ES^3: Evolving Self-Supervised Learning of Robust Audio-Visual Speech Representations

Yuanhang Zhang[1,2], Shuang Yang[1,2], Shiguang Shan[1,2], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
zhangyuanhang15@mails.ucas.ac.cn, {shuang.yang, sgshan, xlchen}@ict.ac.cn

Abstract

We propose a novel strategy, ES^3, for self-supervised learning of robust audio-visual speech representations from unlabeled talking face videos. While many recent approaches for this task primarily rely on guiding the learning process using the audio modality alone to capture information shared between audio and video, we reframe the problem as the acquisition of ***shared***, ***unique*** *(modality-specific) and* ***synergistic*** *speech information to address the inherent* asymmetry *between the modalities. Based on this formulation, we propose a novel "evolving" strategy that progressively builds joint audio-visual speech representations that are strong for both uni-modal (audio & visual) and bi-modal (audio-visual) speech. First, we leverage the more easily learnable audio modality to initialize audio and visual representations by capturing audio-unique and shared speech information. Next, we incorporate video-unique speech information and bootstrap the audio-visual representations on top of the previously acquired shared knowledge. Finally, we maximize the total audio-visual speech information, including synergistic information to obtain robust and comprehensive representations. We implement ES^3 as a simple Siamese framework and experiments on both English benchmarks and a newly contributed large-scale Mandarin dataset show its effectiveness. In particular, on LRS2-BBC, our smallest model is on par with SoTA models with only 1/2 parameters and 1/8 unlabeled data (223h).*

1. Introduction

The task of learning audio-visual speech representations has gained increasing importance due to its potential to support a wide range of relevant applications. Building on the success in other domains [9, 11, 12, 14, 22, 31], several effective *self-supervised* approaches have recently emerged for this task by pre-training models on paired or mixed audio, visual and text data with pretext tasks like speech unit prediction [30, 59, 80]. However, learning representations in the audio-visual speech domain faces unique challenges brought by the inherent *asymmetry* between the two modalities involved. First, audio and video often convey distinct and complementary *types* of speech information. For example, /m/ and /n/ sound similar, but are visually different; on the contrary, /p/ and /b/ are visually similar but can be distinguished with audio [70]. Even more interesting is the presence of *synergistic* information – information that can further enhance speech perception [55] but only emerges when audio and video are considered together, and cannot be inferred when either modality is absent. Second, they manifest different *amounts* of noise and information. Specifically, in most existing audio-visual speech datasets, audio is characterized by high sampling rate and relatively clean speech samples, making audio representations readily learnable from these datasets. In contrast, the visual modality often presents several challenging factors, including missing (*e.g.* due to self-occlusion in profile faces or low image quality), redundant (*e.g.* due to appearance variations, makeup and beards etc.) and ambiguous (*e.g.* due to the lower sampling rate and the presence of homophenes) visual information. Their presence complicates learning meaningful and robust representations solely from video. The difficulties encountered in the continuous efforts to improve the performance of visual speech recognition (VSR) [15, 24, 40, 41] models illustrate this persistent obstacle. Additionally, these differences between the two modalities lead to an imbalance between their learning dynamics, requiring additional manual settings [3, 39, 42, 61].

In this work, we break down the problem of learning audio-visual speech representations into the learning of three different types of information: *shared* (across modalities), *unique* (modality-specific) and *synergistic* speech information. Based on our formulation, we propose a novel strategy called ES^3 (short for **E**volving **S**elf-**S**upervised Learning of Robust Audio-Visual **S**peech Representations)

to tackle the asymmetry between the two modalities for learning robust audio-visual speech representations. Specifically, our ES^3 employs a three-stage progressive learning approach. Initially, it focuses on acquiring the easily learnable audio-*unique* and *shared* speech information. Then, it addresses the more challenging visual-*unique* speech information while incorporating it into the initial audio-visual joint representation. Lastly, it optimizes the final audio-visual representations by leveraging *synergy*, on top of the acquired shared and modality-unique information.

Our contributions can be summarized as follows:

- We propose a new formulation for audio-visual speech representation learning by decomposing the task into the acquisition of *shared* (across modalities), *unique* (modality-specific) and *synergistic* speech information.
- We propose a novel strategy, ES^3 to progressively learn the three types of information in a self-supervised manner with a simple Siamese framework. This *evolving* strategy significantly alleviates the previous burden of tuning balancing parameters.
- Extensive experiments on two widely used benchmarks show that ES^3 rivals previous methods in auditory, visual and audio-visual speech recognition performance.
- We also contribute a new large-scale Mandarin audio-visual speech dataset named CAS-VSR-S101 to the community for evaluation of similar methods.

2. Related Work

Self-supervised learning from multi-modal data. The arguably most successful paradigm for self-supervised multi-modal learning in the past few years is *contrastive learning*, which exploits the common information encoded in different modalities [68], such as audio (A) [6, 7], image (I) [53, 66], video (V) [5, 8, 32, 67] and text (T) [4, 16, 43, 44, 49, 63]. These approaches focus on learning semantically aligned multi-modal representations that are effective for specific downstream tasks like cross-modal retrieval, but discard modality-unique information that is useful for other uni-modal tasks. To solve this problem, Gong et al. [25] combine contrastive learning with the masked prediction objective [28] to learn strong audio-visual representations suitable for both retrieval and joint classification. Liang et al. [35] propose a new algorithm, FACTORCL to learn self-supervised multi-modal representations that capture both shared and unique information and verify its effectiveness on multiple tasks. In this paper, we explicitly address capturing unique, shared and synergistic information for uni- and multi-modal downstream tasks.

Learning audio-visual speech representations. A main challenge of learning audio-visual speech representation is properly handling the aforementioned asymmetry between audio and video. An established approach is to leverage the easily learnable audio modality as a teacher to guide visual or audio-visual speech representation learning. Implementing additional measures to mitigate audio dominance and promote visual features during learning is vital for this paradigm. For instance, Shi et al. [59] and Zhang et al. [76] use different masking strengths for the two modalities, apply modality dropout, and enforce simple audio features by only using a weak linear projection layer. Others have explored freezing pre-trained audio representations [2, 38, 78]. However, such an audio-centric approach is sub-optimal and inefficient. First, it limits the learned representations to capturing only shared, *i.e*. modality-invariant information, or at most, shared and audio-unique information. Recently, a few attempts have been made to go beyond merely using audio as the anchor modality. Haliassos et al. [27] combine multiple audio-visual prediction tasks in place of purely audio-centric ones, and treats A-V asymmetry by having the visual stream predict only auditory but not visual targets. Lian et al. [34] perform a series of trials to determine an optimal schedule for audio and video input combination. Second, the process of exhaustively testing and identifying the optimal set of balancing parameters such as masking ratios, modality schedules and modality dropout rate is extremely time-consuming and computationally demanding. In contrast, our newly proposed evolving strategy identifies different types of task-relevant information and learns them progressively, while avoiding pitfalls and computational burdens (see supp).

3. Our Proposed ES^3

3.1. Preliminaries

The goal of ES^3 is to learn audio-visual speech representations from unlabeled talking face videos. The inputs to our ES^3 are paired audio-visual data $X_{a,v} = \{(X_{a,i}, X_{v,i}) \mid i = 1, 2, \ldots, N\}$. Each of the N samples consists of a sequence of audio frames $X_{a,i} = \text{LogFBank}(\mathbf{a}_i) \in \mathbb{R}^{F \times T_{a,i}}$ and a temporally synchronized video sequence $X_{v,i} = \mathbf{v}_i \in \mathbb{R}^{C \times H \times W \times T_{v,i}}$ showing the lips of a speaker, where F is the filter bank dimension, C is the number of video channels, $T_{a,i}$ and $T_{v,i}$ are the total number of audio and video frames in the i-th sample, and H and W are the width and height of each frame. Audio and video are sampled at 100 and 25 FPS, *i.e*. for a 1-second clip, $T_a = 4T_v = 100$. The goal is to learn a d-dimensional *contextualized* representation $\gamma \in \mathbb{R}^{d \times T_v}$ capturing vital speech and context information within the two modality spaces.

Base architecture. The base architecture used in this paper is a relatively simple one and can be easily extended if needed. It includes two modality-specific encoders φ_a and φ_v, instantiated as a 2-layer CNN-based audio filter bank encoder and a *lightweight* VGG-(2+1)D video encoder [57, 58]. Here, φ_a downsamples X_a by 4x to match the video frame rate; φ_v has a temporal receptive field of 5 frames, or 200ms at 25 FPS, same as the most com-

monly adopted 3D-ResNet18 encoder [41, 59, 64], but has much fewer parameters. A *modality-agnostic* Transformer-based [71] modality fusion encoder θ equipped with rotary positional embeddings (RoPE) [65] follows the modality-specific encoders φ_a and φ_v. To allow θ to accept either audio, video, or audio-visual features, we adopt the simple *sum* operation to compute a fused audio-visual representation F_{fused} from modality features $F_m = \varphi_m(X_m)$, where $m \in \{a, v\}$, to allow easy handling of missing modalities:

$$F_{\text{fused}} = \begin{cases} F_a + F_v & \text{if input is audio-visual,} \\ F_a + \mathbf{0} = F_a & \text{if input is audio,} \\ F_v + \mathbf{0} = F_v & \text{if input is video.} \end{cases} \quad (1)$$

The Transformer encoder θ then yields the final contextual features $E = \theta(F_{\text{fused}})$.

Based on this base architecture, we build our Siamese based learning framework as shown in Fig. 1(a). A student model (*left*) is guided by a teacher model (*right*), which share the same base architecture. The student and teacher accept different versions of the same speech inputs (*e.g.* masked vs unmasked). The teacher model updates its parameters in an exponential moving average (EMA) manner, and the student is trained to regress the teacher outputs in general. In the subsequent sections, our objective is to optimize the student model with the base architecture to obtain the parameterized audio-visual speech representations $\gamma(\mathbf{a}, \mathbf{v}; \varphi_a, \varphi_v, \theta) = E$.

3.2. Problem Reformulation

In the audio-visual speech representation learning task, audio and video inputs can be taken as two different views of the same underlying speech signal, carrying different types of information. The relationship between the speech information in the two modalities can be visualized as the bottom-left Venn diagram in Fig. 1(b). Owing to the asymmetry between the two modalities, there exists both *modality-unique* speech information, *i.e.* audio- and video-unique information U_a and U_v. There is also speech information R *shared* between the two modalities, since they share the same underlying speech production process. At the same time, information that only emerges when both audio and video modality are present is the *synergistic* information, which we denote by S. Our overall objective is to learn a representation that maximizes the total information $R + U_a + U_v + S$ for bi-modal (audio-visual) speech representations, as well as $R+U_a$ and $R+U_v$ for uni-modal (audio & visual) speech representations.

With this in mind, we reframe the task of learning audio-visual speech representations as a mutual information maximization task. Specifically, the two modalities, corresponding to source variables X_a and X_v, have mutual information with the target speech content variable Y, containing information useful for downstream tasks that we wish to capture. Denote the joint mutual information between the sources and the target by $I(X_a, X_v; Y)$. The problem of learning robust audio-visual speech representations reduces to finding the set of parameters that maximizes $I(\gamma(X_a), \gamma(X_v); Y) = I(E_{av}; Y)$[1]. Similarly, the problem of learning robust audio and visual speech representations can be viewed as maximizing the mutual information $I(\gamma(X_a); Y) = I(E_a; Y)$ and $I(\gamma(X_v); Y) = I(E_v; Y)$, respectively. Taking advantage of the Partial Information Decomposition (PID) framework [73], we can precisely decompose the multivariate mutual information $I(E_{av}; Y), I(E_a; Y)$ and $I(E_v; Y)$ into the previous types of information as:

$$R + U_a = I(E_a; Y), \quad (2)$$
$$R + U_v = I(E_v; Y), \quad (3)$$
$$U_a + S = I(E_a; Y \mid X_v), \quad (4)$$
$$U_v + S = I(E_v; Y \mid X_a), \quad (5)$$
$$R - S = I(E_a; E_v; Y), \quad (6)$$
$$R + U_a + U_v + S = I(E_{av}; Y), \quad (7)$$

where R, U_a, U_v and S are non-negative.

To optimize the above mutual information terms with a neural network parameterized by φ_a, φ_v and θ, we divide the optimization process into three progressive and non-conflicting stages. In **stage 1**, we learn audio-unique information U_a and shared information R from the informative audio modality to maximize $I(E_a; Y)$. This process also helps to bootstrap visual representations by learning the shared information R. In **stage 2**, we turn to learning visual-unique information U_v by maximizing $I(E_v; Y) = R + U_v$, which is now easier with the visual information captured via R in stage 1. At the same time, we inject the learned visual information into the audio-visual joint representations to initialize $I(E_{av}; Y)$. Finally, in **stage 3** we optimize the joint audio-visual representation on top of the previously learned information by maximizing $I(E_{av}; Y)$.

3.3. The Complete Learning Framework

To learn network parameters that maximize the mutual information terms $I(E_\cdot; Y)$, we leverage the popular self-supervised *masked prediction task* based on the simple Siamese framework to support the learning process via self-distillation. Inspired by the recent work of Baevski et al. [12], we create multiple masked copies for each training sample, encode the unmasked regions with the Transformer encoder, and train the student to recover contextualized targets of the masked regions with a lightweight 1D convolutional decoder, which are generated by an online EMA teacher. This task can be used to learn any mutual infor-

[1]This equality holds since the architecture implemented in this work only models additive modality synergy stemming from sum fusion.

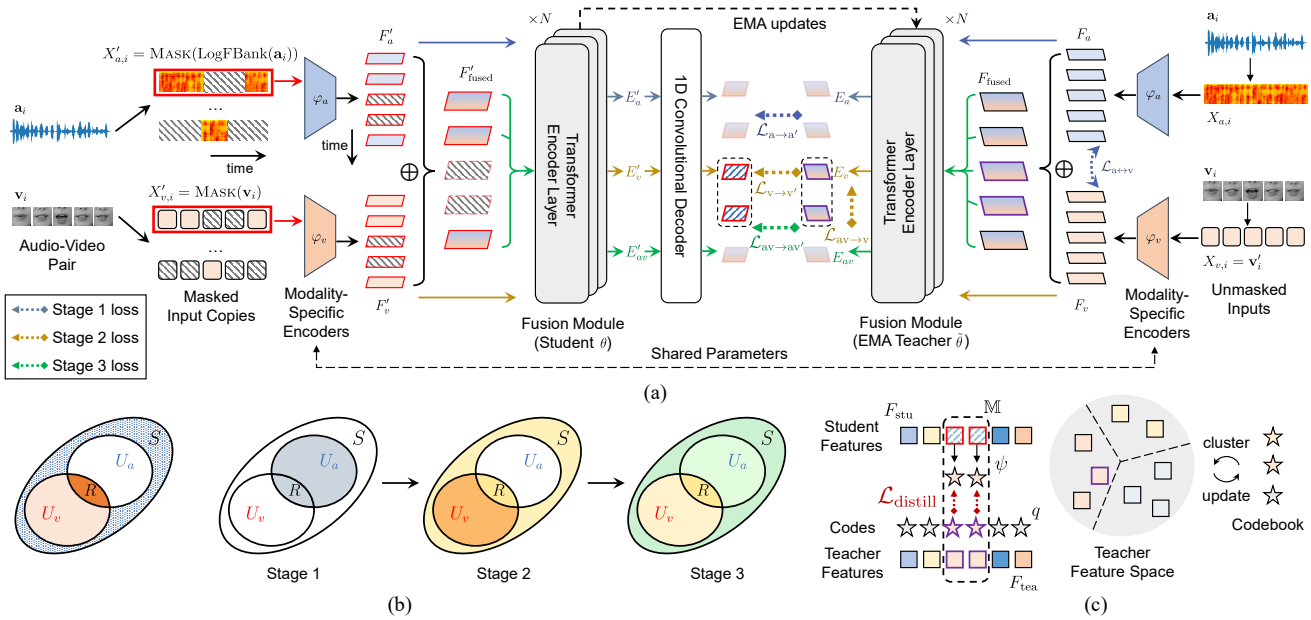

Figure 1. **Overview of the proposed method, ES^3.** (a) We use a Siamese framework with an EMA teacher to implement an iteration-free training procedure. (b) Venn diagrams showing how the three types of information (R, $U_\cdot$ and S) are acquired in an evolving learning process. (c) Illustration of self-distillation with online clustering corresponding to $\mathcal{L}_{\text{distill}}(F_{\text{tea}}, F_{\text{stu}})$.

mation $I(E_\cdot; Y)$ when Y is the downstream speech understanding task. For example, training on auditory or visual speech acts as a proxy to maximize $I(E_a; Y)$ or $I(E_v; Y)$.

Self-distillation with online clustering [36]. We perform distillation from the right-hand side teacher to the left-hand side student in Fig. 1(a). Specifically, as illustrated in Fig. 1(c), for each layer k of the top K Transformer layers in θ with hidden dimension D, we learn a codebook $\mathbf{Q}^k = \{\mathbf{q}_1^k, \ldots, \mathbf{q}_Q^k\}$ with Q codewords $\mathbf{q}_i^k \in \mathbb{R}^D$, which is updated as follows: for each entry j, a set $\tilde{\mathbf{Z}}_{\text{tea}}^k$ of the teacher output features $\tilde{\boldsymbol{z}}_{\text{tea}}^k$ closest to the current representation of q in the codebook is found as

$$\tilde{\mathbf{Z}}_{\text{tea}}^k = \left\{ \tilde{\boldsymbol{z}}_{\text{tea}}^k \;\middle|\; q = \underset{j \in \{1,2,\ldots,Q\}}{\arg\min} ||\tilde{\boldsymbol{z}}_{\text{tea}}^k - \mathbf{q}_j^k||_2 \right\}, \quad (8)$$

where the set indices q are used as discrete pseudo-labels to train the student model. Codewords are then updated in an EMA fashion as a weighted sum of the embeddings in $\tilde{\mathbf{Z}}_{\text{tea}}^k$ with decay rate τ_{code}. The associated loss function is

$$\mathcal{L}_{\text{distill}}(F_{\text{tea}}, F_{\text{stu}}) = \sum_{t \in \mathbb{M}} \sum_{k \in (K-N, K]} \log p_{\psi_k}(q_t | \boldsymbol{z}_t^K), \quad (9)$$

where F_{tea} and F_{stu} are teacher and student features, $\mathbb{M}$ is the set of masked time steps, ψ_k is the softmax activations of linear projections for each target layer k, q_t is the codeword index for time step t of the teacher $\tilde{\theta}$, and $\boldsymbol{z}_t^K$ is the output of the K-th and final layer of the student model θ. The teacher and the online clustering process do not require gradients.

EMA teacher parameterization. To reduce GPU memory usage, we apply EMA updates to the Transformer encoder θ, but not the earlier modality-specific encoders φ_a and φ_v, inspired by [50]. As shown in Fig. 1, the teacher model weights $\tilde{\theta}$ are computed as an exponential moving average of the student model weights θ:

$$\tilde{\theta} \leftarrow \eta\tilde{\theta} + (1 - \eta)\theta, \quad (10)$$

where $\eta = \min\{\eta_{\max}, (1 + \texttt{step})/(10 + \texttt{step})\}$ is a momentum parameter that slowly increases over time until a pre-determined value $\eta_{\max}$ is reached.

3.4. The Learning Process

We now elaborate on individual learning stages, based on the above Siamese framework and distillation loss.

Stage 1: Learning audio-unique information and shared information. In stage 1, we perform distillation from unmasked audio teacher model to the masked audio student model, yielding audio-unique information U_a and shared information R from audio. The distillation loss is:

$$\mathcal{L}_{\text{a}\to\text{a}'} = \mathcal{L}_{\text{distill}}(E_a, E_a'). \quad (11)$$

Being readily learnable, audio has the potential to provide most of R for learning shared information. However, optimizing $I(E_a; Y)$ alone cannot help visual learning. Noting that the InfoNCE loss between two variables has been widely used to maximize a lower bound on the mutual information, we adopt the InfoNCE loss to learn shared information corresponding to $I(E_a; E_v)$ to bootstrap visual learning with the shared information learned from audio.

Specifically, we apply it to a fine-grained frame-wise alignment task, where temporally aligned features from the same video are treated as positives, and unaligned ones negatives [19]. Since phonemes can appear repeatedly within the same video clip, leading to false negatives, we use a negative margin following Park et al. [47], which alleviates the effect of falsely matched pairs by enforcing a looser decision boundary. The objective function involved is:

$$\mathcal{L}_{\text{NCE}}(H) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{(H_{i,i}-m)/\tau}}{e^{(H_{i,i}-m)/\tau}+\sum_{i\neq j}e^{H_{i,j}/\tau}}, \tag{12}$$

where H is a matrix whose (i,j)-th element represents the cosine similarity between L^2-normalized audio and visual features at time step i and j, respectively, τ is a learnable temperature parameter, and m is a pre-determined margin. To promote the learning of shared features between the two modalities, we apply a symmetrized version of the above NCE loss as:

$$\mathcal{L}_{\text{a}\leftrightarrow\text{v}} = \frac{1}{2}(\mathcal{L}_{\text{a}\rightarrow\text{v}}+\mathcal{L}_{\text{v}\rightarrow\text{a}}) = \frac{1}{2}(\mathcal{L}_{\text{NCE}}(S_{av})+\mathcal{L}_{\text{NCE}}(S_{va})). \tag{13}$$

Here the similarity matrices $S_{av} = \langle F_a, F_v\rangle/(||F_a||\cdot||F_v||)$ and $S_{va} = S_{av}^{\top}$ are computed with features F_a and F_v from modality-specific encoders φ_a and φ_v instead of the contextual features from the Transformer encoder θ to avoid learning positional shortcuts. The total loss in stage 1 is

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{a}\rightarrow\text{a}'} + \mathcal{L}_{\text{a}\leftrightarrow\text{v}}. \tag{14}$$

Stage 2: Learning video-unique information. In stage 2, we proceed to learn video-unique information, which is now much easier than learning from scratch since it has been bootstrapped with shared knowledge from the audio in the previous stage. The audio encoder φ_a does not receive any gradients in this stage and is essentially frozen, since it is expected to have already acquired the audio-specific knowledge well in the previous stage. We adopt a similar distillation loss as stage 1:

$$\mathcal{L}_{\text{v}\rightarrow\text{v}'} = \mathcal{L}_{\text{distill}}(E_v, E'_v). \tag{15}$$

On top of the learned visual-unique information, we also perform distillation from unmasked video features to the joint features of unmasked audio-visual inputs in this stage to prepare for the learning of the final robust audio-visual representations in the next stage. For this purpose, the loss in this stage is set as:

$$\mathcal{L}_{\text{av}\rightarrow\text{v}} = \mathcal{L}_{\text{distill}}(E_v, E_{av}), \tag{16}$$

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{v}\rightarrow\text{v}'} + \mathcal{L}_{\text{av}\rightarrow\text{v}}. \tag{17}$$

To avoid repeated learning of only shared information between A and V again, the cross-modal loss $\mathcal{L}_{\text{av}\rightarrow\text{v}}$ is asymmetric, strictly one-way from AV to V, and a learnable temperature τ_v Eq. (16) is introduced to the visual student outputs, *i.e.*, we replace ψ_k in Eq. (9) with softmax activations of *temperature-scaled* linear projections of E_v.

Stage 3: Learning synergistic information. Finally, we aim to learn synergistic audio-visual information which previous methods have not explicitly addressed. Our ultimate goal is to maximize total task-relevant multi-modal information $I(E_{av}; Y)$, *i.e.* the sum of R, U_a, U_v and S, so that the representations are useful for both uni-modal and multi-modal tasks. Naïvely maximizing this objective without prior stages is problematic because the network tends to be biased towards learning audio-based representations (*i.e.* prioritizing R and U_a) without considering the visual modality. By learning most of R, U_a and U_v through the previous stages, our method is less susceptible to learning biased representations in the final stage. At this stage, we always mask identical positions in the audio and video inputs together so that the masked region never contains any unimodal information that can collapse or degrade training. To better recover the masked inputs, the model must cleverly leverage both modalities, which requires learning synergistic information that is not well-captured in prior stages. The loss function in stage 3 can be written as:

$$\mathcal{L}_{\text{stage3}} = \mathcal{L}_{\text{av}\rightarrow\text{av}'} = \mathcal{L}_{\text{distill}}(E'_{av}, E_{av}). \tag{18}$$

By progressively acquiring the three types of useful information, we build strong and robust audio-visual speech representations for both uni- and bi-modal speech, which we shall show in subsequent experiments.

3.5. Fine-tuning Stage

After pre-training, we add an extra linear layer after the fusion encoder θ to fine-tune the model end-to-end with the Lattice-Free Maximum Mutual Information (LF-MMI) criterion [26, 51], which is much more lightweight than seq2seq Encoder-Decoder models and allows highly efficient decoding. We adopt a "compact" [33] 1-state hidden Markov model (HMM) topology, using `k2`[2] for differentiable Weighted Finite State Transducer (WFST)-based training. We use 624 bi-char units from text-based clustering [77] for English and tonal phone-level targets (initials and finals) for Chinese.

4. Experiments

4.1. Datasets

LRS2-BBC [3]. LRS2-BBC is a 224-hour transcribed audio-visual English speech dataset collected from programs of various genres broadcast by the BBC. It is challenging due to the presence of extreme poses, as well as a large and diverse vocabulary. We use the official train / validation split (see supplementary materials) for evaluation.

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

LRS3-TED [1]. LRS3-TED is another audio-visual English speech dataset collected from TED and TEDx talks, with over 5.5k speakers and 439 hours of data. It covers diverse topics and is widely adopted for evaluating audio-visual speech representation learning methods for being the largest publicly available one of its kind. An official train / validation split is not given, so we use the widely adopted one prepared by Shi et al. [59], same as [27, 34, 80].

CAS-VSR-S101. We also collect a new large-scale, in-the-wild Mandarin dataset with 101.1 hours of data. The videos are sourced from broadcast news and conversational programs in Chinese, covering a highly diverse set of topics, speakers and filming conditions (see supplementary).

4.2. Experimental Setup

More implementation details can be found in the supplementary materials.

High- and low-resource setups. On LRS3-TED we follow the widely adopted setup of Shi et al. [59], which excludes the `pre-train` set during low-resource fine-tuning. To validate data scaling properties, previous works used VoxCeleb2-En [59], which is a 1326-hour subset of VoxCeleb2 [18] in English. However, not only are experiments of this scale computationally expensive, but more importantly, VoxCeleb2 is no longer available from its official source as of Feb. 2023. Therefore, we create a similar scaling protocol, using LRS3-TED (433h) as a pre-training source and LRS2-BBC (223h) for fine-tuning. For low-resource fine-tuning, we only use the 28-hour `train` subset of LRS2-BBC, while for high-resource fine-tuning, we include the `pre-train` subset, similar to [17].

Evaluation. Decoding is performed using beam search with a beam size of 50. The common evaluation metrics for ASR, VSR and AVSR are Word Error Rate (WER) and Character Error Rate (CER) for English and Chinese, respectively. As our bi-char LF-MMI models are trained on verbatim transcriptions for ease of lexicon construction, we report WERs both without and with using Whisper-style text normalization [54] (in parentheses) to compensate for innocuous errors related to text normalization.

4.3. Results on LRS2-BBC and LRS3-TED

Low-resource setup. Low-resource results on LRS2-BBC and LRS3-TED can be found in Tabs. 1 and 2. Our smallest BASE* encoder is modality-agnostic, only half the size of the one in [27], which uses two separate encoders to perform AVSR. On LRS3-TED, although we employ the simple LF-MMI instead of cross-entropy (CE) based seq2seq criterion for fine-tuning, our BASE model rivals or acheives SoTA in all cases. When scaling to a LARGE model, we observe significant improvements in both ASR and AVSR. Meanwhile, VSR performance remains comparable to SoTA, with a moderate 2% gap. This may suggest that with our current VGG-based visual encoder φ_v, the benefits of scaling up the Transformer primarily manifest in ASR and AVSR.

High-resource setup. High-resource results on LRS2-BBC and LRS3-TED can be found in Tabs. 1 and 3. On LRS2-BBC, with a 46M model and relying on its own 223h data for pre-training and fine-tuning, we achieve 30.3%/3.6%/2.9% WER for VSR, ASR and AVSR. Notably, these results surpass a highly competitive supervised method [40] and other self-supervised methods that incorporate additional data. This indicates that our progressive learning approach can effectively exploit visual information. Furthermore, when we introduce additional 433h unlabeled data, we achieve even better performance: 29.8%/3.0%/2.4% WER, which outperforms other methods that employ the same amount of data. When we switch to the LARGE model, performance rivals SoTA methods that employ 4x data (1759h) for pre-training. Similarly, we achieve or rival SoTA performance on LRS3-TED.

4.4. Results on CAS-VSR-S101

We also evaluate on our newly collected CAS-VSR-S101 dataset, which is smaller than both of the two previous datasets and in an even more ambiguous language (Chinese has more homophonous syllables and words than English). As shown in Tab. 4, ES^3 still works out-of-the-box under this challenging setup, and yields noticeable performance gains over a strong fully supervised baseline [39]. Interestingly, our fully supervised baseline model beats [39] in ASR and AVSR but slightly lags behind in VSR, again hinting that our strategy can further benefit from more complex visual modeling architectures.

Table 4. **Results on CAS-VSR-S101.** Here we use a BASE* model for the fully-supervised baseline and ES^3. The self-supervised models are pre-trained on training data and then fine-tuned for 50 epochs with the labels. The numbers before and after slashes correspond to CERs on the validation and test sets.

Methods	Criterion	Encoder Size	#Epochs	VSR	ASR	AVSR
Supervised						
Baseline	LF-MMI	46M	100	60.0 / 57.4	**15.4 / 12.6**	**14.6 / 12.1**
E2E Conformer [39]	CTC+CE	79M	100	**58.6 / 56.5**	16.6 / 13.0	16.0 / 12.6
Self-supervised						
ES^3 Stage 1				61.2 / 59.0	14.7 / 11.9	-
ES^3 Stage 2	LF-MMI	46M	50	59.1 / 56.7	-	13.9 / 11.4
ES^3 Stage 3				**58.1 / 55.6**	**14.4 / 11.6**	**13.4 / 11.0**

4.5. Analysis and Discussion

Representation quality probing. In Tab. 5, we demonstrate the effect of each stage on LRS2-BBC. ASR performance between stages 1 and 3 are close, indicating that stage 1 indeed acquires most of audio-unique information. Comparing the VSR results between stages 1 and 2 reveals a significant gap, suggesting that stage 2 also effectively learns visual-unique information, hence the minimal differences between the final VSR result and those of stage 2. Moreover, comparing the results of stage 1-2 and stage

Table 1. **Results on LRS2-BBC.** We pre-train a BASE* model with 223h unlabeled data from LRS2-BBC, as well as BASE and LARGE models with 433h unlabeled data (LRS3-TED) to demonstrate scaling properties. *: external language models.

Methods	Unlabeled AV data	Labeled Data	Encoder Size	Criterion	VSR	ASR	AVSR
Supervised							
Yu et al. [75]	-	1519h	-	LF-MMI[1]	48.9	6.7	5.9
Ma et al. [39]*	-	380h	79M	CTC+CE	37.9	3.9	**3.7**
Ma et al. [41]*	-	818h	186M	CTC+CE	**27.9**	**2.6**	-
Semi-supervised							
Afouras et al. [2]*	777h	223h[2]	-	CTC	51.3	-	-
Ma et al. [41]*	2630h	818h	186M	CTC+CE	**14.6**	**1.5**	**1.5**
Self-supervised (BASE and BASE models)*							
AV-HuBERT [59, 60]	1759h	223h	103M	CE	31.2[3]	-	3.6[3]
VATLM [80]	1759h[4]	223h	107M	CE	**30.6**	-	2.9
RAVEn [27]	433h	223h	97M	CTC+CE	32.1	3.9	-
Pan et al. [45]	-[5]	380h	399M	CTC+CE	43.2	**2.7**	**2.6**
ES^3 (ours)	223h	28h	46M	LF-MMI	40.2 (39.1)	6.0 (5.0)	5.7 (4.9)
	433h	28h	102M		39.3 (38.2)	5.5 (4.8)	5.1 (4.1)
	223h	223h	46M		31.4 (30.3)	4.3 (3.6)	3.8 (2.9)
	433h	223h	102M		**30.7 (29.8)**	**3.4 (3.0)**	**3.2 (2.4)**
Self-supervised (LARGE models)							
ES^3 (ours)	433h	28h	317M	LF-MMI	36.4 (35.4)	5.2 (4.4)	4.7 (4.0)
	433h	223h	317M		**26.7 (25.8)**	**3.1 (2.5)**	**3.1 (2.5)**
AV-HuBERT [59, 60]	1759h	223h	325M	CE	25.5[3]	-	2.5[3]
VATLM [80]	1759h[4]	223h	332M	CE	24.3	-	2.3
RAVEn [27]	1759h	223h	671M	CTC+CE	23.2	2.5	-

[1] Not end-to-end; requires a GMM-HMM alignment stage. [2] Uses an additional ASR model trained on LibriSpeech (960h).
[3] Reproduced by Zhu et al. [80]. [4] Uses additional 3846h audio, 452h audio-text and 600M text data.
[5] Uses additional 60000h audio data and 1.28M unlabeled images.

Table 2. **Low-resource results on LRS3-TED.** We pre-train a BASE and LARGE model with 433h unlabeled data.

Methods	Unlabeled AV data	Labeled Data	Encoder Size	Criterion	VSR	ASR	AVSR
Self-supervised (BASE models)							
AV-HuBERT [59, 60]	433h	30h	103M	CE	51.8	4.9	4.7[1]
VATLM [80]	433h[2]	30h	107M	CE	48.0	-	3.6
RAVEn [27]	433h	30h	97M	CTC+CE	47.0	4.7	-
AV2vec [76][3]	433h	30h	103M	CE	45.0	-	5.8
AV-data2vec [34]	433h	30h	103M	CE	**45.2**	4.4	4.2
ES^3 (ours)	433h	30h	102M	LF-MMI	**45.5 (44.7)**	**3.9 (3.3)**	**3.6 (3.0)**
Self-supervised (LARGE models)							
AV-HuBERT	433h	30h	325M	CE	44.8	4.5	4.2[1]
AV-data2vec	433h	30h	325M	CE	**40.5**	3.7	3.4
ES^3 (ours)	433h	30h	317M	LF-MMI	43.5 (42.5)	**3.8 (2.9)**	**2.9 (2.3)**

[1] Reproduced by Lian et al. [34]. [2] Uses additional 3846h audio, 452h audio-text and 600M text data.
[3] Zhang et al. [76] inject noise during pre-training and obtain better fine-tuning results even with its base model AV-HuBERT (47.1%).

3 demonstrates that the joint information learning in stage 3, which also incorporates synergistic information, benefits both uni-modal (ASR & VSR) and bi-modal (AVSR) tasks. Similar trends can also be observed on the more challenging Mandarin dataset CAS-VSR-S101 in Tab. 4, further supporting the motivation of our proposed strategy.

Table 5. **Representation quality probing on LRS2-BBC.** Results are reported on the validation set using the BASE* model under the low-resource setup for each pre-training stage.

Pre-training Stage	VSR	ASR	AVSR
Stage 1	58.3 (57.3)	11.5 (10.1)	-
Stage 2	48.7 (47.3)	-	12.9 (11.5)
Stage 3 (Full ES^3)	**48.1 (46.9)**	**11.4 (9.8)**	**10.1 (8.6)**

Effect of automatically balancing shared and unique information learning. It is known that in a hardness-aware loss function, the temperature rules the strength of penalties on hard negative samples [72]. We track the automatically learned temperature parameter τ used for cross-modal learning objectives $\mathcal{L}_{\mathrm{a}\leftrightarrow\mathrm{v}}$ in Eq. (12) and $\mathcal{L}_{\mathrm{av}\rightarrow\mathrm{v}}$ in Eq. (16) when pre-training on LRS2-BBC, as shown in Fig. 2. In stage 1, τ continues to fall until it reaches $\tau_{\min}$, while in stage 2 it briefly drops and then rises throughout training. The trend is in line with our intuition: in stage 1, audio information is easily learnable, so the model can fully exploit shared information by lowering τ without sacrificing audio-unique information learning; in stage 2, after learning some residual shared information, an increasing τ progressively relaxes $\mathcal{L}_{\mathrm{av}\rightarrow\mathrm{v}}$ to yield more freedom to $\mathcal{L}_{\mathrm{v}\rightarrow\mathrm{v}'}$, *i.e.* video-unique information learning.

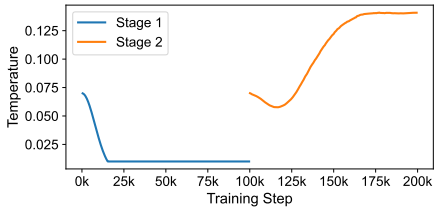

Figure 2. The temperature parameter paces learning automatically.

Table 3. **High-resource results on LRS3-TED.** We fine-tune on 433h labeled data of LRS3. †: uses test-time augmentation. *: uses external language models. ‡: noise injection during pre-training. We list more methods for comparison in the supplementary material.

Methods	Year	Unlabeled AV data	Labeled Data	Backbone	Encoder Size	Criterion	VSR	ASR	AVSR
Supervised									
Ma et al. [39]*	2021	-	595h	Conformer	79M	CTC+CE	43.3	**2.3**	**2.3**
Prajwal et al. [52]†*	2022	-	698h	Transformer	32M	CE	40.6	-	-
Ma et al. [40]*	2022	-	818h	Conformer	79M	CTC+CE	**34.7**	-	-
Semi-Supervised									
Shillingford et al. [62]*	2019	-	3886h	RNN	-	CTC	55.1	-	-
Makino et al. [42]	2019	-	31kh	RNN	43M	Transducer	33.6	4.8	4.5
Afouras et al. [2]*	2020	344h	433h[1]	Jasper (CNN)	-	CTC	59.8	-	-
Serdyuk et al. [57]	2021	-	90kh	Transformer	-	Transducer	25.9	-	2.3
Ma et al. [40]*	2022	641h	818h	Conformer	79M	CTC+CE	31.5	-	-
Serdyuk et al. [58]	2022	-	90kh	Conformer	-	Transducer	17.0	1.6	1.6
Ma et al. [41]*	2023	2630h	818h	Conformer	186M	CTC+CE	19.1	**1.0**	**0.9**
Chang et al. [15]	2023	-	100kh	Conformer	98M	Transducer	**12.8**	-	**0.9**
Self-supervised (Base Models)									
AV-HuBERT [59, 60]	2022	433h	433h	Transformer	103M	CE	44.0	3.0	2.8[2]
AV2vec [76]‡	2023	433h	433h	Transformer	103M	CE	39.9	-	2.6
RAVEn [27]	2023	433h	433h	Transformer	97M	CTC+CE	39.1	2.2	-
ES^3 (ours)	2023	433h	433h	Transformer	102M	LF-MMI	40.3 (39.2)	2.9 (2.4)	2.5 (2.0)
AV-data2vec [34]	2023	433h	433h	Transformer	103M	CE	**39.0**	**2.0**	**1.8**
Self-supervised (Large Models)									
AV-HuBERT [59, 60]	2022	433h	433h	Transformer	325M	CE	41.6	2.7	2.5[2]
AV-data2vec [34]	2023	433h	433h	Transformer	325M	CE	37.4	**1.9**	**1.7**
ES^3 (ours)	2023	433h	433h	Transformer	317M	LF-MMI	**37.1 (36.7)**	2.8 (2.1)	**2.1 (1.7)**

[1] Uses an additional ASR model trained on LibriSpeech (960h). [2] Reproduced by Lian et al. [34].

Visualization of our learned psuedo-label sequences. We randomly select a sample from the LRS3-TED validation set and visualize the pseudo-label sequences, *i.e.* code indices q generated by the audio, visual and audio-visual EMA teacher $\tilde{\theta}$ in each stage. Ground truth phones for each audio frame (10ms) are derived by forced alignment [79]. Fig. 3 shows that our learned codebook yields high-quality pseudo-labels that strongly correlate with phonemic units.

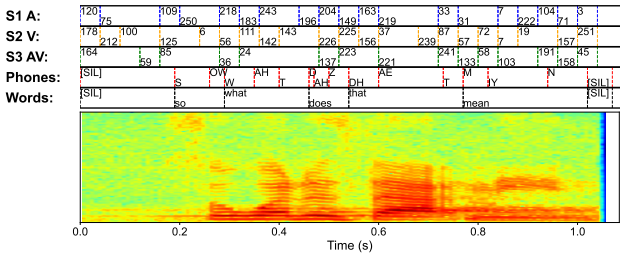

Figure 3. **Pseudo-label sequences from each stage.** Example of pseudo-labels generated from the last Transformer layer for the utterance *"so what does that mean"* by the audio, visual, and audio-visual teachers in each stage (BASE model). At the bottom we show the spectrogram for reference. Please note that our pseudo-labels are derived at 40ms intervals, limited by video frame rate.

Training efficiency comparison with SoTA methods. Learning audio-visual speech representations has been extremely compute-intensive. Here we briefly compare ES^3 with prominent SoTA methods. Pre-training a BASE/LARGE model on LRS3-TED using our ES^3 for 100 epochs per stage only takes about 6/8.3 days in total with just 8 NVIDIA A40 GPUs. In contrast, AV-data2vec is trained for 1000k steps using 64 NVIDIA V100 GPUs for 4-5/6-7 days using BASE/LARGE with a batch size of 1280/2560s, or equivalently, 821/1642 epochs. They also experiment extensively with hyperparameters such as modality schedulers, which further increases the cost of training. RAVEn performs an equivalent of 6.25/50 days of training on 8 A100s with BASE*/LARGE. It uses separate encoders for audio and video, so the resulting model cannot perform audio-visual fusion out-of-the-box. This also requires keeping track of four sets of parameters, which is very memory-intensive. By focusing on designing a clever learning strategy, we circumvent such manual tuning altogether and greatly reduce training costs. Finally, we use the simple LF-MMI criterion, and expect better performance from tuning beam size, including a larger external language model (see supplementary), or switching to the common heavy Attention-based Encoder-Decoder model.

5. Conclusions

We present a novel strategy, ES^3, for self-supervised learning of audio-visual speech representations. Unlike existing methods that primarily focus on learning information shared between audio and video, we take an "evolving" approach that also captures unique and synergistic information. Our experimental results demonstrate that ES^3 indeed progressively learns unique, shared and synergistic information. Moreover, in terms of downstream performance, it rivals SoTA methods in terms of ASR, VSR and AVSR accuracy without burdensome manual balancing efforts.

Acknowledgments. This work is partially supported by the National Natural Science Foundation of China (No. 62276247, 62076250). We are grateful to the ICT computing platform, the Nanjing Institute of InforSuperBahn, and Capitalonline Data Service Co., Ltd for providing additional compute resources. We also thank Chenhao Wang, Mingmin Yang and Jingyun Xiao for helping to prepare CAS-VSR-S101, Bingquan Xia and Songtao Luo for help with Conformer and seq2seq baselines, Bingquan Xia and Feixiang Wang for help during the rebuttal phase, and Feixiang Wang and Zhaoxin Yuan for an early version of Fig. 1.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR*, abs/1809.00496, 2018. 6

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. ASR is all you need: Cross-modal distillation for lip reading. In *ICASSP*, pages 2143–2147. IEEE, 2020. 2, 7, 8, 4, 5

[3] Triantafyllos Afouras, Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):8717–8727, 2022. 1, 5, 4

[4] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. In *NeurIPS*, 2020. 2

[5] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, 2020. 2

[6] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, pages 609–617. IEEE Computer Society, 2017. 2

[7] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV (1)*, pages 451–466. Springer, 2018. 2

[8] Yuki Markus Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *NeurIPS*, 2020. 2

[9] Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. Effectiveness of self-supervised pre-training for speech recognition. *CoRR*, abs/1911.03912, 2019. 1

[10] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020. 3

[11] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, pages 1298–1312. PMLR, 2022. 1

[12] Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *ICML*, pages 1416–1429. PMLR, 2023. 1, 3

[13] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *ICCV*, pages 1021–1030. IEEE Computer Society, 2017. 2

[14] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision Transformers. In *ICCV*, pages 9630–9640. IEEE, 2021. 1

[15] Oscar Chang, Hank Liao, Dmitriy Serdyuk, Ankit Shahy, and Olivier Siohan. Conformer is all you need for visual speech recognition. In *ICASSP*, pages 10136–10140. IEEE, 2024. 1, 8, 5

[16] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie W. Boggust, Rameswar Panda, Brian Kingsbury, Rogério Feris, David Harwath, James R. Glass, Michael Picheny, and Shih-Fu Chang. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *ICCV*, pages 7992–8001. IEEE, 2021. 2

[17] Xize Cheng, Tao Jin, Linjun Li, Wang Lin, Xinyu Duan, and Zhou Zhao. OpenSR: Open-modality speech recognition via maintaining multi-modality alignment. In *ACL (1)*, pages 6592–6607. Association for Computational Linguistics, 2023. 6

[18] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In *INTERSPEECH*, pages 1086–1090. ISCA, 2018. 6

[19] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *ICASSP*, pages 3965–3969. IEEE, 2019. 5

[20] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. *CoRR*, abs/2307.08691, 2023. 2

[21] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *NeurIPS*, 2022. 2

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019. 1, 2

[23] Angela Fan, Edouard Grave, and Armand Joulin. Reducing Transformer depth on demand with structured dropout. In *ICLR*. OpenReview.net, 2020. 2

[24] Dalu Feng, Shuang Yang, and Shiguang Shan. An efficient software for building lip reading models without pains. In *ICME Workshops*, pages 1–2. IEEE, 2021. 1

[25] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *ICLR*. OpenReview.net, 2023. 2

[26] Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur. End-to-end speech recognition using lattice-free MMI. In *INTERSPEECH*, pages 12–16. ISCA, 2018. 5

[27] Alexandros Haliassos, Pingchuan Ma, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Jointly learning visual and auditory speech representations from raw data. In *ICLR*. OpenReview.net, 2023. 2, 6, 7, 8, 1, 4, 5

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 15979–15988. IEEE, 2022. 2

[29] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *SPECOM*, pages 198–208. Springer, 2018. 2

[30] Wei-Ning Hsu and Bowen Shi. u-HuBERT: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality. In *NeurIPS*, 2022. 1

[31] 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 Trans. Audio Speech Lang. Process.*, 29:3451–3460, 2021. 1

[32] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, pages 7774–7785, 2018. 2

[33] Aleksandr Laptev, Somshubra Majumdar, and Boris Ginsburg. CTC variations through new WFST topologies. In *INTERSPEECH*, pages 1041–1045. ISCA, 2022. 5

[34] Jiachen Lian, Alexei Baevski, Wei-Ning Hsu, and Michael Auli. AV-data2vec: Self-supervised learning of audio-visual speech representations with contextualized target representations. In *ASRU*. IEEE, 2023. 2, 6, 7, 8, 1, 5

[35] Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Y. Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *NeurIPS*, 2023. 2

[36] Alexander H. Liu, Heng-Jui Chang, Michael Auli, Wei-Ning Hsu, and Jim Glass. DinoSR: Self-distillation and online clustering for self-supervised speech representation learning. In *NeurIPS*, 2023. 4, 2

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net, 2019. 2

[38] Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W. Schuller, and Maja Pantic. LiRA: Learning visual speech representations from audio through self-supervision. In *Interspeech*, pages 3011–3015. ISCA, 2021. 2

[39] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with Conformers. In *ICASSP*, pages 7613–7617. IEEE, 2021. 1, 6, 7, 8, 4, 5

[40] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages in the wild. *Nat. Mac. Intell.*, 4(11):930–939, 2022. 1, 6, 8, 2, 4, 5

[41] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-AVSR: Audio-Visual speech recognition with automatic labels. In *ICASSP*, pages 1–5. IEEE, 2023. 1, 3, 7, 8, 4, 5

[42] Takaki Makino, Hank Liao, Yannis M. Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *ASRU*, pages 905–912. IEEE, 2019. 1, 8, 5

[43] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640. IEEE, 2019. 2

[44] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. Speech2Action: Cross-modal supervision for action recognition. In *CVPR*, pages 10314–10323. Computer Vision Foundation / IEEE, 2020. 2

[45] Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin. Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition. In *ACL (1)*, pages 4491–4503. Association for Computational Linguistics, 2022. 7, 4

[46] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *INTERSPEECH*, pages 2613–2617. ISCA, 2019. 2

[47] Sooyoung Park, Arda Senocak, and Joon Son Chung. MarginNCE: Robust sound localization with a negative margin. In *ICASSP*, pages 1–5. IEEE, 2023. 5, 2

[48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019. 2

[49] Mandela Patrick, Yuki Markus Asano, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *ICCV*, pages 9557–9567. IEEE, 2021. 2

[50] Trung Pham, Chaoning Zhang, Axi Niu, Kang Zhang, and Chang D. Yoo. On the pros and cons of momentum encoder in self-supervised visual representation learning. *CoRR*, abs/2208.05744, 2022. 4

[51] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *INTERSPEECH*, pages 2751–2755. ISCA, 2016. 5

[52] K. R. Prajwal, Triantafyllos Afouras, and Andrew Zisserman. Sub-word level lip reading with visual attention. In *CVPR*, pages 5152–5162. IEEE, 2022. 8, 4, 5

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2

[54] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, pages 28492–28518. PMLR, 2023. 6

[55] Jordi Robert-Ribes, Jean-Luc Schwartz, Tahar Lallouache, and Pierre Escudier. Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of french oral vowels in noise. *The Journal of the Acoustical Society of America*, 103(6):3677–3689, 1998. 1

[56] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *INTERSPEECH*, pages 3465–3469. ISCA, 2019. 3

[57] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. Audio-visual speech recognition is worth $32 \times 32 \times 8$ voxels. In *ASRU*, pages 796–802. IEEE, 2021. 2, 8, 5

[58] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video. In *INTERSPEECH*, pages 2833–2837. ISCA, 2022. 2, 8, 5

[59] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *ICLR*. OpenReview.net, 2022. 1, 2, 3, 6, 7, 8, 4, 5

[60] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. Robust self-supervised audio-visual speech recognition. In *INTERSPEECH*, pages 2118–2122. ISCA, 2022. 7, 8, 4, 5

[61] Bowen Shi, Abdelrahman Mohamed, and Wei-Ning Hsu. Learning lip-based audio-visual speaker embeddings with AV-HuBERT. In *INTERSPEECH*, pages 4785–4789. ISCA, 2022. 1

[62] Brendan Shillingford, Yannis M. Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Misha Denil, Ben Coppin, Ben Laurie, Andrew W. Senior, and Nando de Freitas. Large-scale visual speech recognition. In *INTERSPEECH*, pages 4135–4139. ISCA, 2019. 8, 5

[63] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogério Feris, David Harwath, James R. Glass, and Hilde Kuehne. Everything at once - multi-modal fusion Transformer for video retrieval. In *CVPR*, pages 19988–19997. IEEE, 2022. 2

[64] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with LSTMs for lipreading. In *INTERSPEECH*, pages 3652–3656. ISCA, 2017. 3

[65] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864, 2021. 3

[66] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*. OpenReview.net, 2020. 2

[67] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *ICCV*, pages 7463–7472. IEEE, 2019. 2

[68] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV (11)*, pages 776–794. Springer, 2020. 2

[69] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image Transformers. In *ICCV*, pages 32–42. IEEE, 2021. 2

[70] Kristin J Van Engen, Avanti Dey, Mitchell S Sommers, and Jonathan E Peelle. Audiovisual speech perception: Moving beyond mcgurk. *The Journal of the Acoustical Society of America*, 152(6):3216–3225, 2022. 1

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 3, 2

[72] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, pages 2495–2504. Computer Vision Foundation / IEEE, 2021. 7

[73] Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010. 3

[74] Bo Xu, Cheng Lu, Yandong Guo, and Jacob Wang. Discriminative multi-modality speech recognition. In *CVPR*, pages 14421–14430. Computer Vision Foundation / IEEE, 2020. 5

[75] Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu. Audio-visual recognition of overlapped speech for the LRS2 dataset. In *ICASSP*, pages 6984–6988. IEEE, 2020. 7, 4

[76] Jing-Xuan Zhang, Genshun Wan, Zhen-Hua Ling, Jia Pan, Jianqing Gao, and Cong Liu. Self-supervised audio-visual speech representations learning by multimodal self-distillation. In *ICASSP*, pages 1–5. IEEE, 2023. 2, 7, 8, 1, 5

[77] Xiaohui Zhang, Vimal Manohar, David Zhang, Frank Zhang, Yangyang Shi, Nayan Singhal, Julian Chan, Fuchun Peng, Yatharth Saraf, and Mike Seltzer. On lattice-free boosted MMI training of HMM and CTC-based full-context ASR models. In *ASRU*, pages 1026–1033. IEEE, 2021. 5

[78] Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song. Hearing lips: Improving lip reading by distilling speech recognizers. In *AAAI*, pages 6917–6924. AAAI Press, 2020. 2

[79] Jian Zhu, Cong Zhang, and David Jurgens. Phone-to-audio alignment without text: A semi-supervised approach. In *ICASSP*, pages 8167–8171. IEEE, 2022. 8

[80] 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. 1, 6, 7, 4, 5