Interspeech 2024
1-5 September 2024, Kos, Greece

Cross-Modal Denoising: A Novel Training Paradigm for Enhancing Speech-Image Retrieval

Lifeng Zhou, Yuke Li, Rui Deng, Yuting Yang, Haoqi Zhu*

Netease Yidun AI Lab, Hangzhou, China

{hzzhoulifeng,liyuke,dengrui01,yangyuting04,zhuhaoqi}@corp.netease.com

Abstract

The success of speech-image retrieval relies on establishing an effective alignment between speech and image. Existing methods often model cross-modal interaction through simple cosine similarity of the global feature of each modality, which fall short in capturing fine-grained details within modalities. To address this issue, we introduce an effective framework and a novel learning task named cross-modal denoising (CMD) to enhance cross-modal interaction to achieve finer-level cross-modal alignment. Specifically, CMD is a denoising task designed to reconstruct semantic features from noisy features within one modality by interacting features from another modality. Notably, CMD operates exclusively during model training and can be removed during inference without adding extra inference time. The experimental results demonstrate that our framework outperforms the state-of-the-art method by 2.0% in mean R@1 on the Flickr8k dataset and by 1.7% in mean R@1 on the SpokenCOCO dataset for the speech-image retrieval tasks, respectively. These experimental results validate the efficiency and effectiveness of our framework.

Index Terms: speech-image retrieval, cross-modal fine-grained alignment, cross-modal denoising

1. Introduction

By harnessing plentiful labeled data and computational resources, speech processing systems have demonstrated remarkable performance [1][2]. Unfortunately, the scarcity of labeled data for the majority of languages, coupled with the expensive nature of transcribing large volumes of speech data, has fueled a rising interest in the development of techniques capable of extracting valuable insights from unlabeled data [3][4].

Recently, there has been a notable emergence of self-supervised learning (SSL) methods as a prominent strategy for acquiring representations from unlabeled audio data, as evidenced by studies such as [5], [6], [7] and [8]. These methods have garnered attention for their effectiveness in this area, as demonstrated by the work of [1] and [9]. Furthermore, the exploration of multimodal data and the extraction of valuable information from it have been investigated as an alternative approach to improving the performance of speech processing systems. Pairing images with speech has been widely utilized to improve speech processing, ultimately resulting in the advancement of visually grounded speech (VGS) models, as exemplified in the work of [10]. These models have demonstrated their utility across a range of applications, such as speech recognition [11], [12], word discovery [13], and multilingual spoken language processing [14]. Typically, VGS models undergo training and evaluation in the context of speech-image retrieval tasks

The development of VGS models has significantly improved the accuracy of speech-image retrieval systems, highlighting the potential of speech-image retrieval as a standalone application. The FaST-VGS method [15] uses a unique training and retrieval approach, combining dual-encoder and cross-attention architectures. This enables a single model to achieve rapid and accurate speech-image retrieval capabilities. SpeechCLIP, as described in [16], leverages a speech encoder initialized with a pre-trained speech self-supervised learning (SSL) model [17], to align with a frozen CLIP image encoder using paired speech-image data. This alignment of the speech and image embedding spaces enables SpeechCLIP to perform state-of-the-art speech-image retrieval tasks.

While these methods have demonstrated effectiveness, they do have certain limitations. For example, in FaST-VGS, the utilization of an object detector as the image encoder may limit its expressive power, as it is constrained by the capabilities of the object detector and its predetermined visual vocabulary. SpeechCLIP replaces the object detector with an image encoder from CLIP to extract image features. It encodes speech and images separately, utilizing contrastive learning as the training objective. However, in contrastive learning, the interaction between modalities is managed solely through the cosine similarity of the speech and image features, which may pose challenges in achieving fine-grained alignment. As a result, this approach may lead to false positive matching during inference when images and speech share similar semantics but differ in details.

Therefore, designing a better interaction between modalities to facilitate fine-grained alignment between modalities is of crucial significance for the performance of speech-image retrieval systems. This paper introduces an innovative framework and a novel learning task cross-modal denoising (CMD) to enhance cross-modal interaction and achieve fine-grained cross-modal alignment. CMD is a denoising task designed to reconstruct semantic features from noisy features within one modality by interacting features from another modality. The objective of the CMD is to enhance speech representations, enabling them to focus on specific image-patch contexts, thereby achieving fine-grained cross-modal alignment.

Our main contributions can be summarized as follows:

- We propose a simple yet powerful framework with only 14M trainable parameters to achieve effective alignment between speech and image, ultimately leading to more accurate speech-image retrieval.
- We introduce a novel cross-modal learning task CMD to enhance cross-modal fusion, thereby achieving fine-grained cross-modal alignment. Importantly, CMD operates exclusively during model training and can be removed during inference without adding extra inference time.

*Corresponding author.

10.21437/Interspeech.2024-745

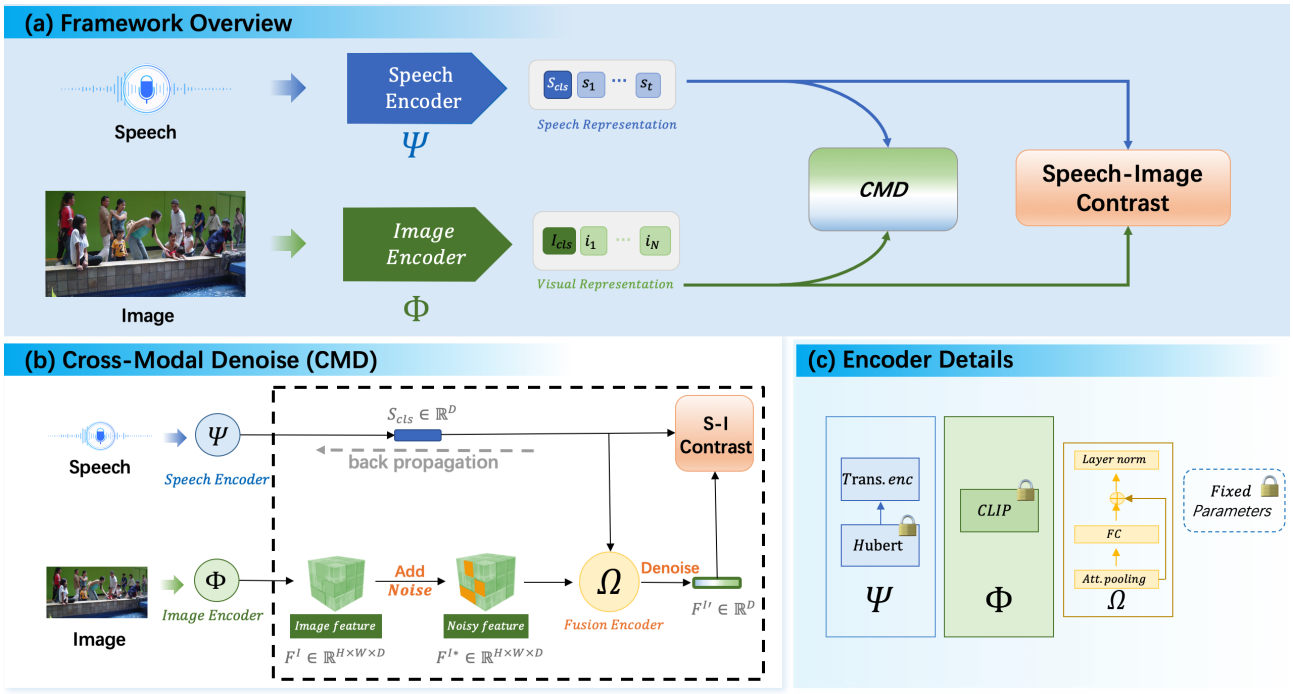

Figure 1: *The overview of our proposed framework. Figure (a) showcases that our framework is optimized with speech-image contrastive learning tasks and CMD tasks. Figure (b) provides details of the CMD tasks, while Figure (c) presents the specifics of the three encoders used in our framework.*

- Our framework has exhibited a significant improvement of 2.0% in mean R@1 on the benchmark dataset of Flickr8k Audio Capitions Coupus and 1.7% in mean R@1 on the SpokenCOCO dataset, surpassing the performance of the current state-of-the-art approach.

2. Methods

2.1. Preliminaries

In this section, we will provide a brief explanation of the two pre-trained models, HuBERT and CLIP, that are utilized in our framework.

Hidden-unit BERT (HuBERT) [18]. HuBERT is a self-supervised learning speech model that utilizes a masked prediction objective, similar to the renowned BERT [19] model. It predicts masked speech frames by considering the surrounding context. It comprises a CNN feature extractor followed by a transformer encoder. It can effectively extract valuable speech representations for various downstream tasks [20].

CLIP [21]. CLIP leverages contrastive learning to pre-train visual models at a large scale using natural language supervision [22][23][24], which is derived from paired image-text data. By employing two separate encoders for processing images and text, CLIP seeks to align semantically similar images and text captions. This enables CLIP to seamlessly transfer across a range of computer vision tasks with minimal supervision.

In our framework, pre-trained HuBERT and CLIP models are frozen and serve as feature extractors.

2.2. Architecture

As illustrated in Figure 1, we employ the speech encoder Ψ and the image encoder Φ to extract speech features $S = \{S_{cls}, s_1, ..., s_T\}$ and image features $I = \{I_{cls}, i_1, ..., i_N\}$, where S_{cls} and I_{cls} represent the normalized global speech and image semantic features, and s_i and i_i represent the frame-level speech feature and patch-level image feature. The framework is optimized with multitasks, including speech-image contrastive learning and CMD tasks. The speech-image contrastive learning task is designed to align the speech and image features at a coarse level. The CMD task aims to achieve fine-grained alignment between speech and image modality.

As shown (b) in Figure 1, CMD can be viewed as a combination of feature denoising and speech-image contrastive learning tasks. Specifically, we first introduce noise to the image features F^I to get noisy image features F^{I*}, where $F^I \in \mathbb{R}^{H\times W\times D}$ is reshaped from patch-level image features $i_1, ..., i_N$. The noise addition process is as follows: for a given image feature F^I, we randomly select certain patch features and then replace them with patch features from other image features within the same batch. Subsequently, the noisy image features F^{I*} along with the speech semantic features S_{cls} interact in the multimodal fusion encoder Ω to reconstruct image global semantic feature $F^{I'}$. This process can be presented as follows:

$$Attn(I^* \mid s) = Softmax\left(\frac{Q_s K_{I^*}^T}{\sqrt{D}}\right) V_{I^*}, \quad (1)$$

$$F^{I'} = LN\left(FC\left(Attn\left(I^* \mid s\right)\right) + Attn\left(I^* \mid s\right)\right)^T, \quad (2)$$

where Q, K, V denote the query, key, and value embeddings, $Softmax$ refers to the normalization function, LN stands for the layer normalization layer, and FC represents a fully connected layer. The objective of the multimodal fusion encoder is to improve speech representations, enabling them to concentrate on specific image-patch contexts in order to achieve fine-grained cross-modal alignment. Lastly, we use speech-image contrastive learning to put paired speech semantic feature S_{cls} and denoised image semantic feature $F^{I'}$ close together in the latent space and to pull them apart from the other features.

Speech Encoder	Speech Encoder	Image Encoder	Fusion Encoder	Trainable Params	Total Params
HuBERT Large (316M)	Transformer encoder **(13.4M)**	VIT-L/14 (422M)	FC **(0.6M)**	**14M**	752M

Table 1: *The model details of our framework.*

The details of the three encoders used in our framework are shown (c) in Figure 1, the speech encoder Ψ comprise a self-supervised learning speech model HuBERT [18] and 1 layer transformer encoder. Inspired by SUPERB [20], we combine the CNN output of HuBERT and the hidden representations from its transformer encoder using learnable weights. This weighted sum of HuBERT's output forms a sequence of speech features. These speech features, along with the CLS token, are then fed to the transformer encoder to extract speech embeddings $S = \{S_{cls}, s_1, ..., s_N\}$. Additionally, we utilize the image encoder of CLIP [21] as Φ to extract image features. The multimodal fusion encoder Ω comprises an attentive pooling layer, a fully connected layer, and a layer normalization layer. Notably, the parameters of HuBERT in Ψ and Φ are fixed during training process.

2.3. Training Objectives

Our model is jointly trained with two main objectives: speech-image contrastive learning on the unimodal encoders and CMD on the multimodal fusion encoder.

Speech-Image Contrastive Learning aims to align the speech and the image features at a coarse level, making it easier for the multimodal fusion encoder to perform cross-modal learning. It learns a similarity function $s = S_{cls}^T I_{cls}$, such that parallel speech-image pairs have higher similarity scores compared to non-parallel pairs.

For each speech, we calculate the softmax-normalized similarity between the speech and image features as follows:

$$p_j^{\mathrm{s2i}}(S) = \frac{\exp\left(s\left(S, I_j\right)/\tau\right)}{\sum_{j=1}^{B} \exp\left(s\left(S, I_j\right)/\tau\right)}, \tag{3}$$

where τ is a learnable temperature parameter, B is the mini-batch size. For each image, the softmax-normalized image and speech similarity is calculated as:

$$p_j^{\mathrm{i2s}}(I) = \frac{\exp\left(s\left(I, S_j\right)/\tau\right)}{\sum_{j=1}^{B} \exp\left(s\left(I, S_j\right)/\tau\right)}. \tag{4}$$

Let $\boldsymbol{y}^{\mathrm{s2i}}(S)$ and $\boldsymbol{y}^{\mathrm{i2s}}(I)$ represent the ground-truth one-hot similarity, where negative pairs have a probability of 0, and the positive pair has a probability of 1. The speech-image contrastive loss is defined as the cross-entropy H between $\boldsymbol{p}$ and $\boldsymbol{y}$ as follows:

$$\mathcal{L}_{\mathrm{sic}} = \frac{1}{2}[\mathrm{H}\left(\boldsymbol{y}^{\mathrm{s2i}}(S), \boldsymbol{p}^{\mathrm{s2i}}(S)\right) + \mathrm{H}\left(\boldsymbol{y}^{\mathrm{i2s}}(I), \boldsymbol{p}^{\mathrm{i2s}}(I)\right)] \tag{5}$$

Cross-modal denoising is a denoising task which aims to align the speech and image features at a fine-grained level. As illustrated in Figure 1(b), CMD can be viewed as a combination of feature denoising and speech-image contrastive learning tasks. The training loss for CMD is similar with Equation 5 as follows:

$$\mathcal{L}_{\mathrm{cmd}} = \frac{1}{2}[\mathrm{H}\left(\boldsymbol{y}^{\mathrm{s2i}}(S), \boldsymbol{p}^{\mathrm{s2i}'}(S)\right) + \mathrm{H}\left(\boldsymbol{y}^{\mathrm{i2s}}(I), \boldsymbol{p}^{\mathrm{i}'2\mathrm{s}}(I)\right)]. \tag{6}$$

The full pre-training objective of our framework is donated as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{sic}} + \alpha\mathcal{L}_{\mathrm{cmd}}, \tag{7}$$

where α is a hyper-parameter used to balance $\mathcal{L}_{\mathrm{sic}}$ and $\mathcal{L}_{\mathrm{cmd}}$.

3. Experiment

3.1. Setup

Dataset. Our model is trained and evaluated with speech-image retrieval on Flickr8k Audio Captions Corpus [13] and SpokenCOCO dataset [27]. Each image in both datasets is paired with five spoken captions produced by humans uttering text captions. Flickr8k consists of 8k images and 46 hours of speech, while SpokenCOCO has 123k images and 742 hours of speech. Following FaST-VGS [15], we use the Karpathy [28] split for SpokenCOCO.

Setup. The speech encoder Ψ consists of HuBERT and a single-layer transformer encoder. The HuBERT model utilized in our experiments is HuBERT-Large, while the transformer encoder has eight attention heads, and the hidden dimension of the transformer encoder is the same as that of HuBERT. As for the CLIP image encoder Φ, we used ViT-L/14. Both the parameters of HuBERT and CLIP are kept frozen throughout the training process. Additionally, the input and output dimensions of the fully connected layer utilized in the fusion encoder Ω are both set at 768. For detailed model configurations, please refer to Table 1. During the noise addition process, we randomly select 30% of image patch-level features to add noise. Since the two datasets contain multiple speech for each image, we change the ground-truth label of contrastive learning to consider multiple positives during training, where each positive has a ground-truth probability of $1/n$, where n is the number of positive samples. All models are trained with Adam optimizer with a weight decay of 10^{-6}, batch size of 128, and 60k steps in total. The learning rate linearly increases to 10^{-4} in the first 4k steps and decreases to 10^{-8} afterward. All experiments are conducted on a machine with 8 32GB V100 GPUs. During inference, we remove the multimodal fusion encoder Ω and only compute the feature similarity score between speech and image semantic features S_{cls} and I_{cls} for all speech-image pairs.

Evaluation Metric. We select the widely used Recall at K (R@K) metric, where a higher value indicates better performance, to evaluate the cross-modal retrieval performance of our framework. We presented the results for both speech-to-image retrieval and image-to-speech retrieval.

3.2. Speech-Image Retrieval

In this section, we assess the performance of our framework in speech-image retrieval tasks, thereby demonstrating the effectiveness of our models in aligning speech with image features. The cross-modal retrieval performance of our method is presented in Table 2. In comparison to previous methods, we have achieved the best retrieval performance in both speech-to-image retrieval and image-to-speech retrieval tests. Our model has shown significant improvements over the previous best model [16], with increases of 2.0% in mean R@1, 2.4% in mean R@5, and 1.9% in mean R@10 on the Flickr8k dataset. Besides, our model has demonstrated improvements of 1.7% in mean R@1, 0.7% in mean R@5, and 0.7% in mean R@10 on the SpokenCOCO dataset. These improvements can be mainly attributed to the ability of our model, jointly trained with contrastive learning and CMD tasks, to not only identify the shared semantics between images and speech but also capture the subtle differ-

Method	Speech → Image			Image → Speech			Mean		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
					Flickr8k				
FaST-VGS$_{\text{CO}}$ [15]	26.6	56.4	68.8	36.2	66.1	76.5	31.4	61.3	72.6
FaST-VGS$_{\text{CTF}}$ [15]	29.3	58.6	71.0	37.9	68.5	79.9	33.6	63.6	75.5
MILAN [25]	33.2	62.7	73.9	49.6	79.2	87.5	41.4	71.0	80.7
Cascaded SpeechCLIP [16]	14.7	41.2	55.1	21.8	52.0	67.7	18.3	46.6	61.4
Parallel SpeechCLIP [16]	39.1	72.0	83.0	54.5	84.5	93.2	46.8	78.3	88.1
Ours	**40.7**	**75.1**	**85.8**	**56.8**	**86.2**	**94.2**	**48.8**	**80.7**	**90.0**
					SpokenCOCO				
ResDAVEnet [11]	17.3	41.9	55.0	22.0	50.6	65.2	19.65	46.3	60.1
FaST-VGS$_{\text{CO}}$ [15]	31.8	62.5	75.0	42.5	73.7	84.9	37.2	68.1	80.0
FaST-VGS$_{\text{CTF}}$ [15]	35.9	66.3	77.9	48.8	78.2	87.0	42.4	72.3	82.5
Cascaded SpeechCLIP [16]	6.4	20.7	31.0	9.6	27.7	39.7	8.0	24.2	35.4
Seg. SpeechCLIP [26]	28.2	55.3	67.5	28.5	56.1	68.9	28.4	55.7	68.2
Parallel SpeechCLIP [16]	35.8	66.5	78.0	50.6	80.9	89.1	43.2	73.7	83.5
Ours	**37.5**	**67.3**	**78.6**	**52.3**	**81.4**	**89.7**	**44.9**	**74.4**	**84.2**

Table 2: *Recall scores for speech-image retrieval on Flickr8k and SpokenCOCO testing sets.*

Method	Speech → Image			Image → Speech			Mean		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
supervised	40.7	75.1	85.8	56.8	86.2	**94.2**	48.8	80.7	90.0
zero-shot	**48.9**	**78.3**	**87.5**	**61.4**	**88.2**	93.8	**55.1**	**83.2**	**90.7**

Table 3: *Recall scores for zero-shot speech-image retrieval on Flickr8k testing sets.*

ences between them.

3.3. Zero-Shot Speech-Image Retrieval

In order to evaluate the generalization ability of our framework, we performed zero-shot retrieval by directly assessing the model trained on SpokenCOCO on the testing sets of Flickr8K. To the best of the author's knowledge, this is the first time the exploration of the generalization capability of the speech-image retrieval model has been proposed. The result is shown in Table 3, where *supervised* indicates the model trained on Flickr8k training sets. Surprisingly, the model trained on the SpokenCOCO training sets outperforms the model trained on Flickr8k training sets by a large margin. This demonstrates the excellent generalization ability of our model. The superior performance attributes to the model being trained on a larger SpokenCOCO dataset in comparison to the Flickr8k dataset. In other words, our model demonstrates good scalability. We strongly believe that training on a larger corpus will further enhance its generalization capabilities.

3.4. Ablation Studies

In this section, we conduct ablation studies and report the results in mean R@1 on two datasets for simplicity.

Effectiveness of CMD. Table 4 studies the effect of CMD on cross-modal retrieval. In comparison to training without CMD, the inclusion of the CMD training task resulted in a 2.3% improvement on the Flickr8k dataset and a 1.9% improvement on the SpokenCOCO dataset, indicating the effectiveness of CMD.

The balance hyper-parameter α. The hyper-parameter determines the weight of CMD task. To assess its impact, we explore various scale ranges for α within the interval of [0.0, 1.0] on two datasets. The results, depicted in Figure 2, indicate that the optimal value of α varies across different datasets.

Training task	Flickr8k	SpokenCOCO
w/o CMD	46.5	43.0
w/ CMD	**48.8**	**44.9**

Table 4: *Ablation study of CMD*

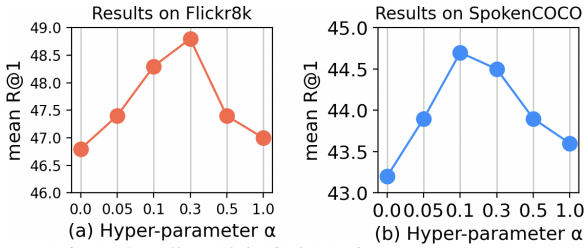

Figure 2: *Effect of the balance hyper-parameter α*

4. Conclusions

We propose a simple yet powerful framework to enhance the alignment between speech and image, thereby leading to more accurate speech-image retrieval. The framework is trained with cross-modal contrastive learning and cross-modal denoising (CMD) tasks. Specifically, CMD is a novel denoising task designed to enhance speech representations, enabling them to focus on specific image-patch contexts, thereby achieving fine-grained cross-modal alignment. Importantly, CMD operates solely during model training and can be removed during inference without adding any inference time. The experimental results demonstrate that our framework outperforms the state-of-the-art method by 2.0% in mean R@1 on the Flickr Audio Captions Corpus and by 1.7% in mean R@1 on the SpokenCOCO dataset for the speech-image retrieval tasks, respectively. These experimental results validate the efficiency and effectiveness of our framework. In our future works, we aim to continue advancing speech-image retrieval performance, as the accuracy of speech-image retrieval systems has lagged behind their image-text counterparts.

5. References

[1] Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *arXiv preprint arXiv:1911.08460*, 2019.

[2] Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, et al. Transformer-based acoustic modeling for hybrid speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6874–6878. IEEE, 2020.

[3] Leonardo Badino, Claudia Canevari, Luciano Fadiga, and Giorgio Metta. An auto-encoder based approach to unsupervised learning of subword units. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7634–7638. IEEE, 2014.

[4] Saurabhchand Bhati, Shekhar Nayak, and Kodukula Sri Rama Murty. Unsupervised speech signal to symbol transformation for zero resource speech applications. In *Interspeech*, 2017.

[5] Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

[6] Alexander H Liu, Yu-An Chung, and James Glass. Non-autoregressive predictive coding for learning speech representations from local dependencies. *arXiv preprint arXiv:2011.00406*, 2020.

[7] Jiawei Yao, Tong Wu, and Xiaofeng Zhang. Improving depth gradient continuity in transformers: A comparative study on monocular depth estimation with cnn. *arXiv preprint arXiv:2308.08333*, 2023.

[8] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Neural Information Processing Systems,Neural Information Processing Systems*, Jun 2020.

[9] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech 2019*, Sep 2019.

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

[11] Wei-Ning Hsu, David Harwath, and James Glass. Transfer learning from audio-visual grounding to speech recognition. In *Interspeech 2019*, Sep 2019.

[12] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, Oct 2020.

[13] David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 237–244. IEEE, 2015.

[14] David Harwath, Galen Chuang, and James Glass. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4969–4973. IEEE, 2018.

[15] Puyuan Peng and David Harwath. Fast-slow transformer for visually grounding speech. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7727–7731, 2022.

[16] 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. *IEEE SLT*, 2022.

[17] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, page 1505–1518, Oct 2022.

[18] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung 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*, PP:1–1, 10 2021.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, Jan 2019.

[20] 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, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. Superb: Speech processing universal performance benchmark. In *Interspeech 2021*, Aug 2021.

[21] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[23] 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*, pages 12888–12900. PMLR, 2022.

[24] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[25] Ramon Sanabria, Austin Waters, and Jason Baldridge. Talk, don't write: A study of direct speech-based image retrieval. *arXiv: Computation and Language,arXiv: Computation and Language*, Apr 2021.

[26] Saurabhchand Bhati, Jesús Villalba, Laureano Moro-Velazquez, Thomas Thebaud, and Najim Dehak. Leveraging pretrained image-text models for improving audio-visual learning. *arXiv preprint arXiv:2309.04628*, 2023.

[27] Wei-Ning Hsu, David Harwath, Tyler Miller, Christopher Song, and James Glass. Text-free image-to-speech synthesis using learned segmental units. In *Proceedings of the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Jan 2021.

[28] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 664–676, Apr 2017.