SPEECHCLIP+: SELF-SUPERVISED MULTI-TASK REPRESENTATION LEARNING FOR SPEECH VIA CLIP AND SPEECH-IMAGE DATA

Hsuan-Fu Wang[1], *Yi-Jen Shih*[2], *Heng-Jui Chang*[3], *Layne Berry*[2], *Puyuan Peng*[2], *Hung-yi Lee*[1], *Hsin-Min Wang*[4], *and David Harwath*[2]

[1]National Taiwan University, Taiwan
[2]The University of Texas at Austin, USA
[3]Massachusetts Institute of Technology, USA
[4]Institute of Information Science, Academia Sinica, Taiwan

ABSTRACT

The recently proposed visually grounded speech model SpeechCLIP is an innovative framework that bridges speech and text through images via CLIP without relying on text transcription. On this basis, this paper introduces two extensions to SpeechCLIP. First, we apply the Continuous Integrate-and-Fire (CIF) module to replace a fixed number of CLS tokens in the cascaded architecture. Second, we propose a new hybrid architecture that merges the cascaded and parallel architectures of SpeechCLIP into a multi-task learning framework. Our experimental evaluation is performed on the Flickr8k and SpokenCOCO datasets. The results show that in the speech keyword extraction task, the CIF-based cascaded SpeechCLIP model outperforms the previous cascaded SpeechCLIP model using a fixed number of CLS tokens. Furthermore, through our hybrid architecture, cascaded task learning boosts the performance of the parallel branch in image-speech retrieval tasks.

Index Terms— SpeechCLIP, visually grounded, self-supervised learning, multimodal.

1. INTRODUCTION

Recent research in speech processing has focused on self-supervised learning (SSL), which pre-trains upstream models on different pretext tasks [1], such as generation or reconstruction [2, 3], contrastive learning [4, 5], prediction [6, 7], and knowledge distillation [8, 9]. These pre-trained models have been shown to outperform supervised models on certain downstream tasks, like speech recognition [6]. In addition to unimodal SSL methods, it may be beneficial to leverage different modalities, e.g., contextual image information can help speech models recognize similar-sounding words, thereby boosting the model's performance on speech recognition [10]. Therefore, some studies have utilized image-speech or image-text pairs for multi-modal training.

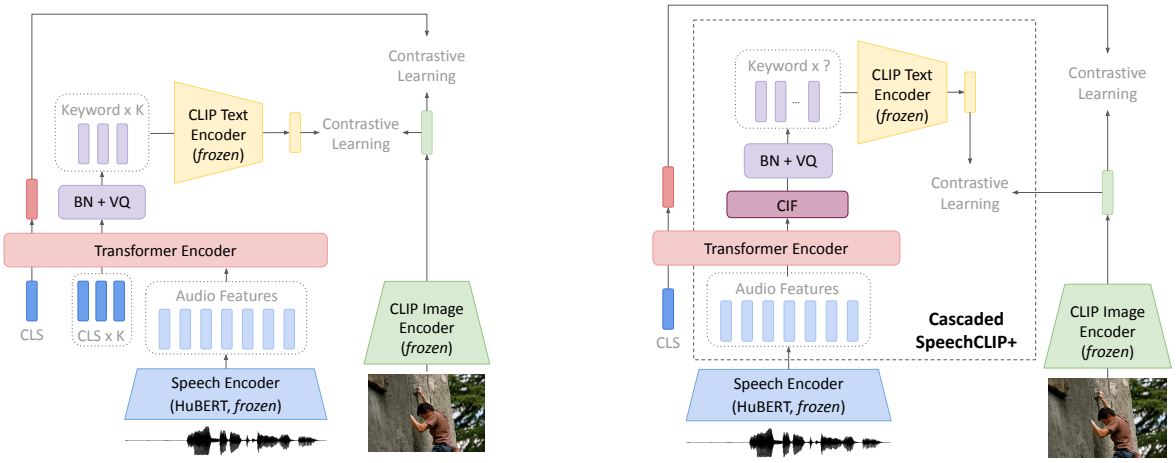
Speech processing models trained using image-speech paired data are known as Visually Grounded Speech (VGS) models. Many studies [11, 12] have found that these VGS models are beneficial for many tasks. The recently proposed SpeechCLIP [13] is a VGS model that leverages the pre-trained CLIP [14] image-text model and the speech SSL model HuBERT [6]. HuBERT is trained with a masked language modeling strategy, which provides good initialization for general speech processing tasks [15]. CLIP employs contrastive learning to pre-train robust image and text encoders by aligning semantically related images and text captions. By aligning speech-image pairs during training, SpeechCLIP learns to transform speech representations into the same embedding space as the pre-trained CLIP, thereby aligning speech and text together without transcriptional supervision. Moreover, M-SpeechCLIP [16], an extension of SpeechCLIP, demonstrates the capability of multilingual speech-to-image retrieval.

SpeechCLIP employs two architectures: parallel and cascaded. Parallel SpeechCLIP aligns semantically related images and spoken captions with utterance-level information, while cascaded SpeechCLIP aligns them with intermediately extracted subword-level information. Each architecture has its strengths: cascaded SpeechCLIP is capable of extracting a fixed number of keywords directly from speech without any text supervision and image tagging system. On the other hand, parallel SpeechCLIP excels at image-speech retrieval tasks. In this paper, we aim to further improve two types of architectures in terms of their strengths. Regarding the keyword extraction task, we have identified some issues in [13]. First, the K CLS tokens in the original cascaded architecture tend to attend to similar segments in the input speech, resulting in duplicate keywords in the output. Additionally, the fixed number of CLS tokens may not be flexible enough when the duration of the input speech varies significantly.

To address these issues, we apply the Continuous Integrate-and-Fire (CIF) [17] module to replace a fixed number of CLS tokens in the original cascaded architecture. We believe that CIF's monotonic alignment method will alleviate the issue of duplicate keywords. Furthermore, CIF enables the ability to output a dynamic number of keywords. The experimental results also manifest improvement when we do not consider duplicate keywords in the speech keyword extraction task. For the image-speech retrieval tasks, we posit that meaningful subwords' information could be beneficial in addition to the summary of the entire utterance. Thus, we propose a hybrid architecture, a multi-task learning framework that integrates the learning objectives of SpeechCLIP from both cascaded and parallel architectures. Experimental results manifest the improvements on Flickr8k [18].

2. METHOD

As with the training of SpeechCLIP [13], the input to the model is a batch of paired images and audio waveforms. Images are passed to CLIP's image encoder to extract image features, and audio waveforms are passed to a pre-trained SSL speech encoder. CLIP's text and image encoders are frozen during SpeechCLIP training. They serve as projectors of the pre-trained image-text semantic space. We follow SpeechCLIP by using HuBERT as our speech encoder. We



(a) hybrid SpeechCLIP

(b) hybrid SpeechCLIP+ with cascaded SpeechCLIP+

Fig. 1: Illustration of the proposed models. BN and VQ denote batch normalization and vector quantization processes, respectively. (a) In hybrid SpeechCLIP, the training loss combines the contrastive loss between the leftmost CLS token and the output image representation of the CLIP image encoder (the parallel branch same as parallel SpeechCLIP [13]) and the contrastive loss between the output speech representation of the CLIP text encoder for the remaining K CLS tokens and the output image representation of the CLIP image encoder (the cascaded branch same as cascaded SpeechCLIP [13]). (b) In cascaded SpeechCLIP+, instead of extracting keyword information through a fixed number of learnable CLS tokens, CIF is used to segment frame-level features into subword-level keyword sequences. In hybrid SpeechCLIP+, the parallel branch is based on parallel SpeechCLIP, and the cascaded branch is based on cascaded SpeechCLIP+.

also freeze it during model training but employ a set of learnable weights to perform a weighted sum of all its hidden states for audio feature extraction.

2.1. Continuous Integrate-and-Fire (CIF)

CIF is a soft, monotonic alignment method used to segment input sequential features with assigned weights. Its effectiveness has been demonstrated in previous speech segmentation tasks [19, 20]. Given sequential features $\mathbf{x} = (x_1, x_2, \cdots, x_T)$, CIF uses a convolution layer followed by a feed-forward layer with a sigmoid function to generate the corresponding weights $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_T)$. Segmentation boundaries are determined by accumulating α_t. As a result, CIF generates a sequence $\mathbf{c} = (c_1, c_2, \cdots, c_L)$ based on its segmentation and aggregation process.

During training, for each input speech, a target segmentation number L is required, which determines the desired output length of $\mathbf{c}$. Therefore, in [17], the following quantity loss $\mathcal{L}_{\text{QUA}}$ is used to encourage CIF to produce the correct cumulative sum of $\boldsymbol{\alpha}$,

$$\mathcal{L}_{\text{QUA}} = \left| \sum_{t=1}^{T} \alpha_t - L \right|. \tag{1}$$

Additionally, since the segmentation number may sabotage the training stability, in [17], a scaling strategy is used to generate $\boldsymbol{\alpha}' = (\alpha'_1, \alpha'_2, \cdots, \alpha'_T)$, where

$$\alpha'_j = \frac{\alpha_j}{\sum_{t=1}^{T} \alpha_t} \times L, \tag{2}$$

for performing segmentation and aggregation. The scaling strategy ensures that the output length of $\mathbf{c}$ is adjusted to the desired L.

2.2. Cascaded SpeechCLIP+

As shown in Fig. 1b, in cascaded SpeechCLIP+, instead of extracting keyword information through a fixed number of learnable CLS tokens, we apply CIF mechanism to segment frame-level features into a subword-level keyword sequence. During training, we set the target length of the quantity loss in Eq. (1) to 5% of the input length, based on the average ratio between the length of the Byte-Pair Encoding (BPE) token sequence and the output feature length of HuBERT, as observed during our experimentation. Additionally, we apply the scaling strategy in Eq. (2) to enhance training stability in the first 5k steps. After the CIF mechanism, the same as cascaded SpeechCLIP, we employ batch normalization and vector quantization to generate BPE tokens. These tokens are then input into the CLIP text encoder to extract representations, which are used to generate representations for computing contrastive loss $\mathcal{L}_{\text{cascaded}}$ with image features. The overall loss $\mathcal{L}$ for training cascaded SpeechCLIP+ is the linear combination of $\mathcal{L}_{\text{cascaded}}$ and $\mathcal{L}_{\text{QUA}}$,

$$\mathcal{L} = \lambda_c \times \mathcal{L}_{\text{cascaded}} + \lambda_q \times \mathcal{L}_{\text{QUA}}, \tag{3}$$

where λ_c and λ_q are adjustable weights.

2.3. Hybrid SpeechCLIP

Two types of SpeechCLIP are expected to benefit speech encoders from semantically related images through different alignment methods. Parallel SpeechCLIP enables speech encoders to benefit from utterance summarization, while cascaded SpeechCLIP enables speech encoders to benefit from capturing subword-level information from utterances. The advantages of combining the two approaches are quite intuitive. As shown in Fig. 1a, there is a total $K + 1$ CLS tokens in hybrid SpeechCLIP, the parallel branch refers to the path with the leftmost single CLS, while the cascaded branch refers to the path with the remaining K CLS tokens. The parallel branch is the same as the parallel SpeechCLIP, the first CLS representation of the transformer encoder's output is used to compute the contrastive loss $\mathcal{L}_{\text{parallel}}$ with image features. The cascaded branch is also the same as the cascaded SpeechCLIP, the remaining K CLS representations of the transformer encoder's output will be batch-normalized to match the mean and variance of CLIP's BPE

token embeddings and vector-quantized into BPE tokens and then used as input for the CLIP text encoder. The output of the text encoder is used to compute the contrastive loss $\mathcal{L}_{\text{cascaded}}$ with image features. The parallel and cascaded branches are trained jointly, and the overall loss $\mathcal{L}$ is the linear combination of $\mathcal{L}_{\text{parallel}}$ and $\mathcal{L}_{\text{cascaded}}$,

$$\mathcal{L} = \lambda_p \times \mathcal{L}_{\text{parallel}} + \lambda_c \times \mathcal{L}_{\text{cascaded}}, \quad (4)$$

where λ_p and λ_c are adjustable weights.

2.4. Hybrid SpeechCLIP+

As shown in Fig. 1b, in hybrid SpeechCLIP+, the parallel branch is the parallel SpeechCLIP model, and the cascaded branch is the cascaded SpeechCLIP+ model. In the parallel branch, we apply one CLS token to compute the contrastive loss $\mathcal{L}_{\text{parallel}}$ with image features. In the cascaded branch, same as cascaded SpeechCLIP+, we apply CIF to segment frame-level features into a subword-level keyword sequence. During training, we set the target length of the quantity loss in Eq. (1) to 5% of the input length and apply the scaling strategy in Eq. (2) in the first 5k steps. After the CIF mechanism, we employ batch normalization and vector quantization to generate BPE tokens. These tokens are then input into the CLIP text encoder to extract representations, which are subsequently used to generate representations for computing the contrastive loss $\mathcal{L}_{\text{cascaded}}$ with image features. The parallel and cascaded branches are trained jointly with the overall loss $\mathcal{L}$ as follows,

$$\mathcal{L} = \lambda_p \times \mathcal{L}_{\text{parallel}} + \lambda_c \times \mathcal{L}_{\text{cascaded}} + \lambda_q \times \mathcal{L}_{\text{QUA}}, \quad (5)$$

where λ_p, λ_c, and λ_q are adjustable weights.

3. EXPERIMENTS AND RESULTS

3.1. Experimental setups

Datasets. Our models are trained on two image-audio datasets, namely the Flickr8k Audio Captions Corpus [18] and SpokenCOCO [21]. Each image in both datasets is paired with five spoken captions collected by humans reciting the corresponding text captions. The training, development, and test sets in Flickr8k contain 30k, 5k, and 5k utterances, respectively, while SpokenCOCO has 565k, 25k, and 25k utterances in its training, development, and test sets. Following SpeechCLIP [13], we use the Karpathy split for SpokenCOCO [21].

Compared models. In the following experiments, the baseline cascaded SpeechCLIP, the proposed cascaded SpeechCLIP+, cascaded SpeechCLIP in hybrid SpeechCLIP, and cascaded SpeechCLIP+ in hybrid SpeechCLIP+ are denoted as Cascaded, Cascaded+, Cascaded(h) and Cascaded(h)+, respectively. Parallel SpeechCLIP can be trained jointly with cascaded SpeechCLIP in hybrid SpeechCLIP or cascaded SpeechCLIP+ in hybrid SpeechCLIP+. The model trained in the former way is represented as Parallel(h) to distinguish it from the Parallel(h)+ model trained in the latter way. Both models are compared with the baseline parallel SpeechCLIP (denoted as Parallel). All models are trained using the base or large model size setting in [13]. When the large model size setting is used, the model is marked "Large", e.g., Cascaded+ Large.

Implementation details. We use the same transformer encoder and the CLIP [14] model as in SpeechCLIP [13]. The convolution layer of CIF consists of a single one-dimensional convolution with d_{model} channels, a stride of 1, and a kernel width of 3. Here, d_{model} is set to 768 for the base models and 1024 for the large models. The one-dimensional convolution is followed by a dropout with a probability of 0.5 and a ReLU activation function. Regarding the loss weights in Eqs. (3), (4), and (5), we set $\lambda_c = \lambda_p = 1.0$ and $\lambda_q = 0.25$. We employ the same optimization strategy as in SpeechCLIP [13]. All models are trained with a batch size of 256 using two NVIDIA RTX 3090 GPUs with 24GB of memory. The hybrid+ Large model trained on SpokenCOCO took 4 days to converge, while model training under other conditions took less than two days.

Table 1: BPE extraction performance on the Flickr8k and SpokenCOCO test sets.

	w/o stop words			w/ stop words		
Model	R	P	F1	R	P	F1
	Flickr8k					
Cascaded [13]	7.39	0.94	1.66	8.3	1.68	2.79
Cascaded+	**27.16**	**3.55**	**6.29**	**18.11**	**3.62**	**6.04**
Cascaded(h)	7.01	0.85	1.52	5.41	1.06	1.77
Cascaded(h)+	16.52	2.16	3.82	11.40	2.23	3.73
Cascaded Large [13]	5.27	0.75	1.31	14.84	3.10	5.13
Cascaded+ Large	**27.27**	**3.73**	**6.56**	**20.74**	**4.2**	**6.99**
Cascaded(h) Large	6.73	0.86	1.53	6.18	1.25	2.08
Cascaded(h)+ Large	21.69	2.92	5.15	15.48	3.08	5.14
	SpokenCOCO					
Cascaded Large [13]	2.08	0.39	0.65	12.96	2.89	4.72
Cascaded+ Large	20.30	2.74	4.83	18.42	3.39	5.73
Cascaded(h) Large	1.68	0.45	0.71	20.37	4.48	7.35
Cascaded(h)+ Large	**20.56**	**3.73**	**6.32**	**26.21**	**4.95**	**8.33**

3.2. Keyword extraction

As discussed in SpeechCLIP [13], we evaluate the keyword extraction ability of cascaded SpeechCLIP+ through qualitative and quantitative analyses. Although there are other some previous works [22, 23] on unsupervised Bag of Words prediction, it is challenging to compare them in this task because we use different levels of tokens (BPE tokens) for training, whereas they use word tokens. Moreover, they require an image tagging system, while we do not. So we only compare with [13] in this section.

For qualitative analysis, from the example from the Flickr8k test set in Fig. 2, we can observe the ability of cascaded SpeechCLIP+ to capture semantically aligned keywords and their corresponding boundaries in the input speech. For example, our model successfully extracted the words "MAN" and "ROCK" and found their fragments in the audio waveform. In addition, the extracted word "CLIMBING" is semantically correlated to "CLIMB".

For the quantitative analysis, the results are shown in Table 1. In this experiment, the top 5 BPEs closest in cosine similarity to each CLS token (for Cascaded and Cascaded(h)) or each CIF-segmented token (for Cascaded+ and Cascaded(h)+) are retrieved. Performance is evaluated in terms of recall, precision, and F1-score to the BPE sequence of the transcription of the spoken caption. Considering that stop words in a sentence usually have little impact on the semantics of the sentence, we provide complete evaluation results (see w/ stop words) and evaluation results that ignore stop words (see w/o stop words). Stop words are pronouns, articles, prepositions, and conjunctions. We use the stop word set from the nltk python package.

It is clearly seen from the table that the proposed Cascaded+ models significantly outperform their corresponding baseline Cascaded models (Cascaded+ vs Cascaded, Cascaded(h)+ vs Cas-

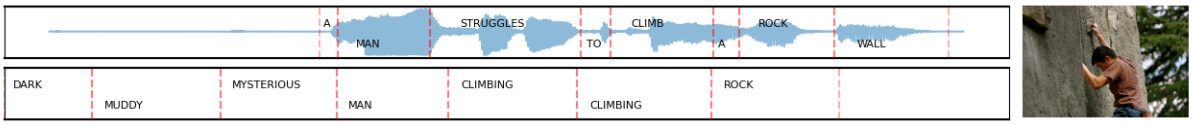


Fig. 2: An example of keywords extracted by Cascaded SpeechCLIP+ from the Flickr8k test set, showing the image, spoken caption, and extracted keywords with corresponding segments.

Table 2: Word and BPE extraction performance of Cascaded+ Large on the Flickr8k test set. "Type" refers to the granularity of the units considered, while "Top K" represents the top K BPEs retrieved by each CLS.

		w/o stop words			w/ stop words		
Type	Top K	R	P	F1	R	P	F1
Word	1	19.81	**12.06**	**14.99**	14.12	**13.66**	**13.89**
	2	22.80	6.84	10.53	16.54	7.86	10.65
	3	25.07	4.95	8.27	18.74	5.85	8.91
	4	26.76	3.92	6.83	19.97	4.61	7.49
	5	**27.59**	3.20	5.74	**20.86**	3.82	6.45
BPE	1	19.17	**12.99**	**15.49**	13.79	**13.97**	**13.88**
	2	22.29	7.53	11.26	16.29	8.25	10.96
	3	24.52	5.54	9.03	18.46	6.24	9.32
	4	26.31	4.46	7.63	19.77	5.01	7.99
	5	**27.27**	3.73	6.56	**20.74**	4.2	6.99

caded(h), Cascaded+ Large vs Cascaded Large, and Cascaded(h)+ Large vs Cascaded(h) Large). Surprisingly, joint training of parallel and cascaded branches did not always bring performance improvements to Cascaded and Cascaded+ models. The reason remains to be further studied.

Table 2 shows the performance of word extraction. Since the previous experiment shows that cascade SpeechCLIP+ Large (Cascaded+ Large) performs best on the Flickr8k test set, we use it in this experiment. We use the extracted neighboring BPEs to construct words. Top K BPEs with K from 1 to 5 are evaluated. The corresponding BPE extraction performance is also provided for reference. As can been seen from Table 2, Top 1 BPE gives the best F1 score for word extraction with or without considering stop words. An increase in K can slightly improve the recall rate (R), but decrease the precision rate (P). BPE extraction performance has the same trend as word extraction performance.

3.3. Image-Speech retrieval

Next, we evaluate Parallel(h)+ and Parallel(h) on image-speech retrieval tasks. The "Speech $\rightarrow$ Image" task is to retrieve the corresponding image given a spoken caption, and the "Image $\rightarrow$ Speech" task is to retrieve the corresponding spoken caption given an image.

From Table 3, we can see that on the Frickr8k dataset, the parallel branches in hybrid SpeechCLIP are mostly better than the corresponding baseline parallel models (Parallel(h) vs Parallel, Parallel(h)+ vs Parallel, Parallel(h) Large vs Parallel Large, and Parallel(h)+ Large vs Parallel Large). The results show that through a hybrid architecture, cascaded task learning improves the performance of parallel branches in image-speech retrieval tasks. The performance of Parallel(h) and Parallel(h)+ is comparable, suggesting that both cascaded SpeechCLIP and cascaded SpeechCLIP+ can effectively enhance the parallel branch in joint training. However, on the

Table 3: Image-speech retrieval performance on the Flickr8k and SpokenCOCO test sets. Parallel(h) refers to the parallel branch in hybrid SpeechCLIP (Fig. 1a), while Parallel(h)+ refers to the parallel branch in hybrid SpeechCLIP+ (Fig. 1b).

	Speech $\rightarrow$ Image			Image $\rightarrow$ Speech		
Model	R@1	R@5	R@10	R@1	R@5	R@10
	Flickr8k					
Parallel [13]	26.7	57.1	70.0	**41.3**	**73.9**	**84.3**
Parallel(h)	**29.8**	**60.8**	73.5	40.9	71.6	83.9
Parallel(h)+	29.3	60.1	**73.7**	39.7	72.8	83.4
Parallel Large [13]	39.1	72.0	83.0	**54.5**	84.5	93.2
Parallel(h) Large	**43.1**	**75.6**	**85.2**	54.3	85.1	93.5
Parallel(h)+ Large	41.7	73.7	84.1	54.2	**86.8**	**94.2**
	SpokenCOCO					
FaST-VGS$_{\text{CTF}}$ [24]	35.9	66.3	77.9	48.8	78.2	87.0
Parallel Large [13]	35.8	**66.5**	**78.0**	50.6	**80.9**	**89.1**
Parallel(h) Large	32.5	60.9	72.9	44.2	73.9	83.8
Parallel(h)+ Large	**36.5**	66.3	77.9	**51.0**	80.0	88.5

SpokenCOCO dataset, the performance of Parallel(h) Large and Parallel(h)+ Large is disappointing, especially Parallel(h) Large. The results in Table 1 show that Cascaded(h) Large and Cascaded(h)+ Large trained on SpokenCOCO have a relatively better ability to extract stop word information than models under other experimental settings. This may make Parallel(h) Large (or Parallel(h)+ Large) dominated by stop words captured by Cascaded(h) Large (or Cascaded(h)+ Large) jointly trained under the hybrid architecture.

4. CONCLUSIONS

In this paper, we attempt to boost the performance of pre-trained speech models on downstream tasks by leveraging visual content and other pre-trained modalities. We propose two extensions to SpeechCLIP. First, we apply the Continuous Integrate-and-Fire (CIF) module to replace a fixed number of CLS tokens in the cascaded architecture. Second, we propose a new hybrid architecture that merges the cascaded and parallel architectures of SpeechCLIP into a multi-task learning framework. In the keyword extraction task, our cascaded SpeechCLIP+ model significantly outperforms the baseline cascaded SpeechCLIP model [13]. Experimental results show that using CIF-segmented representations are more effective and flexible than adding a fixed number of CLS tokens when extracting subword and word information in speech. In the image-speech retrieval task, experimental results show that both cascaded SpeechCLIP and cascaded SpeechCLIP+ can effectively enhance the parallel branch in hybrid SpeechCLIP through joint training. In future work, we will investigate other unsupervised speech segmentation [19] and multi-task learning methods [25].

5. REFERENCES

[1] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaloe, et al., "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[2] Aaron Van Den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," in *NIPS*, 2017.

[3] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass, "An unsupervised autoregressive model for speech representation learning," in *Interspeech*, 2019.

[4] Alexei Baevski, Steffen Schneider, and Michael Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *ICLR*, 2020.

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

[6] 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 Transactions on Audio, Speech, and Language Processing*, 2021.

[7] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[8] Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee, "DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit BERT," in *ICASSP*, 2022.

[9] 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*, 2022.

[10] Felix Sun, David Harwath, and James Glass, "Look, listen, and decode: Multimodal speech recognition with images," in *SLT*, 2016.

[11] Puyuan Peng and David Harwath, "Word discovery in visually grounded, self-supervised speech models," *arXiv preprint arXiv:2203.15081*, 2022.

[12] Puyuan Peng and David F. Harwath, "Self-supervised representation learning for speech using visual grounding and masked language modeling," *arXiv:2202.03543*, 2022.

[13] Yi-Jen Shih, Hsuan-Fu Wang, Heng-Jui Chang, Layne Berry, Hung-yi Lee, and David Harwath, "SpeechCLIP: Integrating speech with pre-trained vision and language model," in *SLT*, 2023.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[15] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al., "SUPERB: Speech processing universal performance benchmark," in *Interspeech*, 2021.

[16] Layne Berry, Yi-Jen Shih, Hsuan-Fu Wang, Heng-Jui Chang, Hung-yi Lee, and David Harwath, "M-SpeechCLIP: Leveraging large-scale, pre-trained models for multilingual speech to image retrieval," in *ICASSP*, 2023.

[17] Linhao Dong and Bo Xu, "CIF: Continuous integrate-and-fire for end-to-end speech recognition," in *ICASSP*, 2020.

[18] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier, "Collecting image annotations using Amazon's Mechanical Turk," in *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[19] Yen Meng, Hsuan-Jui Chen, Jiatong Shi, Shinji Watanabe, Paola Garcia, Hung-yi Lee, and Hao Tang, "On compressing sequences for self-supervised speech models," in *SLT*, 2023.

[20] Zhiyun Fan, Linhao Dong, Meng Cai, Zejun Ma, and Bo Xu, "Sequence-level speaker change detection with difference-based continuous integrate-and-fire," *IEEE Signal Processing Letters*, 2022.

[21] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[22] Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu, "Visually grounded learning of keyword prediction from untranscribed speech," *arXiv preprint arXiv:1703.08136*, 2017.

[23] Leanne Nortje and Herman Kamper, "Towards visually prompted keyword localisation for zero-resource spoken languages," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023.

[24] Puyuan Peng and David Harwath, "Fast-slow transformer for visually grounding speech," in *ICASSP*, 2022.

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022.