_h}$, and D_h represents the channel dimension of the SFM. Multi-layer representations are first instance normalized and then averaged to obtain the final teacher representations $\mathbf{H}^T = [\mathbf{h}_1^T, \dots, \mathbf{h}_T^T] \in \mathbb{R}^{T \times D_h}$ as

$$\mathbf{H}^T = \frac{1}{k} \sum_{i=L-k}^L \text{IN}(\mathbf{H}^i), \quad (4)$$

where $\text{IN}(\cdot)$ represents instance normalization.

Research has shown that self-supervised SFMs encode different aspects of knowledge across various layers [39]. Consequently, averaging representations from multiple layers can enrich the target knowledge for more effective student distillation. Instance normalization is applied prior to averaging to prevent any single layer from dominating the combined representations. Our experimental results in Section 4.3.5 confirmed the efficacy of multi-layer averaging for the self-supervised teacher. Interestingly, for the supervised teacher, using representations solely from the last layer yielded optimal results. This discrepancy likely reflects differences in the hierarchical encoding patterns of information, which arise from the distinct training objectives of self-supervised and supervised SFMs.

3.3. Pretraining

3.3.1. Distillation loss function

During the pretraining phase, the student is optimized using unlabeled audio-visual speech data through cross-modal knowledge distillation from SFM’s representations. The total knowledge distillation loss is a combination of a feature regression loss and a KL divergence loss, as illustrated in Figure 2. For feature regression loss, a linear projection layer is employed to the student’s outputs to predict the teacher’s representations as

$$L_{\text{reg}} = \frac{1}{T} \sum_{t=1}^T \|\mathbf{W}\mathbf{o}_t - \mathbf{h}_t^\top\|_2^2, \quad (5)$$

where \mathbf{W} are trainable parameters.

In AV-HuBERT [6], latent representations are clustered using k-means algorithm, and then quantized by selecting the nearest cluster centroid to obtain pseudo-labels. However, this approach loses the hidden knowledge about the distances from current representation frame to other non-nearest cluster centroids. To extract richer knowledge, we propose constructing soft labels that capture the relationships between the current frame and all cluster centroids. Let the cluster centroids be $\mathbf{c}_1 \dots \mathbf{c}_N$, where N represents the total number of clusters. The i -th value of soft label $\mathbf{l}_t^{\text{soft}} \in \mathbb{R}^N$ for the t -th frame is computed as

$$\mathbf{l}_t^{\text{soft}}(i) = \frac{\exp(-\|\mathbf{h}_t^\top - \mathbf{c}_i\|_2^2 / (\tau' I))}{\sum_{n=1}^N \exp(-\|\mathbf{h}_t^\top - \mathbf{c}_n\|_2^2 / (\tau' I))}, \quad (6)$$

where I represents the inertia computed by k-means clustering, i.e., the sum of squared distances of samples to their closest cluster centriod. The hyper parameter τ' controls the smoothness of soft labels. The student is then trained using a KL divergence loss as

$$L_{\text{kld}} = \frac{1}{T} \sum_{t=1}^T \text{KLD}(\text{softmax}(\cos(\mathbf{U}\mathbf{o}_t, \mathbf{E})/\tau), \mathbf{l}_t^{\text{soft}}), \quad (7)$$

where \mathbf{U} and \mathbf{E} are trainable parameters, $\text{KLD}(\cdot)$ represents the KL divergence loss, and $\cos(\cdot)$ represents the cosine function.

The total knowledge distillation loss function incorporates both L_{reg} and L_{kld} , framing the problem as a multitask learning scenario. To balance these two losses, we employ a multitask learning method handles with multiple training objective functions,

specifically Aligned-MTL-UB [40]. In this approach, the gradients of each loss with respect to the student’s outputs \mathbf{O} are first computed. These gradients are then used to determine the weighting parameters for gradient aggregation. Aligned-MTL aligns the orthogonal components of the linear system of gradients, thereby enhancing the training stability in multitask learning. Unlike previous studies [6, 13, 11] that only compute loss only over masked regions, our method considers both masked and unmasked regions. This method facilitates more effective knowledge distillation, as demonstrated in our experiments in Section 4.3.4.

3.3.2. Multi-teacher ensemble

Our method further proposes employing an ensemble of multiple teachers to enhance the student’s generalization capability. Specifically, representations from multiple N^T teachers are extracted. The student uses separate projection heads $\{\mathbf{W}_1, \mathbf{U}_1, \mathbf{E}_1\}, \dots, \{\mathbf{W}_{N^T}, \mathbf{U}_{N^T}, \mathbf{E}_{N^T}\}$ to compute the knowledge distillation loss for each teacher. These losses are then combined to pretrain the student using the Aligned-MTL algorithm.

3.4. Finetuning

During finetuning, a randomly initialized Transformer-based decoder is employed, with the student model serving as the encoder. The encoder is frozen for the first n_{freeze} updates, after which the encoder-decoder model is optimized jointly using the labeled data for the remaining updates. In addition to the primary training objective of the task, our proposed knowledge distillation loss is still retained as an auxiliary training objective. The knowledge distillation loss is first scaled by a factor of λ then summed with the primary loss to produce the total loss for finetuning.

4. Experiments and results

4.1. Implementation details

¹ **Datasets.** For pretraining, LRS3 dataset² was utilized, which comprises approximately 433 hours of English audio-visual speech data. Additionally, we employed a combination dataset of LRS3 and English-only version of VoxCeleb2 [41], curated by Shi et al. [6], totaling about 1759 hours of audio-visual speech data. Our ablation studies were conducted exclusively on the LRS3 dataset. For finetuning, we adopted two subsets of LRS3: a 30-hour subset and the full 433-hour training set. A validation set of 1000 utterances was reserved, and the results were reported on LRS3 test set. We used the facial images provided in the original datasets without cropping to lip region-of-interest (ROI). The MUSAN dataset³ was employed for noise augmentation. Following the protocol outlined in [42], we sampled and added noise to the audio input of the student model during pretraining. Noise augmentation was also applied during finetuning for the AVSR task. To evaluate our proposed method in AVSR under noisy conditions, we constructed test sets with various noise types, including music, natural sounds, speech, and babble, at signal-to-noise ratios (SNRs) of -10, -5, 0, 5, and 10 dB, respectively.

Model configurations. In our experiments, we employed two speech foundation models: a self-supervised model, WavLM, and a supervised model, iFLYTEK-speech. WavLM [1], based on a Transformer backbone, utilizes a training objective of predicting pseudo-labels of simulated noisy speech in masked regions. We specifically used WavLM-Large⁴, which comprises 24 Transformer blocks, 317M parameters and was trained on 94,000 hours of speech data. iFLYTEK-speech, an in-house model developed by iFLYTEK company, is built on a Conformer backbone and trained on large-scale multilingual speech data. It is designed for various speech-to-text tasks, primarily speech recognition and speech translation. This model employs an encoder-decoder ar-

¹Our code will be available at <https://github.com/jxzhanggg/DistillAV>.

²https://mmai.io/datasets/lip_reading/

³<https://www.openslr.org/17/>

⁴<https://huggingface.co/microsoft/wavlm-large>

chitecture, of which only the encoder is utilized for extracting speech representations. iFLYTEK-speech contains 16 Conformer blocks, approximately 300M parameters and was trained on over 200,000 hours of proprietary speech data.

The student model architecture followed previous studies [6, 13], employing a feed-forward network for the audio frontend and a 3DCNN-ResNet18 network for the visual frontend. Two configurations were implemented for the Transformer encoder: a **Base** version with 12 layers and a **Large** version with 24 layers. Their embedding dimension/feed forward dimension/attention heads were 768/3072/12 and 1024/4096/16, respectively. The WavLM model generates representations at a frame rate of 50 fps, which is double that of the student model’s outputs. To address this discrepancy, we projected each student output frame onto two consecutive teacher frames when computing the knowledge distillation (KD) loss. In contrast, the frame rate of the iFLYTEK-speech encoder representations matches that of the student model. The value of k in Equation 4 was set to 8 for the WavLM-based teacher and 1 for the iFLYTEK-speech-based teacher. These values were optimized using our validation set, as detailed in Section 4.3.5.

During pretraining, 80% of audio features and 30% of video features were masked for the student. p_m and p_a were both set to 0.5. The temperature parameters τ and τ' were both set to 0.1. The inertia values of k-means model for WavLM and iFLYTEK-speech representations were 352.6 and 19.1 respectively. The number of k-means clusters, N , was set to 2000. During finetuning, a decoder Transformer with 6 layers was employed. Its embedding dimension/feed forward dimension/attention heads were 768/3072/4. We utilized subword units with a vocabulary size of 1000 as the targets. The value of n_{freeze} was optimized using our validation set. Our experiments revealed that freezing the encoder throughout the fine-tuning process yielded the best performance when using the 30-hour labeled dataset. Consequently, the auxiliary KD loss for regularizing the encoder was not employed in this scenario. When fine-tuning with the larger 433-hour labeled dataset, we incorporated the auxiliary KD loss with a λ value of 0.1.

Training costs and stability. Unlike AV-HuBERT [6] and u-HuBERT [12], our proposed method eliminates the need for iterative processing between feature cluster-

ing and mask prediction steps. The computational cost of our entire pretraining process is comparable to a single iteration of the AV-HuBERT model, resulting in significant reduction in training resources. 8 Tesla A100 GPUs were utilized in our experiments. For the 433-hour dataset, the pretraining process (400k steps) took approximately 5.3 days and 6.2 days for base and large versions, respectively. When extending to the 1759-hour dataset (800k steps), the pretraining time increased to 8.2 days and 10.3 days for base and large models. The finetuning process is notably efficient, completing within 16 hours.

Another category of AVSSL methods, such as our previously proposed AV2vec [13], requires an online-updated teacher module. These approaches typically necessitate techniques like feature normalization and careful selection of EMA hyperparameters to prevent target collapse [13, 10]. In contrast, our method leverages representations from well-established speech foundation models for knowledge distillation, effectively eliminating the risk of feature collapse. Our experimental results demonstrated robust training stability with no convergence issues.

4.2. Comparison with baselines

4.2.1. Results on clean test set

Table 1 and Table 2 summarize the results on the clean LRS3 test set for models finetuned with 30 and 433 hours of labeled data, respectively. When using low-resource labeled data (30 hours), our proposed method outperformed baselines across VSR, ASR, and AVSR tasks, as shown in Table 1. Notably, the superior performance in the VSR task demonstrates the effectiveness of cross-modal knowledge transfer from the audio to the visual modality. When using high-resource labeled data (433 hours), our method outperformed all self-supervised baselines in VSR task except Lip2vec pre-trained with 1759h data, which achieved comparable performance, as shown in Table 2. While the supervised baseline by Chang et al. [43] obtained the lowest word error rate (WER), it utilized a substantial dataset of approximately 100,000 hours of audio-visual speech data. For ASR and VSR tasks, our method outperformed or matched the baselines. However, AVSR results on the clean LRS3 test set have approached saturation, diminishing the significance of inter-model comparisons. Therefore, we proceeded to

Methods	Unlabeled Data	Label Data	Encoder Size	Criterion	VSR	ASR	AVSR
<i>Self-supervised model (Base)</i>							
AV-HuBERT [6]			103M	CE	51.8	4.9	4.7 ²
RAVen [10]			97M	CTC+CE	47.0	4.7	–
VATLM [7] ¹	433h	30h	103M	CE	48.0	–	3.6
AV-data2vec [11]			103M	CE	45.2	4.4	4.2
AV2vec-MLM [13]			103M	CE	39.4	5.6	5.4
AVFusion-Whisper [32]			87M	CE+KD	–	–	2.8
Lip2Vec [30]			240M	CTC+KD	49.5	–	–
Proposed			103M	CE	37.2	3.0	2.8
AV-HuBERT [6]			103M	CE	46.1	4.6	4.0 ²
RAVen [10]			97M	CTC+CE	40.2	3.8	–
VATLM [7] ¹	1759h	30h	103M	CE	42.6	–	3.4
AV-data2vec [11]			103M	CE	37.8	3.7	3.3
Lip2Vec [30]			240M	CTC+KD	40.6	–	–
Proposed			103M	CE	36.9	2.5	2.2
<i>Self-supervised model (Large)</i>							
AV-HuBERT [6]			325M	CE	44.8	4.5	4.2 ²
AV-data2vec [11]	433h	30h	325M	CE	40.5	3.7	3.4
AVFusion-Whisper [32]			257M	CE+KD	–	–	2.3
Lip2Vec [30]			727M	CTC+KD	55.4	–	–
Proposed			325M	CE	33.2	2.4	1.8
AV-HuBERT [6]			325M	CE	32.5	2.9	3.3 ²
RAVen [10]			671M	CTC+CE	33.1	2.6	–
VATLM [7] ¹	1759h	30h	325M	CE	31.6	–	2.7
AV-data2vec [11]			325M	CE	30.8	2.7	2.7
Lip2Vec [30]			727M	CTC+KD	31.2	–	–
Proposed			325M	CE	30.2	2.2	2.1

Table 1: Word error rate (WER) (%) results on LRS3 test set with 30-hour labeled data. ¹VATLM uses additional 3846h audio, 452h audio-text and 600M text data. ²Reproduction results from [11].

Methods	Unlabeled Data	Label Data	Encoder Size	Criterion	VSR	ASR	AVSR
<i>Supervised model</i>							
Afouras et al. [4]	–	1519h	–	CE	58.9	8.3	–
Makino et al. [24]	–	31kh	–	Transducer	33.6	4.8	4.5
Prajwal et al. [27]	–	2726h	–	CE	30.7	–	–
Serdyuk et al. [26]	–	90kh	–	Transducer	17.0	–	1.6
Chang et al. [43]	–	100kh	–	Transducer	12.8	–	0.9
<i>Self-supervised model (Base)</i>							
AV-HuBERT [6]			103M	CE	44.0	3.0	2.8 ²
RAVen [10]			97M	CTC+CE	39.1	2.2	–
AV-data2vec [11]	433h	433h	103M	CE	39.0	2.0	1.8
AV2vec-MLM [13]			103M	CE	34.4	2.7	2.5
Lip2Vec [30]			240M	CTC+KD	42.0	–	–
Proposed			103M	CE+KD	34.3	1.9	1.8
AV-HuBERT [6]			103M	CE	34.8	2.0	1.8 ²
RAVen [10]			97M	CTC+CE	33.1	1.9	–
VATLM [7] ¹	1759h	433h	103M	CE	34.2	–	1.7
AV-data2vec [11]			103M	CE	32.7	1.7	1.4
Lip2Vec [30]			240M	CTC+KD	34.1	–	–
Proposed			103M	CE+KD	31.4	1.9	1.7
<i>Self-supervised model (Large)</i>							
AV-HuBERT [6]			325M	CE	41.6	2.7	2.5 ²
AV-data2vec [11]	433h	433h	325M	CE	37.4	1.9	1.7
Lip2Vec [30]			727M	CTC+KD	50.1	–	–
Proposed			325M	CE+KD	31.5	1.8	1.6
AV-HuBERT [6]			325M	CE	28.6	1.3	1.4 ²
RAVen [10]			671M	CTC+CE	28.2	1.4	–
VATLM [7] ¹	1759h	433h	325M	CE	28.4	–	1.2
u-HuBERT [12] ¹			325M	CE	27.2	1.4	1.2
AV-data2vec [11]			325M	CE	28.5	1.4	1.3
Lip2Vec [30]			727M	CTC+KD	26.0	–	–
Proposed			325M	CE+KD	26.2	1.4	1.3

Table 2: WER (%) results on LRS3 test set with 433-hour labeled data. ¹VATLM uses additional 3846h audio, 452h audio-text and 600M text data, and u-HuBERT uses additional 452h audio data. ²Reproduction results from [11].

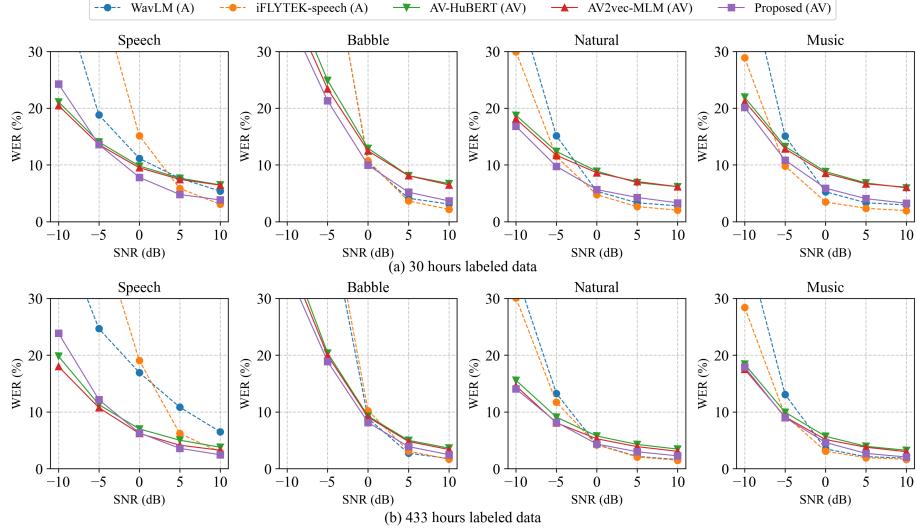


Figure 3: WER (%) results on the test sets corrupted by various types of noise. “A” and “AV” represent models using audio and audio-visual inputs respectively.

evaluate AVSR performance of our proposed method on more challenging noisy test sets.

4.2.2. AVSR results on noisy test sets

We constructed four types of test sets by mixing various noise types (speech, babble, natural, and music) with the clean audio from the LRS3 test set. The teacher models, WavLM and iFLYTEK-speech, were finetuned using audio-only inputs. Two AVSR baselines were adopted, including AV-HuBERT [6] and AV2vec-MLM [13]. AV-HuBERT was reproduced using our configuration with face images as inputs [13]. For pretraining, we utilized 433 hours of data from LRS3. Figure 3 summarizes the experimental results. In low signal-to-noise ratio (SNR) scenarios, models with audio-visual inputs significantly outperformed ASR models, particularly on test sets with speech or babble noise. However, as audio quality improved, the performance of speech foundation models increased dramatically, ultimately achieving the lowest WER under 10 dB SNR conditions across most noisy test sets.

When compared to AVSR baselines, our proposed method demonstrated superior

Labeled Data	Teacher		Base		Large	
	WavLM	iFLYTEK-speech	ASR	VSR	ASR	VSR
30h	✓	✗	3.8	38.9	2.9	33.3
	✗	✓	3.1	38.0	2.5	33.7
	✓	✓	3.0	37.2	2.4	33.2
433h	✓	✗	2.3	35.4	2.0	31.2
	✗	✓	1.9	34.7	1.9	32.9
	✓	✓	1.9	34.3	1.8	31.5

Table 3: WER (%) results on LRS3 test set of our proposed method using WavLM, iFLYTEK-speech-based, and an ensemble of the both as the teacher models.

performance under most conditions, except for low SNR test sets with speech noise. This finding suggests that our method may face challenges of over-reliance on the audio modality when processing data corrupted by heavy speech noise. Orthogonal to this work, such issue in AVSR was addressed by MSRL [44], which could be utilized to further improve AVSR performance in future studies.

4.3. Ablation studies

4.3.1. Multi-teacher ensemble

This section presents an ablation study of our proposed multi-teacher ensemble method. We used the same hyperparameters as those in previous section. 433-hour LRS3 dataset was used for pretraining. We compared the performance of using WavLM, iFLYTEK-Speech, and an ensemble of both as the teacher models. The experimental results are summarized in Table 3. For the ASR task, the student distilled from iFYTEK-speech-based teacher achieved a lower WER than the one distilled from WavLM. In the VSR task, the Base student model performed better when using iFLYTEK-speech as the teacher compared to WavLM. However, this trend was reversed for the Large student model. Notably, the ensemble of both teachers enhanced the generalization of knowledge distillation, resulting in more robust overall performance across both tasks.

4.3.2. Knowledge distillation loss

This section examines our proposed loss function for knowledge distillation. Experiments were conducted using only the iFLYTEK-speech teacher model and the Base version of the student. 433-hour LRS3 dataset was used for pretraining. We compared our proposed method with several variants for representational knowledge distillation. **Reg** represents utilizing only the feature regression loss. **CE** represents employing cross-entropy loss with discrete pseudo-labels, as in AV-HuBERT [6]. **KLD** represents using only the proposed KL divergence loss function with soft labels. **Reg+CE** represents combining cross-entropy and feature regression loss. Our proposed method is denoted by **Reg+KLD**. The results are presented in Table 4. Our proposed KLD loss using soft labels outperformed the CE loss using discrete labels, demonstrating its effectiveness in transferring the teacher’s hidden knowledge. When combined with feature regression loss, the student model achieved the lowest WER in both ASR and VSR tasks. During finetuning with 30 hours of labeled data, the encoder remained frozen, and the auxiliary KD loss was not applied. When finetuning with high-resource labeled data (433 hours), WER was further reduced for both ASR and VSR tasks, particularly ASR. These results demonstrated the effectiveness of our proposed KD loss during the both pretraining and finetuning phases.

4.3.3. Model inputs

This section compares models accepting different types of inputs, including an audio-only model, a visual-only model, and our proposed joint audio-visual model. Experiments were conducted using only the iFLYTEK-speech teacher model and the Base version of the student, with the knowledge distillation loss omitted during fine-tuning. The experimental results are presented in Table 5. The joint audio-visual model achieved equal or better performance compared to the single-modality baselines. Particularly, it significantly improved VSR results, demonstrating the effectiveness of jointly modeling audio and video for cross-modal distillation. We also compared the use of lip ROI versus full face as inputs. The face-based model significantly outperformed the lip ROI-based model in the VSR task, consistent with our previous research [38].

Pretrain	Knowledge Distillation Loss		30h		433h	
	Finetune		ASR	VSR	ASR	VSR
Reg	–		3.2	38.5	2.4	35.6
CE	–		3.4	40.2	2.5	36.6
KLD	–		3.3	38.1	2.4	34.9
Reg+CE	–		3.2	38.9	2.5	35.4
Reg+KLD (<i>proposed</i>)	–		3.1	38.0	2.4	34.8
	Reg		–	–	2.0	34.9
	CE		–	–	2.0	34.9
Reg+KLD (<i>proposed</i>)	KLD		–	–	2.0	34.8
	Reg + CE		–	–	2.1	35.1
	Reg + KLD (<i>proposed</i>)		–	–	1.9	34.7

Table 4: WER (%) results from the ablation study of our proposed knowledge distillation loss. “Reg”, “CE”, and “KLD” represent feature regression, cross-entropy loss using discrete pseudo labels, and KL divergence loss using our proposed soft labels, respectively.

Inputs	Video Type	30h		433h	
		ASR	VSR	ASR	VSR
Teacher (iFLYTEK-ASR)					
✓	✗	–	1.7	–	1.4
Student					
✓	✗	–	3.0	–	2.4
✗	✓	Lip	–	44.4	–
✓	✓	Lip	3.0	40.8	2.5
✓	✓	Face	3.1	38.0	2.4
					34.8

Table 5: WER (%) results from the ablation study of model inputs. “Lip” and “Face” indicate the use of lip region-of-interest and entire face images as visual inputs, respectively. To ensure comparability, we equipped the teacher model with a randomly initialized decoder that shared the same configuration as the student’s.

Masking	KD Region	30h		433h	
		ASR	VSR	ASR	VSR
✗	All	3.5	43.0	2.7	37.2
✓	Mask	3.2	38.3	2.5	35.3
✓	All	3.1	38.0	2.4	34.8

Table 6: WER (%) results from the ablation study of masking strategy. “Masking” indicates whether the span-based feature mask was applied to the student during pretraining. “All” and “Mask” represent computing KD loss over entire region and only the masked region, respectively.

4.3.4. Masking strategy

This section investigates the masking strategy used for the student during pretraining. Experiments were conducted using only the iFLYTEK-speech teacher model and the Base version of the student, with the knowledge distillation loss omitted during finetuning. 433-hour LRS3 dataset was used for pretraining. The results are summarized in Table 6. Applying span masking to the intermediate features enhanced the model performance by compelling the model to utilize contextual information for recovering representations in masked regions. This approach has been extensively explored in previous research [6, 13, 10]. Unlike AV-HuBERT, which predicts only masked regions, our experiments demonstrated that computing KD loss over all regions yielded the best performance. This may be attributed to the high-quality learning targets derived from SFMs, which provided accurate and useful knowledge across all regions.

4.3.5. Multi-layer representations aggregation

This section explores the effect of layer averaging of multi-layer representations from the teacher model. Both WavLM and iFLYTEK-speech-based teacher model were investigated and compared, with the Base version of the student model. Each model was pretrained for 200k updates. 433-hour and 30-hour LRS3 dataset were used for pretraining and finetuning, respectively. Knowledge distillation loss was not used during finetuning. The results on our validation set are presented in Figure 4. For the WavLM-based teacher, the WER initially decreased but then increased as the number of averaged layers grew. Eventually, averaging the last 8 layers achieved the lowest WER, demonstrating the effectiveness of aggregating multi-layer representations for

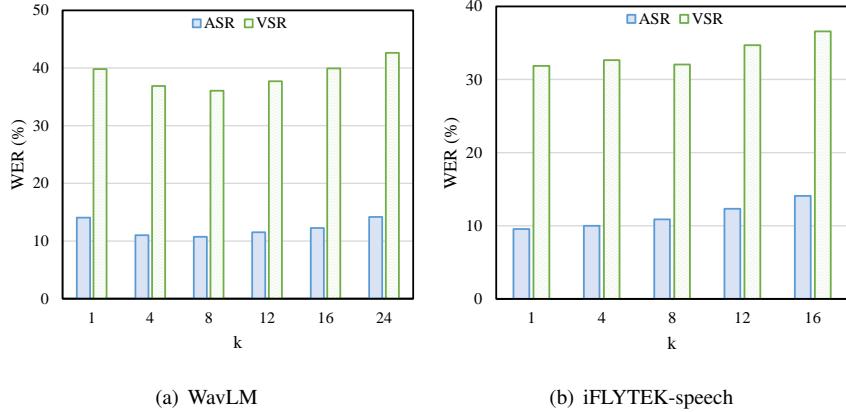


Figure 4: WER (%) results on our validation set as a function of k (number of last teacher layers averaged) for WavLM and iFLYTEK-speech based teachers.

distillation. In contrast, for the iFLYTEK-speech-based teacher model, the WER consistently increased as the number of averaged layers increased. This divergent behavior between the two teachers can be attributed to their distinct methods of encoding knowledge across layers, which are influenced by their respective training objectives [39].

4.4. Representations visualization

In this section, our proposed model with “Base” configuration and an ensemble of two teachers was used to extract the hidden representations from either audio or visual inputs. The Montreal Forced Aligner tool⁵ was then adopted to align the speech with the phoneme sequences using test set, where phoneme sequences was obtained via a grapheme-to-phoneme (G2P) model. A phoneme set of 39 was utilized. Hidden representations for each phoneme was collected and then averaged. Using these representations, Euclidean distances between each pairs of phonemes were computed. A hierarchical clustering algorithm was subsequently employed to cluster the phonemes according to their distances. The distance matrices are visualized in Figure 5. As expected, audio representations exhibit better discriminability for phonemes compared to visual ones. Interestingly, the visual representations in the final layer showed clear

⁵<https://montreal-forced-aligner.readthedocs.io/en/latest/index.html>

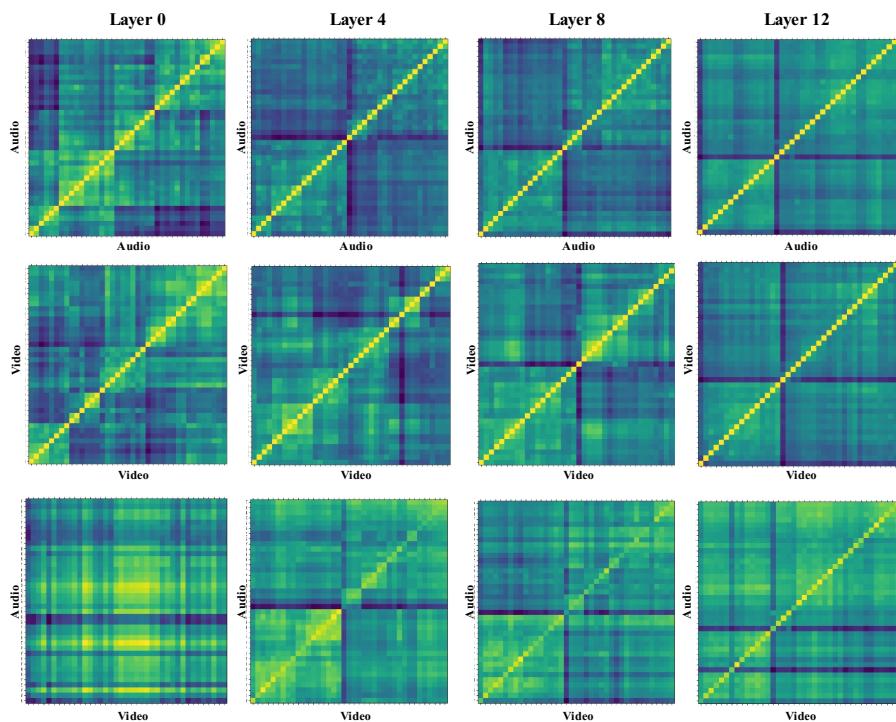


Figure 5: Distance matrices between phonemes using their corresponding representations from the student. Hierarchical clustering was applied to cluster phonemes based on their distances. “Audio” and “Video” refer to phoneme representations extracted from audio and visual inputs respectively. “Layer N” indicates representations from the N-th layer of the student model.

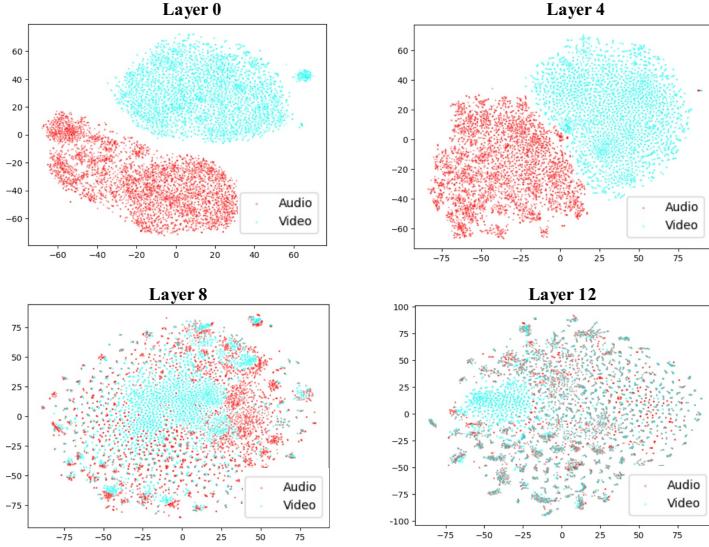


Figure 6: t-SNE visualization of audio and visual representations from the student model. Layer N” indicates representations from the N-th layer of the student model.

separation among different phonemes, which explains our model’s strong performance in the VSR task. The last row of this figure shows that the audio and visual representations of the same phoneme became closer in deeper layers, indicating that both modality shared the same space in deeper layers. Audio and visual representations were also projected into a two-dimensional space using the t-SNE algorithm [45] and visualized in Figure 6. This figure further confirms that the two distinct modality inputs were progressively processed and encoded within a unified space. These results align with recent studies demonstrating a trend towards representational convergence across modalities [46].

5. Conclusion

In this work, we propose an audio-visual representation learning method through cross-modal distillation using speech foundation models. Our method requires only unlabeled data during pretraining. Multi-layer representations from multiple speech

foundation models are extracted to distill the student model. Our proposed knowledge distillation loss combines a feature regression loss and a KL divergence loss with designed soft labels. The distillation loss is also applied during finetuning to enhance the performance of downstream tasks. By distilling from speech foundation models, our method effectively transfers the rich knowledge acquired from a vast volume of audio data to the video modality. We used two teacher models in our experiments: a self-supervised model, WavLM, and a supervised model, iFLYTEK-speech. The experimental results demonstrated that our proposed method achieved superior or at least comparable performance to baselines from previous studies. Ablation studies confirmed the effectiveness of several key components of our method, including the multi-teacher ensemble, the proposed distillation loss, and the joint modeling of audio and visual modalities. Furthermore, visualization of learned representations indicated that our model effectively encodes audio and visual modalities within a unified space.

Our work has several limitations that remain to be addressed in future research. First, in the lipreading task, while our method achieves lower WER than self-supervised learning baselines, it still falls short of the current best supervised models that utilize a large amount of labeled audio-visual data. This underscores the critical role of dataset scale in achieving optimal lipreading performance. As a future direction, we plan to collect more extensive audio-visual datasets to train lipreading models. Second, in the AVSR task, although our model generally outperforms baselines, its performance deteriorates when the input audio is corrupted with speech noise at -10dB SNR. This observation may suggest that our model has an over-reliance on the audio modality. Given that we adopted a conventional channel-wise concatenation strategy for audio and visual fusion, which was not the primary focus of this work, we intend to further investigate this issue. In future work, we aim to design more flexible and effective methods for integrating audio and visual modalities to achieve more robust audio-visual speech recognition.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62401348), Fundamental Research Funds for the Central Universities (Grant No. GK202406005).

7. Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used iFLYTEK's Xinghuo⁶ in order to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

- [1] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., WavLM: Large-scale self-supervised pre-training for full stack speech processing, *IEEE Journal of Selected Topics in Signal Processing* 16 (6) (2022) 1505–1518.
- [2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: *Proceedings of the International Conference on Machine Learning (ICML)*, PMLR, 2023, pp. 28492–28518.
- [3] J. Shin, J. Lee, D. Kim, Real-time lip reading system for isolated korean word recognition, *Pattern Recognition* 44 (3) (2011) 559–571.
- [4] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, A. Zisserman, Deep audio-visual speech recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

⁶<https://xinghuo.xfyun.cn/>

- [5] X. Yu, S. Sun, Y. Tian, Self-distillation and self-supervision for partial label learning, *Pattern Recognition* 146 (2024) 110016.
- [6] B. Shi, W.-N. Hsu, K. Lakhotia, A. Mohamed, Learning audio-visual speech representation by masked multimodal cluster prediction, in: Proceedings of the International Conference on Learning Representations (ICLR), 2022, pp. 1–12.
- [7] Q. Zhu, L. Zhou, Z. Zhang, S. Liu, B. Jiao, J. Zhang, L. Dai, D. Jiang, J. Li, F. Wei, VATLM: Visual-audio-text pre-training with unified masked prediction for speech representation learning, *IEEE Transactions on Multimedia* 26 (2024) 1055–1064.
- [8] S. Ma, Z. Zeng, D. McDuff, Y. Song, Contrastive learning of global and local video representations, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34, 2021, pp. 7025–7040.
- [9] Z.-Q. Zhang, J. Zhang, J.-S. Zhang, M.-H. Wu, X. Fang, L.-R. Dai, Learning contextually fused audio-visual representations for audio-visual speech recognition, in: *Proceedings of the International Conference on Image Processing (ICIP)*, IEEE, 2022, pp. 1346–1350.
- [10] A. Haliassos, P. Ma, R. Mira, S. Petridis, M. Pantic, Jointly learning visual and auditory speech representations from raw data, in: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023, pp. 1–15.
- [11] J. Lian, A. Baevski, W.-N. Hsu, M. Auli, AV-data2vec: Self-supervised learning of audio-visual speech representations with contextualized target representations, in: *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.
- [12] W.-N. Hsu, B. Shi, u-HuBERT: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022) 21157–21170.

- [13] J.-X. Zhang, G. Wan, Z.-H. Ling, J. Pan, J. Gao, C. Liu, Self-supervised audio-visual speech representations learning by multimodal self-distillation, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.
- [14] P. Ma, R. Mira, S. Petridis, B. W. Schuller, M. Pantic, LiRA: Learning visual speech representations from audio through self-supervision, in: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), 2021, pp. 3011–3015.
- [15] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, Y. Bengio, Multi-task self-supervised learning for robust speech recognition, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 6989–6993.
- [16] N. Liu, K. Wei, Y. Yang, J. Tao, X. Sun, F. Yao, H. Yu, L. Jin, Z. Lv, C. Fan, Multi-modal cross-lingual summarization for videos: A revisit in knowledge distillation induced triple-stage training method, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46 (12) (2024) 10697–10714.
- [17] T. Afouras, J. S. Chung, A. Zisserman, ASR is all you need: Cross-modal distillation for lip reading, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 2143–2147.
- [18] Y. Zhao, R. Xu, X. Wang, P. Hou, H. Tang, M. Song, Hearing lips: Improving lip reading by distilling speech recognizers, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vol. 34, 2020, pp. 6917–6924.
- [19] K. P. Huang, T.-H. Feng, Y.-K. Fu, T.-Y. Hsu, P.-C. Yen, W.-C. Tseng, K.-W. Chang, H.-Y. Lee, Ensemble knowledge distillation of self-supervised speech models, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.
- [20] J. Xue, C. Fan, J. Yi, J. Zhou, Z. Lv, Dynamic ensemble teacher-student distil-

lation framework for light-weight fake audio detection, *IEEE Signal Processing Letters* 31 (2024) 2305–2309.

- [21] C. Fan, M. Ding, J. Tao, R. Fu, J. Yi, Z. Wen, Z. Lv, Dual-branch knowledge distillation for noise-robust synthetic speech detection, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024) 2453–2466.
- [22] P. Ma, S. Petridis, M. Pantic, End-to-end audio-visual speech recognition with conformers, in: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 7613–7617.
- [23] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, M. Pantic, Audio-visual speech recognition with a hybrid CTC/Attention architecture, in: *Proceedings of the Spoken Language Technology Workshop (SLT)*, IEEE, 2018, pp. 513–520.
- [24] T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, O. Siohan, Recurrent neural network transducer for audio-visual speech recognition, in: *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019, pp. 905–912.
- [25] B. Martinez, P. Ma, S. Petridis, M. Pantic, Lipreading using temporal convolutional networks, in: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6319–6323.
- [26] D. Serdyuk, O. Braga, O. Siohan, Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video, in: H. Ko, J. H. L. Hansen (Eds.), *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2022, pp. 2833–2837.
- [27] K. Prajwal, T. Afouras, A. Zisserman, Sub-word level lip reading with visual attention, in: *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5162–5172.
- [28] J. Hong, M. Kim, D. Yoo, Y. Ro, Visual context-driven audio feature enhancement for robust end-to-end audio-visual speech recognition, in: *Proceedings of*

- the Annual Conference of the International Speech Communication Association (INTERSPEECH), 2022, pp. 2838–2842.
- [29] J. Li, C. Li, Y. Wu, Y. Qian, Unified cross-modal attention: Robust audio-visual speech recognition and beyond, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024) 1941–1953.
- [30] Y. A. D. Djilali, S. Narayan, H. Boussaid, E. Almazrouei, M. Debbah, Lip2Vec: Efficient and robust visual speech recognition via latent-to-latent visual to audio representation mapping, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13790–13801.
- [31] K. R. Prajwal, T. Afouras, A. Zisserman, Speech recognition models are strong lip-readers, in: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2024, pp. 2425–2429.
- [32] C. Simic, T. Bocklet, Self-supervised adaptive AV fusion module for pre-trained ASR models, in: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12787–12791.
- [33] A. Rouditchenko, Y. Gong, S. Thomas, L. Karlinsky, H. Kuehne, R. Feris, J. Glass, Whisper-Flamingo: Integrating visual features into whisper for audio-visual speech recognition and translation, in: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2024, pp. 2420–2424.
- [34] U. Cappellazzo, M. Kim, H. Chen, P. Ma, S. Petridis, D. Falavigna, A. Brutti, M. Pantic, Large language models are strong audio-visual speech recognition learners, *arXiv preprint arXiv:2409.12319* (2024).
- [35] R. Bommasani, D. A. Hudson, E. A. et al., On the opportunities and risks of foundation models, *ArXiv* (2021).
URL <https://crfm.stanford.edu/assets/report.pdf>
- [36] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, HuBERT: Self-supervised speech representation learning by masked prediction

- of hidden units, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021) 3451–3460.
- [37] Y. Zhang, S. Yang, J. Xiao, S. S. Shan, X. Chen, Can we read speech beyond the lips? Rethinking ROI selection for deep visual speech recognition, in: *IEEE International Conference on Automatic Face and Gesture Recognition*, 2020, pp. 356–363.
 - [38] J.-X. Zhang, G. Wan, J. Pan, Is lip region-of-interest sufficient for lipreading?, in: *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, 2022, p. 368–372.
 - [39] A. Pasad, J.-C. Chou, K. Livescu, Layer-wise analysis of a self-supervised speech representation model, in: *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 914–921.
 - [40] D. Senushkin, N. Patakin, A. Kuznetsov, A. Konushin, Independent component alignment for multi-task learning, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 20083–20093.
 - [41] J. S. Chung, A. Nagrani, A. Zisserman, Voxceleb2: Deep speaker recognition, in: *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2018, p. 1086–1090.
 - [42] B. Shi, W. Hsu, A. Mohamed, Robust self-supervised audio-visual speech recognition, in: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2022, pp. 2118–2122.
 - [43] O. Chang, H. Liao, D. Serdyuk, A. Shah, O. Siohan, Conformer is all you need for visual speech recognition, in: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10136–10140.
 - [44] C. Chen, Y. Hu, Q. Zhang, H. Zou, B. Zhu, E. S. Chng, Leveraging modality-specific representations for audio-visual speech recognition via reinforcement learning, in: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 37, 2023, pp. 12607–12615.

- [45] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, *Journal of machine learning research* 9 (11) (2008).
- [46] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, et al., Image as a foreign language: BEiT pretraining for vision and vision-language tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19175–19186.