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

Supplementary Material

Please note that the references in this file follow the same numbering as those in the main paper.

1. Schematic Overview

Fig. 4 illustrates graphically the deficiencies of existing audio-visual speech representation learning techniques and the merits of our newly proposed ES^3, which acquires unique, shared and synergistic information in an evolving learning process.

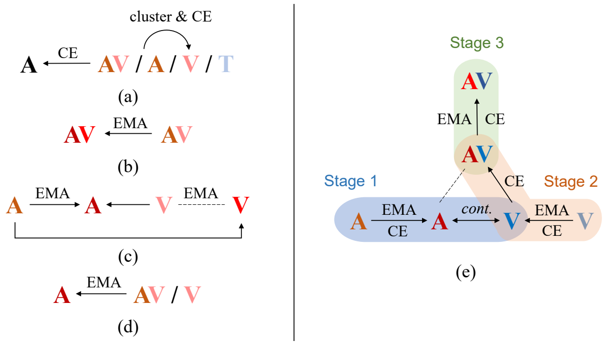

Figure 4. **A schematic overview of existing techniques and our proposed strategy, ES^3.** Existing methods mostly focus on bootstrapping visual and audio-visual representations from audio, leading to inadequate learning of visual-unique and synergistic information (see main text). *Left*: (a) AV-HuBERT [59], u-HuBERT [30] & VATLM [80]. (b) AV2vec [76]. (c) RAVEn [27]. (d) AV-data2vec [34]. *Right*: (e) our proposed strategy, ES^3. Letters in lighter shades represent masked versions of the specified modality. Arrowheads on solid lines point towards the *teachers* (or *targets*) providing the (self-)supervision signals. **A**: audio; **V**: video; **T**: text; ***cont.***: contrastive loss; **CE**: cross-entropy loss; **EMA**: exponential moving average updates.

2. Datasets

LRS2-BBC. The LRS2-BBC dataset (224h) is officially divided into four parts: a `pre-train` set (195h), a `train` set (28h), a development set (0.6h), and a test set (0.5h). The validation and test sets contain $1,082$ and $1,243$ utterances, respectively. The difference between the `pre-train` and `train` sets is that the former contains longer sentence excerpts while the latter only contains sentences clipped to 100 characters or 10 seconds.

CAS-VSR-S101. We collect a new large-scale, in-the-wild Mandarin dataset, CAS-VSR-S101 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. The lengths of the utterances are naturally distributed between 0.01s and 10.57s, and image qualities and resolutions vary. News accounts for 82.4% of the programs. 70.4% of the utterances depict news anchors, hosts and correspondents, while 29.6% are those of interviewees and guests. In addition, at a ratio of approximately $1.5 : 1$, male and female appearances are relatively balanced. It is divided into train, validation and test sets by TV channels to minimize speaker overlap, and at a ratio of roughly $8 : 1 : 1.5$ in terms of duration. The validation and test sets are composed of programs broadcast on provincial TV channels. A random visual sample of the dataset can be found in Fig. 5. The dataset is available for academic use under a license[3].

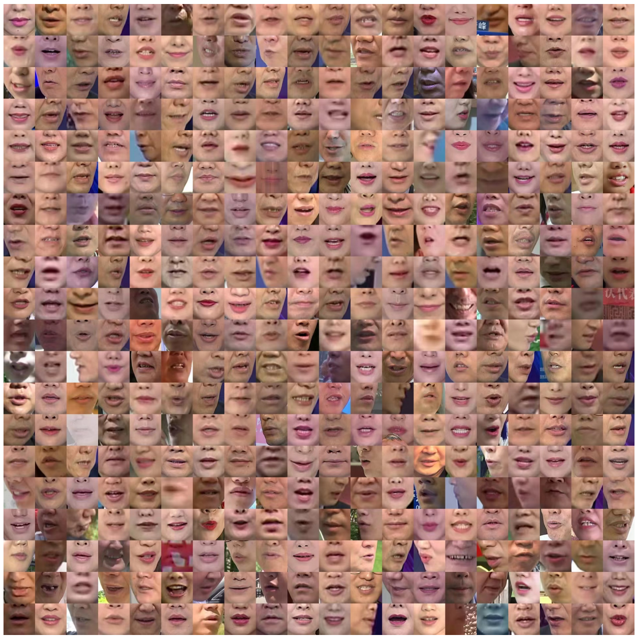

Figure 5. **A random sample of extracted mouth regions-of-interest from CAS-VSR-S101.** Our dataset is highly diverse in terms of speaker identity, visual conditions and image quality.

3. Implementation Details

Pre-processing. 80-dimensional log-Mel filter-bank features are extracted on-the-fly from 16kHz audio using the

[3] Please direct your inquiries to lipreading@vipl.ict.ac.cn.

kaldifeat package[4]. Similar to prior work, a 96 × 96 grayscale crop centered on the speaker's mouth is extracted from each frame based on pre-computed facial landmarks [13] after Procrustes-based registration to a canonical face template with a similarity transformation. The resulting crops are normalized to $[0, 1]$ with mean 0.45 and variance 0.225. All utterances longer than 15 seconds are segmented by silence or word boundary and batched dynamically by length to minimize padding.

Training details. For the Transformer encoder, there are two choices in terms of number of layers, hidden dimension and feedforward layer dimension: BASE (12/768/3072) and LARGE (24/1024/4096) [22, 71]. Following [27], a smaller version of BASE (12/512/2048, denoted by BASE*) that is suitable for fast prototyping is used by default for LRS2-BBC and CAS-VSR-S101. We implement our models with PyTorch [48], PyTorch Lightning and FlashAttention-2 [20, 21]. We train all models using Automatic Mixed Precision (AMP) with the AdamW optimizer [37] with $\beta = (0.9, 0.98)$ and weight decay of 10^{-6} on servers with 8 GPUs (Geforce RTX 4090, NVIDIA A40 or A100 PCIe 40G). We use a cosine learning rate scheduler, warming up linearly for the first 30% of the training up to a maximum learning rate of 0.0005 (BASE & BASE*) / 0.0002 (LARGE) and decaying thereafter. We tune the maximum number of frames in a batch to fit on different devices; a typical setup is $6400/9600$ frames (256/384s) per device for pre-training and fine-tuning on $8\times$A100 40G. The number of masked input copies is set to 8 for the first stage, and 2 for the rest due to GPU memory constraints. The margin m in Eq. (12) is set to -0.2 [47]. Learnable temperatures τ are initialized to 0.07 and clipped to a minimum of 0.01. The codebook has $Q = 256$ codewords, and the codebook decay rate τ_{code} is set to 0.9. We set the number of target layers N to 8 for BASE and BASE* models [36], and 16 for LARGE models. During pre-training, we perform data augmentation in the form of horizontal flipping and random cropping to 88×88 consistently applied to all frames. A central crop is taken during evaluation. Filter banks are masked with a fixed random embedding, and video by random substitution with frames from the same segment [59]. Mask ratio is simply set to $M = 50\%$ for all input modalities and stages. We use LayerScale [69] with $\varepsilon = 0.1$ following [27] and LayerDrop [23] with per-layer probability $p = 0.05$ following [59]. For the audio inputs, we apply SpecAugment [46] in the frequency domain, with a maximum masking range of 40 filter-banks. We also apply adaptive time masking [27, 40] with a probability of 0.2 and mean duration of 0.2s per second to both audio and video using the fixed random embedding and zeros, respectively. We fine-tune for 75 epochs under a high-resource setup, and 50 epochs under a low-resource setup. We pre-train for 30k steps per stage and fine-tune for 50 epochs on CAS-VSR-S101. We do not freeze any layers during fine-tuning.

We want to reemphasize that none of the aforementioned hyperparameters are associated with *modality balancing* (*e.g.* modality input schedule, modality dropout, asymmetric mask ratios for the two modalities). Moreover, they are either directly taken from **existing literature** or chosen **empirically without further tuning**. The only exception is the fine-tuning learning rate, which we do find to be influential on downstream performance and therefore tune separately for VSR, ASR and AVSR. However, it is important to note that tuning the learning rate is a standard procedure in any neural network training. Despite the limited tuning performed and adopting the simple LF-MMI criterion without external language models for decoding, our results either match or rival the state-of-the-art performance on the evaluated datasets. We firmly believe that this alleviation in the burden of parameter tuning will prove invaluable to the community.

4. Low- and High-Resource Results for LRS2-BBC and LRS3-TED

In this section, we present more comprehensive results on LRS2-BBC and LRS3-TED with more existing works listed for comparison, as well as numbers obtained by decoding with larger external language models, which we unfortunately could not fit into the main text due to page limits.

4.1. Decoding with External Language Models

External language model for LRS2-BBC. For LRS2-BBC, we collect an in-house British TV program corpus coming from the same domain as LRS2-BBC totaling 117.4M words. Each sentence has a minimum of 4 words and a maximum of 100. The texts have been filtered by checking for overlap with the LRS2-BBC development and test sets. We then construct an external language model by combining dataset transcripts with the corpus.

External language model for LRS3-TED. For LRS3-TED, we leverage the TED-LIUM 3 corpus [29] transcripts with 4.9M words. We aggressively filter the text by checking for overlap with the LRS3-TED development and test sets. The perplexities of the language models constructed from the expanded textual corpora and the original LRS2-BBC and LRS3-TED datasets are provided in Tab. 6.

Decoding. With LF-MMI training, we follow common practice and perform HLG decoding with beam search on a 3-gram decoding graph, using a fixed beam size of 50. For English datasets, we additionally re-score the lattice using a 4-gram language model (trained on the same dataset).

[4] https://github.com/csukuangfj/kaldifeat

Corpora	LRS2-BBC		LRS3-TED	
	3-gram	*4-gram*	*3-gram*	*4-gram*
Dataset Only	218.15	211.66	197.54	191.80
Expanded	222.53	204.20	150.66	134.17

Table 6. **Perplexities of our *n*-gram language models for LRS2-BBC and LRS3-TED.** The numbers are computed on the corpora used to train the respective language models.

We tune language model scale among $\{0.1, 0.2, \ldots, 2.0\}$ and word insertion penalty among $\{0, 0.5, \ldots, 5.0\}$ on the validation sets. Note that this is done automatically through a quick grid search and considered common practice when evaluating CTC-like models following speech representation pre-training [10, 56].

4.2. Complete Comparison

As can be seen in Tabs. 7 to 9, decoding with a larger external language model leads to significant performance improvements and even SoTA results, especially for VSR, where the external LMs improve WER by about 2%. For ASR and AVSR, on LRS3-TED, we observe that with the external LM our low-resource models can be as good as a high-resource model without external LM. We therefore expect similar better performance if we move to a more complex Encoder-Decoder model, which is known to build strong internal language models, or hybrid CTC+CE models. Finally, it is worth noting that with a strong external LM, we observe diminishing returns by scaling up the Transformer on LRS3-TED under a high-resource setup, particularly for VSR and AVSR. As mentioned in the main text, this points to the limited modeling capacity of our *lightweight* visual encoder φ_v. Adopting more complex modeling architectures can be explored in future work.

4.3. Analysis of VSR Transcription Errors

We provide a few examples of VSR errors on LRS2-BBC in Tab. 10. The quality of the transcriptions clearly improve as we scale up the model and include external language models. Interestingly, the BASE* model pre-trained on 223h of in-domain LRS2-BBC data decoded the last sentence better than the 433h LRS3-TED pre-trained BASE model, which we believe is due to a slight mismatch in terms of content and pronunciation habits (British vs general English).

5. Ethical Considerations

Our work involves audio-visual human speech, which raises ethical considerations regarding privacy and potential misuse of these models. We acknowledge these issues and emphasize the need for responsible data use and ethical considerations in the development and deployment of systems built on our work. The datasets that we use in this work are either publicly available or datasets that are licensed for academic use.

Table 7. **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. †: test-time augmentation. *: external language models.

Methods	Unlabeled AV data	Labeled Data	Encoder Size	Criterion	VSR	ASR	AVSR
Supervised							
Afouras et al. [3]†*	-	1519h	71M	CE	48.3	9.7	8.5
Yu et al. [75]	-	1519h	-	LF-MMI[1]	48.9	6.7	5.9
Ma et al. [39]*	-	223h	79M	CTC+CE	39.1	4.3	4.2
	-	380h			37.9	3.9	**3.7**
Ma et al. [40]*	-	223h	79M	CTC+CE	32.9	-	-
	-	380h			28.7	-	-
	-	818h			**27.3**	-	-
Prajwal et al. [52]†*	-	698h	32M	CE	28.9	-	-
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. [40]*	641h	818h	79M	CTC+CE	25.5	-	-
Prajwal et al. [52]†*	1204h[3]	1472h	32M	CE	22.6	-	-
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[4]	-	3.6[4]
VatLM [80]	1759h[5]	223h	107M	CE	**30.6**	-	2.9
RAVEn [27]	433h	223h	97M	CTC+CE	32.1	3.9	-
Pan et al. [45]	-[6]	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)**
ES^3 (ours)*	223h	28h	46M	LF-MMI	38.0 (36.5)	4.5 (3.8)	4.3 (3.5)
	433h	28h	102M		37.1 (35.8)	4.4 (3.6)	4.1 (3.3)
	223h	223h	46M		29.3 (28.0)	3.4 (2.5)	3.0 (2.3)
	433h	223h	102M		**28.7 (27.6)**	**3.0 (2.2)**	**2.8 (1.9)**
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)**
ES^3 (ours)*	433h	28h	317M	LF-MMI	35.0 (33.9)	4.0 (3.1)	3.8 (3.0)
	433h	223h	317M		**24.6 (23.7)**	**2.5 (1.9)**	**2.4 (1.8)**
AV-HuBERT [59, 60]	1759h	223h	325M	CE	25.5[4]	-	2.5[4]
VatLM [80]	1759h[5]	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] We consider the TEDx_{ext} dataset to be unlabeled, since the automatic captions have not gone through additional verification.

[4] Reproduced by Zhu et al. [80].

[5] Uses additional 3846h audio, 452h audio-text and 600M text data.

[6] Uses additional 60000h audio data and 1.28M unlabeled images.

Table 8. **Low-resource results on LRS3-TED.** We pre-train a Base and Large model with 433h unlabeled data. *: uses external language models.

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)**
ES^3 (ours)*	433h	30h	102M	LF-MMI	**43.9 (43.2)**	**3.0 (2.4)**	**2.8 (2.1)**
			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)**
ES^3 (ours)*	433h	30h	317M	LF-MMI	41.6 (40.7)	**2.7 (2.1)**	**2.3 (1.7)**

[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, leading to better fine-tuning results even with its base model AV-HuBERT (47.1%).

Table 9. **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.

Methods	Year	Unlabeled AV data	Labeled Data	Backbone	Encoder Size	Criterion	VSR	ASR	AVSR
				Supervised					
Afouras et al. [3]†*	2018	-	1519h	Transformer	71M	CE	58.9	8.3	7.2
Xu et al. [74]	2020	-	590h	RNN	-	CE	57.8	7.2	6.8
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	-	-
		-	438h				37.9	-	-
Ma et al. [40]*	2022	-	595h	Conformer	79M	CTC+CE	35.1	-	-
		-	818h				**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	-	-
Prajwal et al. [52]†*	2022	1204h[2]	1472h	Transformer	32M	CE	30.7	-	-
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[3]
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**
ES^3 (ours)*	2023	433h	433h	Transformer	102M	LF-MMI	**37.9 (37.0)**	**2.5 (1.9)**	**2.0 (1.4)**
				Self-supervised (Large Models)					
AV-HuBERT [59, 60]	2022	433h	433h	Transformer	325M	CE	41.6	2.7	2.5[3]
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)**
ES^3 (ours)*	2023	433h	433h	Transformer	317M	LF-MMI	37.7 (37.1)	**2.4 (1.9)**	**2.1 (1.6)**

[1] Uses an additional ASR model trained on LibriSpeech (960h).
[2] We consider the TEDx$_{\text{ext}}$ dataset to be unlabeled, since the automatic captions have not gone through additional verification.
[3] Reproduced by Lian et al. [34].

Model	Transcription
Ground Truth	sort of second half of october
LR, Base*, 223h	should have sent out for october
HR, Base*, 223h	so you've set up of october
HR, Base, 223h	so you've sent of october
HR, Base, 223h, + extLM	sort of sendoff october
HR, Large, 223h, + extLM	sort of second half of october
Ground Truth	it's not all about size
LR, Base*, 223h	it's ~~not~~ all about size
HR, Base*, 223h	it's ~~not~~ all about starts
HR, Base, 223h	it's not all about stars
HR, Base, 223h, + extLM	it's not all about stars
HR, Large, 223h, + extLM	it's not all about size
Ground Truth	the garrison tried to surrender before he could try out his new toy
LR, Base*, 223h	the garrison tried to surrender before he can try out his new toy
HR, Base*, 223h	the garrison tried to surrender before he could try out a new toy
HR, Base, 223h	the garrison tried to ventilate before he can try out his new toy
HR, Base, 223h, + extLM	the carringtons tried to regulate before he could try out his new toy
HR, Large, 223h, + extLM	the garrison tried to surrender before he could try out his new toy

Table 10. **VSR Transcription errors**. **Green**: correctly recognized words; **Red**: substitution errors; **Blue**: deletion errors.