AV-SUPERB: A MULTI-TASK EVALUATION BENCHMARK FOR AUDIO-VISUAL REPRESENTATION MODELS

Yuan Tseng[1], *Layne Berry*[2*], *Yi-Ting Chen*[3*], *I-Hsiang Chiu*[1*], *Hsuan-Hao Lin*[1*], *Max Liu*[1*], *Puyuan Peng*[2*], *Yi-Jen Shih*[1*], *Hung-Yu Wang*[1*], *Haibin Wu*[1*], *Po-Yao Huang*[4], *Chun-Mao Lai*[1], *Shang-Wen Li*[4], *David Harwath*[2], *Yu Tsao*[3], *Shinji Watanabe*[5], *Abdelrahman Mohamed*[6], *Chi-Luen Feng*[1], *Hung-yi Lee*[1]

[1] National Taiwan University, Taiwan [2] University of Texas at Austin, USA
[3] Academia Sinica, Taiwan [4] Meta AI [5] Carnegie Mellon University, USA [6] Rembrand
r11942082@ntu.edu.tw

ABSTRACT

Audio-visual representation learning aims to develop systems with human-like perception by utilizing correlation between auditory and visual information. However, current models often focus on a limited set of tasks, and generalization abilities of learned representations are unclear. To this end, we propose the AV-SUPERB benchmark that enables general-purpose evaluation of unimodal audio/visual and bimodal fusion representations on 7 datasets covering 5 audio-visual tasks in speech and audio processing. We evaluate 5 recent self-supervised models and show that none of these models generalize to all tasks, emphasizing the need for future study on improving universal model performance. In addition, we show that representations may be improved with intermediate-task fine-tuning and audio event classification with AudioSet serves as a strong intermediate task. We release our benchmark with evaluation code[1] and a model submission platform[2] to encourage further research in audio-visual learning.

***Index Terms*—** Audio-Visual Learning, Representation Learning, Evaluation, Self-Supervised Learning

1. INTRODUCTION

Emulating the seamless integration of multiple tasks in human cognition, such as spoken language comprehension, sound event detection, and visual object recognition has been a long-standing goal of computational research. Prior research demonstrates that the pretrain-then-finetune paradigm is an effective and scalable method of building multitasking algorithmic systems for speech [1, 2], audio [3, 4], and vision [5, 6]. In the pretraining stage, models can often learn meaningful representations from unlabelled data alone through optimization of contrastive, masked prediction, or other self-supervised loss functions. These pretrained representations can then be applied to diverse tasks just by fine-tuning minimal additional parameters.

In order to better measure progress in representation learning, previous works have established multitask benchmarks in speech [7, 8], audio [9], and vision [10, 11]. However, these benchmark predominantly evaluate performance in isolation within single modalities. This approach overlooks the inherent multimodal nature of human perception, which synergistically integrates auditory and visual cues [12, 13]. While audio-visual representation learning has made significant progress [14, 15, 16, 17, 18], the assessment of these models tends to be task-specific, leaving the broader generalization capabilities across various audio-visual challenges less understood. This complicates comparitive analysis of different models and training strategies, impeding the development of more robust and versatile audio-visual representation learning approaches.

To address this issue, we propose AV-SUPERB, a standardized benchmark for comprehensively evaluating representations across seven distinct datasets involving five speech and audio processing tasks. AV-SUPERB comprises of three tracks to assess audio, video, and audio-visual fusion representations. We envision that these distinct tracks will allow researchers in speech, audio, and video representation learning alike to compare learning strategies across models and modalities, enabling broader analysis of their effectiveness.

Our contributions are four-fold: (1) **Diverse-domain evaluation**: We propose the first audio-visual learning benchmark that encompasses multiple datasets and tasks, covering both speech and audio domains. (2) **Easy and reproducible benchmarking**: We release evaluation code and a dedicated model submission platform that ensures reproducible evaluation on dynamic Youtube datasets and reduces computational entry barriers. (3) **Intermediate-task fine-tuning**: Our work emphasizes the potential benefits of full fine-tuning on intermediate tasks for improving performance on out-of-domain downstream tasks. (4) **Layer-wise analysis**: We show that different layers contribute variably to task performance, suggesting that simply using representations of the final layer is suboptimal, motivating the weighted-sum evaluation approach.

2. RELATED WORK

Recognizing how the close relation between audition and vision facilitates multimodal human perception, many audio-visual datasets have been gathered for action recognition [19, 15, 20, 21], speech recognition [22, 23, 24], speaker recognition [25, 26], and a variety of other tasks to study audio-visual learning. However, most models are trained and evaluated on different datasets with different experiment settings, which increases comparison difficulty and obfuscates the broad applicability of proposed methods. Hence in the AV-SUPERB benchmark, we select a diverse set of datasets from multiple tasks to comprehensively compare works in audio-visual representation learning.

Past multitask benchmarks in speech [7, 8], audio [9], and video

[*] Equal contribution; sorted alphabetically

[1] https://github.com/roger-tseng/av-superb

[2] https://av.superbbenchmark.org

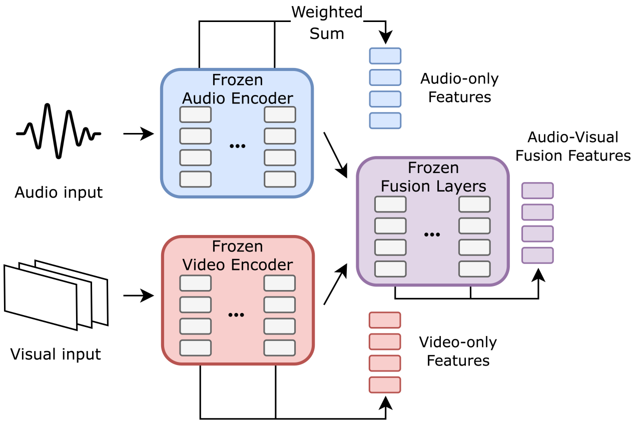

Fig. 1. We consider three evaluation scenarios: extracting features using inputs from one or both modalities. Following [7], the weighted-sum of features from Transformer layers (if applicable) are used as input for fine-tuning a small downstream model for each individual task. Details of selected tasks are given in Section 3.1.

representation learning [27, 10, 11] allow for fairer comparison of different models and promote research towards general approaches that are applicable to a variety of real-world tasks. SUPERB [7, 28] and SUPERB-SG [8] evaluate speech representation models on a wide range of downstream tasks covering content, speaker, and other different aspects of speech. Additionally, the HEAR benchmark [9] evaluates audio representations in diverse domains beyond speech, such as music and environmental sounds. For video representations, the SEVERE-benchmark [10] compares video self-supervised learning models on a diverse set of datasets to measure model sensitivity to different properties of downstream tasks. Feichtenhofer et al. [27] extend 4 image self-supervised learning methods to video representations and compare their efficacy on several downstream datasets, while Kumar et al. [11] focus on the effects of different factors in self-supervised video pretraining. However, these works focus on individual domains and cannot make use of the relationship between paired audio/visual inputs.

Previous multitask multimodal benchmarks focus on egocentric videos [29], vision-and-language domains [30, 31] or general multimodal learning[32]. In contrast, AV-SUPERB specializes in audio-visual tasks from speech and audio processing, allowing for more holistic assessment of representation models of audio and video alike.

3. BENCHMARK DETAILS

As shown in Figure 1, audio-visual models typically consist of two separate unimodal encoders followed by multimodal fusion layers. Based on this design, we setup three evaluation tracks in AV-SUPERB to benchmark representations from the two encoder and fusion layers, referred as audio-only, video-only, and audio-visual fusion features. This also allows for easy comparison with previous unimodal representation models.

Instead of striving for best possible performance for each task, the goal of our benchmark is to provide insight on the generalization capabilities of pretrained representations; therefore, we freeze the parameters of the task-invariant pretrained representation model (hereby referred as upstream model), and only fine-tune the parameters of the task-specific model (hereby referred as downstream model), following previous work [7]. Downstream models are designed to be simple and lightweight in order to purely evaluate representation abilities. Following the spirit of representation evaluation, we also limit hyperparameter tuning for downstream tasks. Although, we recognize that different representations may have vastly different loss landscapes, hence we search for the best performing learning rate from 10^{-1} to 10^{-5} in log-scale.

3.1. Downstream Task Selection

To keep computational costs reasonable, we mainly focus on utterance-level classification tasks in speech and audio processing, with the addition of ASR.

For audio processing, we select two audio classification tasks that highlight the relevance of different modalities, audio event classification (AEC) and action recognition (AR). Since audio events are often directly caused by actions, these tasks are complementary, and utilizing both audio and visual information can lead to better representations. This enables the possibility of learning better representations from multimodal input compared to unimodal baselines.

For speech processing, we select three audio-visual speech processing tasks where visual information is known to be beneficial [33, 34, 35], automatic speech recognition (ASR), automatic speaker verification (ASV), and emotion recognition (ER), in order to assess model capabilities on three fundamental aspects of speech: content, speaker, and paralinguistic information.

In designing the architecture for the downstream models, we generally follow the setup used for utterance-level tasks in the SUPERB benchmark. Specifically, the downstream model consists of a two-layer fully-connected network. This network takes the mean of features extracted from the frozen upstream model as input, and outputs class probabilities. However, as we also include the frame-level ASR task, we employ a two-layer BiLSTM model that takes the whole representation sequence as input and outputs characters.

3.2. Pretrained Upstream Models

To showcase the utility of our benchmark, we opt for the base version of four audio-visual upstream models, AV-HuBERT [36], RepLAI [37], Lee et al.'s model [38] (hereby referred as AVBERT throughout this paper), and MAViL [39]. These models were specifically chosen because they each excel at different tasks, underscoring the current gap in multi-tasking capabilities within existing audio-visual models. They vary substantially in terms of architecture, training objectives, and preprocessing techniques. We also conduct experiments on the base HuBERT [2] model, an unimodal speech representation model with similar design as AV-HuBERT, to make a fairer comparison between audio & audio-visual features.

Additionally, we incorporate two baselines that use handcrafted features as input for downstream models. Specifically, we employ log mel filterbank (FBANK) for audio and histogram of oriented gradients (HoG) for video, respectively.

4. EXPERIMENTAL RESULTS AND DISCUSSION

Previous work has shown that simply using representations extracted at the last layer of a frozen self-supervised model often results in suboptimal performance [40, 7]. Hence, we take a learnable weighted-sum of representations extracted over different Transformer layers as the final representation for each downstream task. For the audio-only and video-only tracks, only unimodal input and the relevant layers are used for extracting representations. For the audio-visual fusion track, both of the unimodal encoders plus fusion layers are used. As the size of representations extracted from fusion Transformer layers differ from those of unimodal layers, we take the weighted-sum for fusion Transformer layers only.

Representation Type	Params.	Overall Score	Audio-Visual				Speech-Visual		
			AEC		AR		ASR	ASV	ER
			AS-20K	VGGSound	Kinetics-Sounds	UCF101	LRS3-TED	VoxCeleb2	IEMOCAP
			(mAP ↑)	(Acc. ↑)	(Acc. ↑)	(Acc. ↑)	(CER ↓)	(EER ↓)	(Acc. ↑)
Audio-only									
FBANK	0	36.88	2.8	7.76	24.73	19.91	21.43	27.16	51.52
HuBERT	95M	53.66	14.3	30.21	51.46	36.06	**2.96**	15.58	**62.14**
AV-HuBERT*	90M	53.20	12.6	31.14	49.02	38.58	3.01	**14.45**	58.54
RepLAI	5M	39.70	12.3	27.01	45.90	33.85	66.09	32.58	57.53
AVBERT	10M	44.81	20.5	37.67	55.28	43.26	80.23	23.74	60.94
MAViL	86M	54.11	**21.6**	**39.91**	**57.28**	**45.68**	24.43	20.71	59.46
Video-only									
HoG	0	25.39	1.5	3.81	18.70	25.67	71.46	36.32	35.83
AV-HuBERT*	103M	33.48	2.4	5.90	24.73	37.55	**50.91**	**11.90**	26.59
RepLAI	15M	36.40	5.5	13.5	46.68	56.69	71.33	36.95	40.72
AVBERT	37M	47.69	11.5	28.73	62.67	77.42	72.29	20.00	**45.8**
MAViL	87M	49.70	**18.0**	**32.08**	**74.01**	**79.37**	74.03	24.58	43.03
Audio-visual fusion									
AV-HuBERT	103M	53.42	13.3	32.69	52.23	41.46	**2.75**	**9.46**	46.45
AVBERT	43M	54.85	22.9	44.54	71.31	71.76	70.12	18.31	**61.87**
MAViL	187M	62.36	**26.7**	**47.22**	**79.51**	**77.98**	30.18	19.67	54.94

*In order to fairly compare HuBERT & AV-HuBERT, we set features of the opposing modality to 0 and extract features from the 12-layer fusion Transformer for audio-only and video-only tracks.

Table 1. Main results. Best results for each track are highlighted in bold. Second-best results are underlined. We observe that MAViL excels at audio processing tasks, while HuBERT and AV-HuBERT are better for speech processing tasks.

4.1. Downstream Datasets and Training Details

Evaluation results for the three tracks are given in Table 1. For AEC, we evaluate on AudioSet [41] and VGGSound [42], and for AR, we select Kinetics-Sounds [15] and UCF101 [20]. Notably, in VGGSound and Kinetics-Sounds, audio and visual information are more correlated. This is reflected in our results, as audio-visual fusion results in larger gains compared to AudioSet and UCF101. We report testing set mean average precision for multi-label classification on AudioSet, and accuracy for the remaining three datasets.

For speech processing, we choose LRS3-TED for ASR, VoxCeleb2 for ASV, and IEMOCAP for ER. For ASR, we optimize CTC loss for character-level ASR, and report character error rate. For ASV, we first train for speaker identification on a subset of the dev split, then calculate cosine similarity to do verification on the test split and report equal error rate. For ER, we follow the conventional evaluation policy of removing unbalanced classes to perform four-way classification (neutral, happy, sad, angry) and report accuracy. Additional details related to datasets and training are given on our submission platform[2].

4.2. Overall Results

We find that existing models generally obtain large gains over handcrafted features, yet none of the five models tested were able to outperform all others in every task. To gauge universal performance across tasks, we provide an overall score calculated as the mean of either task-specific accuracies or the complement of error rates.

For the three speech processing tasks (ASR, ASV, ER), AV-HuBERT performs the best on ASR and ASV, and HuBERT achieves superior performance on ER. Notably, the unimodal HuBERT scores competitively on ASR and ASV as well, despite not being trained to utilize any visual grounding information.

For the four audio processing datasets, MAViL and AVBERT consistently outperforms all other models in all three tracks. We hypothesize that this is largely due to the diversity and large size of AudioSet data used for pretraining. Despite the domain mismatch, AVBERT also performs competitively for the ASV and ER speech tasks, especially in the audio-visual fusion track.

However, MAViL and AVBERT cannot perform ASR well, as simply using handcrafted FBANK features achieves lower error rates. Comparing their scores in the audio-only and fusion tracks, we see that their fusion layers are unable to effectively utilize the additional lip reading information, as performance is reduced when video is provided.

4.3. When does Visual Grounding Improve Audio Representation Learning?

Compared to unimodal audio representation models, audio-visual models may take advantage of information learned from visual grounding to improve audio representations even when only audio input is available at inference. Of the five selected models, HuBERT and AV-HuBERT use similar architectures and optimize the same masked cluster prediction objective using k-means clusters of MFCC features as initial targets. Although HuBERT is only trained on unimodal speech data, AV-HuBERT is trained to predict *multimodal* cluster targets obtained from both audio and visual modalities. By comparing their results on the audio-only track, we see that visual grounding information from multimodal cluster prediction improves representations for VoxCeleb2, VGGSound and UCF101.

		Audio-Visual				Speech-Visual		
	Intermediate Task Fine-tuning Data	AEC		AR		ASR	ASV	ER
		AS-20K (mAP ↑)	VGGSound (Acc. ↑)	Kinetics-Sounds (Acc. ↑)	UCF101 (Acc. ↑)	LRS3-TED (CER ↓)	VoxCeleb2 (EER ↓)	IEMOCAP (Acc. ↑)
AV-HuBERT								
Audio	LRS3-TED (video-text pairs)	12.6(-0.6)	22.83(-8.31)	38.19(-10.83)	28.70(-9.88)	13.89(-10.88)	22.38(-7.93)	53.92(-4.62)
Video		2.5(+0.1)	6.12(+0.22)	25.35(+0.62)	42.03(+4.48)	35.48(+15.43)	11.40(+0.50)	32.69(+6.10)
Fusion		5.1(-8.2)	17.11(-15.58)	38.52(-13.71)	40.74(-0.72)	22.66(-19.91)	11.35(-1.89)	43.58(-2.87)
MAViL								
Audio	AudioSet-2M	28.3(+6.7)	44.79(+4.89)	62.93(+5.65)	50.10(+4.42)	23.99(+0.44)	21.77(-1.06)	58.17(-1.29)
Video		20.9(+2.9)	36.68(+4.58)	77.39(+3.38)	86.93(+7.56)	78.59(-4.56)	23.93(+0.65)	39.15(-3.88)
Fusion		39.1(+12.4)	55.94(+8.72)	84.93(+5.42)	88.07(+10.09)	30.65(-0.47)	18.61(+1.06)	46.35(-8.59)

Table 2. Intermediate-task fine-tuning does not generally improve performance across all tasks. Results after intermediate-task fine-tuning (left) and absolute improvements compared to the original self-supervised model (right) are shown. Fine-tuning data for each model is color-coded to the corresponding downstream dataset.

4.4. Layer-wise Contribution Analysis

After fine-tuning the learnable weighted-sum over all upstream model layers on a downstream task, we may compare layer utilization by examining the weights of each layer in the weighted-sum. [43] Since the magnitude of representations from each layer may differ, we normalize layer weights for each layer by multiplying the weight with the L2-norm of representation values on the training set.

For MAViL, we find the layers that are commonly more dominant are the last three layers in the audio encoder, and the last two layers in the video encoder and fusion layers. Despite this, we observe an exception for emotion recognition on IEMOCAP. For IEMOCAP, the most dominant layer is the 0^{th} layer instead.

For AV-HuBERT, the final layer often contributes little. In the audio-only setup, we see that the layer with the most contribution is the penultimate layer for most speech and audio tasks besides ASR. For ASR, the last two layers are highly dominant on all three tracks. For non-ASR tasks, we note that when additional visual inputs are given, prior layers increase in contribution only when audio-visual fusion outperforms audio-only performance for AV-HuBERT (VGGSound, Kinetics-Sound, UCF101, VoxCeleb2), suggesting that prior layers in AV-HuBERT are more related to *visual* information, while the last few layers contain more *audio* information.

Overall, the variation in layer usage for different tasks, models, and modalities strongly motivates the use of the learnable weighted-sum technique for evaluation, instead of sub-optimally evaluating the final layer alone.

5. HOW DOES INTERMEDIATE-TASK FINE-TUNING AFFECT PERFORMANCE?

Studies in natural language processing show that pretrained language models can be improved by initial fine-tuning on an intermediate task, followed by further fine-tuning on the target task [44, 45].

In previous sections, we focus on assessing models pretrained in a self-supervised manner. However, model creators often release models variants that are fine-tuned further for performing specific downstream tasks. For example, MAViL adds 3 Transformer fusion layers after the audio and video encoders, and the whole model is fine-tuned on (audio&video, class) pairs for audio event classification. We hypothesize that these supervised models variants may provide improved representations for speech/audio tasks after intermediate-task training.

In order to support our hypothesis, we additionally evaluate fully fine-tuned variants of AV-HuBERT and MAViL on our benchmark, to determine when intermediate-task fine-tuning is beneficial. The variant of AV-HuBERT uses the same architecture, and is fine-tuned on 433 hours of (video, text) pairs from LRS3-TED to perform visual speech recognition, whereas the MAViL variant is fine-tuned on the entirety of AudioSet-2M. Experiment results are shown in Table 2.

For AV-HuBERT, we see that visual speech recognition on LRS3-TED is not a suitable intermediate task in general. Video-only representations obtain small gains in generalizability, at the cost of greatly reducing audio-only and fusion performance. We posit that intermediate-task fine-tuning with (video,text) pairs shifts AV-HuBERT Transformer layers to favor video input alone, reducing usability for audio-only and audio-visual inputs.

Contrarily, for audio-visual fusion with MAViL, we see that intermediate-task training on AudioSet-2M not only brings substantial improvements to all AEC and AR datasets, but also improves ASV while maintaining ASR performance. This suggests that fine-tuning on AudioSet-2M may be sufficiently diverse to improve speaker separability of representations without much loss of content information.

6. CONCLUSIONS

We introduce AV-SUPERB, the first benchmark for assessing general-purpose capabilities of audio-visual representations. AV-SUPERB includes a suite of 7 speech and audio processing datasets covering 5 audio-visual tasks. The benchmark is split into three tracks: two unimodal audio-only or video-only representations tracks, as well as a bimodal audio-visual fusion track. This enables easy comparison between unimodal and bimodal learning. Despite advances made in recent years, our experiments show that none of the models tested generalize to all tasks, leading us to conclude that further research is required to develop universal audio-visual models.

As discussed in Section 3.1, although our benchmark aims to comprehensively evaluate audio-visual models, only a limited set of tasks and datasets are included in its current form. For future work, we wish to incorporate more tasks relevant to additional facets of audio-visual processing, such as cross-modal retrieval, audio-visual localization, and sound/video generation, as well as improving the diversity and comprehensiveness of data sources.

7. REFERENCES

[1] Alexei Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[2] Wei-Ning Hsu et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, 2021.

[3] Daisuke Niizumi et al., "Byol for audio: Self-supervised learning for general-purpose audio representation," in *IJCNN*, 2021.

[4] Po-Yao Huang et al., "Masked autoencoders that listen," in *NeurIPS*, 2022.

[5] Alexey Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020.

[6] Kaiming He et al., "Masked autoencoders are scalable vision learners," in *CVPR*, 2022.

[7] Shu wen Yang et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech*, 2021.

[8] Hsiang-Sheng Tsai et al., "Superb-sg: Enhanced speech processing universal performance benchmark for semantic and generative capabilities," in *ACL*, 2022.

[9] Joseph Turian et al., "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*.

[10] Fida Mohammad Thoker et al., "How severe is benchmark-sensitivity in video self-supervised learning?," in *ECCV*, 2022.

[11] Akash Kumar et al., "Benchmarking self-supervised video representation learning," *arXiv preprint arXiv:2306.06010*, 2023.

[12] Harry McGurk and John MacDonald, "Hearing lips and seeing voices," *Nature*, 1976.

[13] Asif A. Ghazanfar and Charles E. Schroeder, "Is neocortex essentially multisensory?," *Trends in Cognitive Sciences*, 2006.

[14] Yusuf Aytar et al., "Soundnet: Learning sound representations from unlabeled video," in *NeurIPS*, 2016.

[15] Relja Arandjelovic and Andrew Zisserman, "Look, listen and learn," in *ICCV*, 2017.

[16] Bruno Korbar et al., "Cooperative learning of audio and video models from self-supervised synchronization," in *NeurIPS*, 2018.

[17] Humam Alwassel et al., "Self-supervised learning by cross-modal audio-video clustering," *NeurIPS*, 2020.

[18] Mandela Patrick et al., "Space-time crop & attend: Improving cross-modal video representation learning," in *ICCV*, 2021.

[19] Will Kay et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[20] Khurram Soomro et al., "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[21] Dima Damen et al., "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," *IJCV*, 2022.

[22] Martin Cooke et al., "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, 2006.

[23] Joon Son Chung et al., "Lip reading sentences in the wild," in *CVPR*, 2017.

[24] Triantafyllos Afouras et al., "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[25] A. Nagrani et al., "Voxceleb: a large-scale speaker identification dataset," in *Interspeech*, 2017.

[26] Joon Son Chung et al., "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018.

[27] Christoph Feichtenhofer et al., "A large-scale study on unsupervised spatiotemporal representation learning," in *CVPR*, 2021.

[28] Tzu-hsun Feng et al., "Superb@ slt 2022: Challenge on generalization and efficiency of self-supervised speech representation learning," in *SLT*, 2022.

[29] Kristen Grauman et al., "Ego4d: Around the world in 3,000 hours of egocentric video," in *CVPR*, 2022.

[30] Wangchunshu Zhou et al., "VLUE: A multi-task multi-dimension benchmark for evaluating vision-language pre-training," in *ICML*, 2022.

[31] Linjie Li et al., "Value: A multi-task benchmark for video-and-language understanding evaluation," in *NeurIPS Track on Datasets and Benchmarks*, 2021.

[32] Paul Pu Liang et al., "Multibench: Multiscale benchmarks for multimodal representation learning," in *NeurIPS Track on Datasets and Benchmarks*, 2021.

[33] G. Potamianos et al., "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, 2003.

[34] Arsha Nagrani et al., "Disentangled speech embeddings using cross-modal self-supervision," in *ICASSP*, 2020.

[35] Egils Avots et al., "Audiovisual emotion recognition in wild," *Machine Vision and Applications*, 2019.

[36] Bowen Shi et al., "Learning audio-visual speech representation by masked multimodal cluster prediction," in *ICLR*, 2022.

[37] Himangi Mittal et al., "Learning state-aware visual representations from audible interactions," in *NeurIPS*, 2022.

[38] Sangho Lee et al., "Parameter efficient multimodal transformers for video representation learning," in *ICLR*, 2021.

[39] Po-Yao Huang et al., "Mavil: Masked audio-video learners," in *NeurIPS*, 2023.

[40] Ankita Pasad et al., "Layer-wise analysis of a self-supervised speech representation model," in *ASRU*, 2021.

[41] Jort F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.

[42] Honglie Chen et al., "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020.

[43] Sanyuan Chen et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[44] Jason Phang et al., "Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks," *arXiv preprint arXiv:1811.01088*, 2018.

[45] Alex Wang et al., "Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling," in *ACL*, 2019.