INTEGRATING SELF-SUPERVISED SPEECH MODEL WITH PSEUDO WORD-LEVEL TARGETS FROM VISUALLY-GROUNDED SPEECH MODEL

Hung-Chieh Fang[1], Nai-Xuan Ye[1*], Yi-Jen Shih[1,2], Puyuan Peng[2], Hsuan-Fu Wang[1], Layne Berry[2], Hung-yi Lee[1], David Harwath[2]*

[1] National Taiwan University, Taiwan
[2] The University of Texas at Austin, USA

ABSTRACT

Recent advances in self-supervised speech models have shown significant improvement in many downstream tasks. However, these models predominantly centered on frame-level training objectives, which can fall short in spoken language understanding tasks that require semantic comprehension. Existing works often rely on additional speech-text data as intermediate targets, which is costly in the real-world setting. To address this challenge, we propose Pseudo-Word HuBERT (PW-HuBERT), a framework that integrates pseudo word-level targets into the training process, where the targets are derived from a visually-ground speech model, notably eliminating the need for speech-text paired data. Our experimental results on four spoken language understanding (SLU) benchmarks suggest the superiority of our model in capturing semantic information.

Index Terms— Self-supervised learning, spoken language understanding, word segmentation, visual grounding

1. INTRODUCTION

Traditional supervised learning methods demand a substantial amount of labeled data, which is challenging to acquire in real-world scenarios, especially in speech processing. Recent research [1] has shifted the focus to self-supervised learning (SSL), which pre-trained an upstream model with unlabeled data to obtain the representations. These representations are subsequently leveraged to train downstream models for tasks with only a limited amount of labeled data.

One branch of SSL study [2] on speech is visually-grounded speech (VGS) models, where the models are trained with speech-image paired data to boost semantic information. VGS models have shown success in a wide range of tasks such as speech recognition [3, 4], speech generation [5], syllable and word discovery [6, 7, 8], keyword localization [9], and speech-image retrieval [10, 11, 12, 13, 14]. The recently proposed VG-HuBERT [8] is a VGS model that reaches state-of-the-art on the word segmentation task. For a given unlabeled utterance, a simple threshold is applied to the Transformer layer's attention weights to predict word segments. A K-Means clustering algorithm is further applied for segment categorization. Although VG-HuBERT has been shown to contain rich word boundary information in its representation [7], our experimental results show its limited performance on general SLU tasks.

While recent state-of-the-art SSL speech models [15, 16] achieve remarkable performance on various downstream tasks, it is worth noting that their training objectives are primarily frame-level. However, this may pose a challenge for SLU tasks, as such tasks often require more coarse-grained, word-level information.

_____
* Equal contribution.

To alleviate the issue, previous works introduce a framework comprised of an acoustic speech recognition (ASR) model and a natural language understanding (NLU) model. However, such an approach often requires paired transcripts, which are expensive to obtain. A recent work [17] proposes a textless approach that utilizes discrete units as intermediate targets to overcome the limitation. Notably, this method is specifically tailored for downstream models.

In this paper, we investigate the question of capturing semantic information in self-supervised speech models. We propose Pseudo-Word HuBERT (PW-HuBERT), a self-supervised model that is trained with pseudo word-level targets. In particular, we leverage word boundaries from word-segmentation models to generate pseudo word-level targets and apply them to HuBERT [15] pre-training. Our motivation is that word boundaries can aid foundation models in understanding the relationships between tokens within the same word. In contrast to the original HuBERT, where each target is independently assigned based on a fixed frequency, thus lacking contextual association.

We introduce two different architectures: single and hierarchical. The architecture of our single PW-HuBERT is similar to HuBERT. We add two extra transformer layers to predict the pseudo word-level targets generated from VG-HuBERT [8]. For the hierarchical PW-HuBERT, we apply the original HuBERT pretraining objective as an additional frame-level loss to further guide the training process. We evaluate our models on multiple SLU benchmarks and downstream tasks, including SLUE, SLUE Phase-2, SNIPS, and ZeroSpeech 2021 semantics, showing great improvement in the experimental results.

Our contributions can be summarized as follows:

- We are the first work to leverage an unsupervised word discovery model for speech SSL model pre-training.
- We demonstrate the benefits of joint training with frame-level and word-level units for capturing semantic information.

2. METHOD

2.1. Preliminaries

Hidden-unit BERT (HuBERT) [15]. HuBERT is a self-supervised model that simulates masked language modeling, where their targets are generated by clustering on acoustic features to provide frame-level supervision. HuBERT produces good feature representations for speech downstream tasks. In this paper, we use HuBERT-Base as our backbone model, which includes a CNN feature extractor and 12 transformer layers.

Visually-grounded HuBERT (VG-HuBERT) [8]. VG-HuBERT is trained with a generic image-speech contrastive loss, which exhibits state-of-the-art word discovery capabilities by applying a binary threshold to the self-attention map.
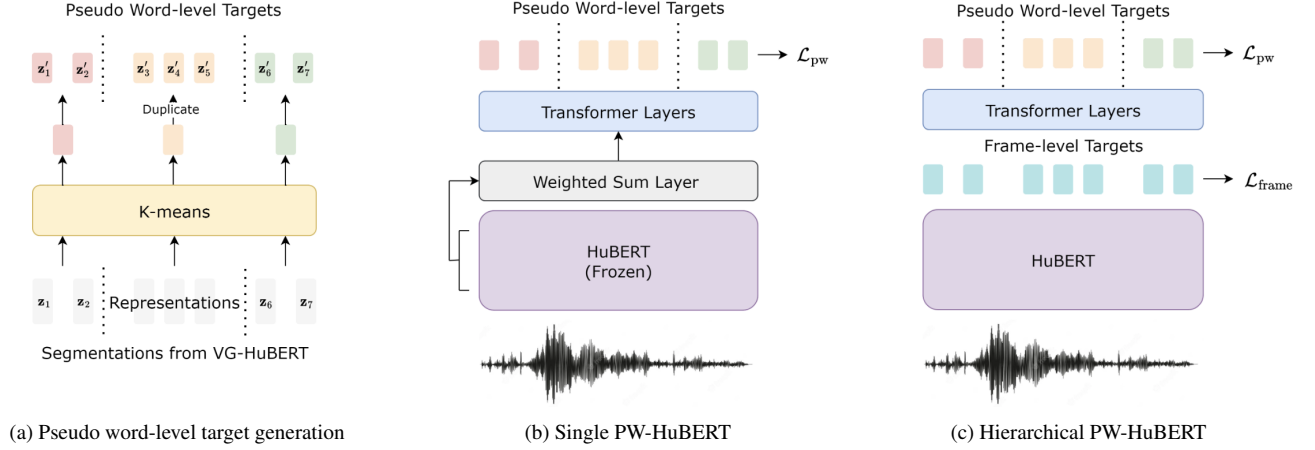
Fig. 1: Overview of pseudo word-level target generation and PW-HuBERT models. (a) With the word segmentations from VG-HuBERT [8], we apply mean-pooling on the representations within the same segment. A K-means model is used to predict the cluster ID of each segment, which is then duplicated to match the length of the corresponding segment. (b) The HuBERT weights are frozen, and the features from each layer are passed to a learnable weighted sum layer to get the final representation. The representation is then passed to another two transformer layers to predict pseudo word-level targets. (c) Hierarchical PW-HuBERT predicts frame-level targets at the 12th layer and predicts pseudo word-level targets at the 14th layer.

2.2. Word-Boundary and Pseudo Target Generation

In this section, we demonstrate our utilization of unsupervised word segmentation to generate pseudo word-level targets. Specifically, we leverage the state-of-the-art method VG-HuBERT for word discovery. It is worth highlighting that this approach can potentially be extended to various unsupervised word segmentation methods.

Fig. 1a illustrates the process of generating pseudo word-level targets. Given a speech utterance $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$ of T frames, we first apply VG-HuBERT to generate the feature sequence $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_T]$ and the attention boundaries $\mathbf{B} = [(s_1, e_1), \ldots, (s_u, e_u)]$, where u is the number of segments. Following the previous work [8], we do not directly use the attention segments as they tend to be narrower than the actual word segments. Instead, we use the midpoints between the adjacent attention boundaries as our pseudo word boundaries $\mathbf{B}' = [(s_1', e_1'), \ldots, (s_u', e_u')]$, where $s_i' = \frac{e_{i-1}+s_i}{2}$ and $e_i' = \frac{e_i+s_{i+1}}{2}$. (The first and the last stay the same: $s_1' = s_1, e_u' = e_u$)

We then apply mean pooling on features within the same segment to construct the pseudo word features [1]$\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_u]$, where $\mathbf{w}_i = \text{mean}(\mathbf{z}_{s_i}, \ldots, \mathbf{z}_{e_i})$. After that, We run the K-means clustering algorithm on $\mathbf{W}$ and assign cluster ID $\mathbf{z}_i'$ to each $\mathbf{w}_i$. In order to match the sequence length T of the input speech utterance, we simply duplicate each cluster ID to its corresponding length. e.g. the cluster ID for $\mathbf{w}_i$ will be duplicated $e_i - s_i + 1$ times. Lastly, we concatenate all the cluster IDs and form the pseudo word-level target sequence $\mathbf{Z}' = [\mathbf{z}_1', \ldots, \mathbf{z}_\mathbf{T}']$

2.3. Single PW-HuBERT

To distill word-level information in the HuBERT model, we propose Single PW-HuBERT illustrated in Fig. 1b. Specifically, we weighted-sum the representations in each layer of HuBERT (same

as [12, 18]) and feed it into two additional learnable transformer layers to learn the word level semantics. The HuBERT model is frozen to reduce computational cost and increase the training stability.

Following the training procedure of HuBERT, we define $\hat{\mathbf{X}}$ as the masked version of input speech sequence $\mathbf{X}$ and $\mathbf{M}$ as the masked indices generated randomly, where embeddings of $\hat{\mathbf{X}}$ at timestep $t \in \mathbf{M}$ are replaced with mask embedding. Our model f has to reconstruct $\mathbf{X}$ from $\hat{\mathbf{X}}$ by predicting probability $f(\cdot|\hat{\mathbf{X}}, t)$ over each masked timestep. We then calculate cross-entropy loss between target tokens $\mathbf{Z}'$ generated in section 2.2 and the model's prediction as $\mathcal{L}_{\text{pw}}$, which is defined as:

$$\mathcal{L}_{\text{pw}} = \sum_{t \in \mathbf{M}} \log f(\mathbf{z}_\mathbf{t}'|\hat{\mathbf{X}}, t) \tag{1}$$

Following the original HuBERT [15], we only calculate the loss on masked segments to be more resilient.

Overall, to retain the performance of HuBERT while incorporating word-level targets into the training objective to capture semantic information, we follow the pre-training convention [12, 18] to freeze HuBERT layers, and pre-trained two more layers on top of it with pseudo-word labels.

2.4. Hierarchical PW-HuBERT

To further explore the synergy between phone-level units (i.e. HuBERT units) and word-level units (i.e. pseudo units from section 2.2) in SSL pre-training, we propose Hierarchical PW-HuBERT to unify the two training objectives illustrated in Fig. 1c. The architecture of Hierarchical PW-HuBERT is similar to Single PW-HuBERT. We add two extra transformer layers after HuBERT-Base model. Compared with Single PW-HuBERT, we omit the weighted-sum layer and unfreeze the HuBERT model in Hierarchical PW-HuBERT.

Prior work [19] has demonstrated that higher-level information usually exists in deeper layers. Therefore, we inject objectives of different granularity on different layers. Specifically, the model aims to predict frame-level targets on the 12th layer and word-level targets

---

[1]Note that it does not necessarily map to real words; instead, it represents higher-level features compared to the original HuBERT.

Table 1: Performance of different models on SLUE, SLUE-2, ZeroSpeech 2021 semantics track, and SNIPS. [†] indicates that the model is trained with a similar task setup to ZeroSpeech semantics.

| Dataset | SLUE | | SLUE Phase-2 | SNIPS | | ZeroSpeech 2021 semantics | |
	SA	NER	NEL	SF	IC	librispeech	synthetic
Metric	F1 ↑	F1 / Label F1 ↑	Frame F1 / Word F1 ↑	F1 ↑		Similarity Judgement ↑	
HuBERT [15]	45.27	51.6 / 64.8	57.54 / 61.14	88.16	98.57	5.71	6.79
HuBERT$_{14}$	44.54	51 / 66.8	58.43 / 61.84	88.18	98.71	5.11	6.63
VG-HuBERT [8]	45.1	41.4 / 52.3	47.11 / 51.52	84.98	98.42	8.42[†]	9.97[†]
Single PW-HuBERT	48.7	52.5 / 67.3	59.44 / 63.51	88.32	98.44	5.16	6.88
Hierarchical PW-HuBERT	49.06	55.3 / 68.6	61.28 / 65.55	88.25	98.85	6.55	9.02

on the 14$^\text{th}$ layer, which aligns with the information scope of these layers.

Similar to Single PW-HuBERT, the pseudo word-level targets $\mathbf{Z}'$ are used to calculate $\mathcal{L}_\text{pw}$. Furthermore, we keep the original training loss in HuBERT [15] as frame-level objective $\mathcal{L}_\text{frame}$. The overall training loss $\mathcal{L}$ is the combination of $\mathcal{L}_\text{pw}$ and $\mathcal{L}_\text{frame}$:

$$\mathcal{L} = \mathcal{L}_\text{pw} + \lambda\mathcal{L}_\text{frame} \qquad (2)$$

where λ is a hyper-parameter. It's worth mentioning that the masked positions in $\mathcal{L}_\text{pw}$ and $\mathcal{L}_\text{frame}$ are generated independently to increase the diversity of training input.

3. EXPERIMENT

3.1. Datasets

To evaluate whether our model has a better understanding of higher level understanding of spoken languages, we run our models on a wide range of spoken language understanding benchmarks, including SLUE [20], SLUE Phase-2 [21], SNIPS [22], and ZeroSpeech 2021 [23].

SLUE is a benchmark for Spoken Language Understanding (SLU) evaluation on natural speech, which contains three tasks: sentiment analysis (SA), named entity recognition (NER), and automatic speech recognition (ASR). In this paper, we only focus on the SLU tasks: SA and NER. SA aims to classify the sentiment of a speech utterance as having negative, neutral, or positive sentiment, and the metric employed is macro-averaged F1 score. NER involves predicting the name entities and their corresponding types in a given speech sequence. The metrics involve F1 and Label F1, which evaluate entities and corresponding types respectively. Label F1 is useful to understand model accuracy despite the possible misspelling and segmentation errors in speech-to-text conversion.

SLUE Phase-2 is a complement of SLUE, which introduces four additional SLU tasks. We only evaluate on the named entity localization (NEL) task as the codebase and datasets of other tasks have not been published yet. The goal of NEL is to locate the position of named entities in the audio by predicting their start and end times. The task evaluates performance via frame F1 and word F1 scores based on the overlap between the predicted and ground truth time segments. We follow the official training set and evaluate on the development set for both SLUE and SLUE Phase-2, since the test sets are not publicly accessible.

ZeroSpeech 2021 is a benchmark designed to evaluate the model's performance on different linguistic levels, and we evaluate on the semantics task. Given pairs of spoken words, the task is to predict the similarity for each word pair and compare the correlation score between model prediction and human judgment.

SNIPS includes a collection of speech utterances along with corresponding annotations for slots and intents. We evaluate on the slot filling (SF) and intent classification (IC) tasks. SF aims to locate and classify entities within the input data. IC involves determining the user's intention given the speech utterance. For both tasks, we use F1 as the metric.

For SLUE, SLUE Phase-2, and SNIPS, we evaluate the tasks with SLUE toolkit[2]. For ZeroSpeech 2021 semantics, we follow the instruction of ZeroSpeech2021 challenge[3] to get the results.

3.2. Baselines

We compare our models with 3 different baselines: HuBERT [15], VG-HuBERT [8], and a 14-layer variant of HuBERT - HuBERT$_{14}$. To investigate whether the improvement of our model comes from pseudo word-level targets rather than additional parameters, we add an extra baseline of a 14-layer HuBERT on all tasks. Specifically, the training procedure is the same as the original HuBERT but with 14 transformer layers.

3.3. Implementation Details

For pseudo word-level target generation, we apply a threshold of 0.8 to the attention weights of layer 9 from VG-HuBERT$_3$[4] on LibriSpeech 100hr to get the word boundaries for our training dataset. Each target ID is obtained from running K-means clustering with 4096 clusters on the features extracted from the VG-HuBERT model and mean-pooled within the boundary. For frame-level target generation, we follow the same procedure of HuBERT [15] to get the labels by applying the K-means algorithm on the 6th transformer layer from the previous iteration model.

As for the training configuration, the total training steps of Single PW-HuBERT and Hierarchical PW-HuBERT are 500k steps on a single GPU and 125k steps on 4 GPUs respectively. We apply a linear scheduler with warmup steps of 32k and a peak learning rate of 1e-4. The two additional transformer layers are randomly initialized and have the same dimensions as the previous layers.

4. RESULTS AND ANALYSIS

4.1. Main Results

Table 1 provides an overview of our experimental results. We can observe that both Single and Hierarchical PW-HuBERT show significant improvement over baseline models in SLUE and SLUE Phase-2, indicating the effectiveness of leveraging pseudo word-level targets. In SNIPS, the hierarchical PW-HuBERT consistently

---

[2] https://github.com/asappresearch/slue-toolkit
[3] https://zerospeech.com/challenge_archive/2021/04_instructions
[4] VG-HuBERT$_x$ refers to reinitializing the weights of the last x transformer layers.

Table 2: The comparison with oracle setting, where we replace attention midpoint boundary with ground truth boundary and replace ID generated from clustering with ID from the BERT tokenizer.

Architecture	Boundary	Target ID	SLUE		SLUE 2	SNIPS
			SA	NER	NEL	SF
HuBERT	frame	cluster	45.27	51.6 / 64.8	57.54 / 61.14	_88.2_
Hierarchical	ground truth	word ID	_47.23_	50.5 / 63.7	55.5 / 57.46	87.4
	ground truth	cluster	44.49	**56.6 / 70.8**	**61.5 / 66.40**	88
	attention midpoint	cluster	**49.06**	_55.3_ / 68.6	_61.28_ / _65.55_	**88.3**

improves on both SF and IC task, although the improvements are relatively modest. In ZeroSpeech 2021 Semantics, the hierarchical PW-HuBERT yields better performance than HuBERT and HuBERT$_{14}$ on both librispeech and synthetic subsets, which demonstrates our model has a better ability to capture semantic information. While our models exhibit lower performance than VG-HuBERT on both subsets, it is worth considering that VG-HuBERT has a similar training setup to ZeroSpeech semantics task, which calculates the similarities between embeddings. While this setup shows advantages on ZeroSpeech semantics task, it can be observed that it has a negative impact on general SLU tasks, including SLUE, SLUE Phase-2, and SNIPS.

Overall, our models' superior performance on SLUE, SLUE Phase-2, and SNIPS demonstrates their effectiveness in SLU downstream tasks. Furthermore, our improvements on ZeroSpeech 2021 semantics indicate the richness of word-level information within our pre-trained representations.

4.2. Comparison with Oracle Setting

In this section, we discuss the effect of introducing ground truth word boundary and word token ID into our model architecture. To do so, we have two different settings, one is to directly replace the word boundary generated from VG-HuBERT with ground truth boundary and run K-means clustering as before, and the other is not only to substitute word boundary but also skip the clustering method by assigning word ID from the BERT tokenizer to each token. Notice that under these two settings, the model has to predict IDs in 4096 and around 30000 different labels respectively. For these two experiments, we employ the same hyper-parameters and model architecture as the Hierarchical PW-HuBERT model.

As shown in Table 2, the improvement of using ground truth boundary is limited and unstable. We hypothesize that utilizing boundaries from attention weights provides the model with more precise locations to focus on. This is attributed to the attention-midpoint boundary typically narrower than the ground truth, consequently contracting more useful information for both K-means clustering and our model. Our findings contradict with [24], who show that LSTM models with boundaries predicted by DP-Parse or gold boundaries perform worse than LSTM models without any boundaries[5]. We think the disparity arises from the usage of word boundaries. Specifically, we aggregate the information within the word boundaries, whereas [24] simply adds a ⟨SPACE⟩ token between words.

As for the worse performance on oracle boundaries and oracle word IDs, we suppose that IDs from clustering results have more semantic relations between each ID entry than ground truth IDs.

---

[5]We consider models training with phone unit from table 5 of [24], as it is similar to our setting.

4.3. Ablation Studies

4.3.1. The Effect of Frame-level Targets

We investigate the effect of incorporating frame-level targets with pseudo word-level targets. We conduct an analysis between Hierarchical PW-HuBERT and its counterpart without frame-level targets. The latter setting is equivalent to Single PW-HuBERT but the weighted-sum layer is omitted and the HuBERT weight is unfrozen.

The results in Table 3 show that the performance consistently improves on almost all datasets when integrating frame-level targets. This suggests that leveraging the synergy between frame-level and word-level targets can further provide guidance for the training process.

Table 3: Ablation studies of hierarchical PW-HuBERT on frame-level supervision.

Architecture	HuBERT Targets	SLUE		SLUE 2	SNIPS
		SA	NER	NEL	SF
Hierarchical	✗	44.94	53.6/67.9	59.4	**88.4**
	✓	**49.06**	**55.3/68.6**	**61.3**	88.3

Table 4: Ablation studies of single PW-HuBERT on freezing HuBERT weights.

Architecture	Freeze	SLUE		SLUE 2	SNIPS
		SA	NER	NEL	SF
Single	✗	44.94	**53.6 / 67.9**	59.44 / **63.51**	**88.4**
	✓	**48.7**	52.5 / 67.3	**59.48** / 62.56	88.3

4.3.2. Freezing HuBERT Weights of Single PW-HuBERT

In this section, we discuss the effectiveness of freezing HuBERT weights in Single PW-HuBERT. As shown in Table 4, freezing HuBERT weights and adding a weighted-sum layer yields comparable performance with the results obtained when HuBERT weights are frozen across all datasets. Notably, this approach has better training efficiency, which prompts us to adopt freezing HuBERT weights as our final method for Single PW-HuBERT.

5. CONCLUSION

In this work, we address the challenge of insufficient semantic information in recent SSL models. We propose PW-HuBERT, a framework that incorporates pseudo word-level targets from a visually-grounded model into the training process without the necessity of speech-text paired data. Our experiments on several SLU tasks show the effectiveness of our models in semantic understanding. Furthermore, we demonstrate that oracle boundary may not be optimal, as

it potentially contains unuseful information. Finally, we show the benefits of joint training with frame-level targets and freezing Hu-BERT weights in the ablation study. In our future work, we plan to explore the impact of introducing intermediate-level units, such as syllable-level information, in speech SSL models and introduce the whole word masking strategy for PW-HuBERT.

6. ACKNOWLEDGEMENT

7. REFERENCES

[1] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al., "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[2] Grzegorz Chrupała, "Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques," *Journal of Artificial Intelligence Research*, vol. 73, pp. 673–707, 2022.

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

[4] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *International Conference on Learning Representations (ICLR)*, 2022.

[5] 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 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Aug. 2021.

[6] David Harwath and James Glass, "Towards visually grounded sub-word speech unit discovery," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3017–3021.

[7] Puyuan Peng, Shang-Wen Li, Okko Räsänen, Abdelrahman Mohamed, and David Harwath, "Syllable segmentation and cross-lingual generalization in a visually grounded, self-supervised speech model," in *Interspeech*, 2023.

[8] Puyuan Peng and David Harwath, "Word discovery in visually grounded, self-supervised speech models," in *Interspeech*, 2022.

[9] Leanne Nortje and Herman Kamper, "Towards visually prompted keyword localisation for zero-resource spoken languages," in *IEEE Spoken Language Technology Workshop (SLT)*, 2022, pp. 700–707.

[10] Ramon Sanabria, Austin Waters, and Jason Baldridge, "Talk, Don't Write: A study of direct speech-based image retrieval," in *Interspeech*, 2021.

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

[12] 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 *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 715–722.

[13] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] Saurabhchand Bhati, Jesús Villalba, Laureano Moro-Velazquez, Thomas Thebaud, and Najim Dehak, "Segmental SpeechCLIP: Utilizing Pretrained Image-text Models for Audio-Visual Learning," in *Proc. INTERSPEECH 2023*, 2023, pp. 431–435.

[15] 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*, vol. 29, pp. 3451–3460, 2021.

[16] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[17] Guan-Wei Wu, Guan-Ting Lin, Shang-Wen Li, and Hung-yi Lee, "Improving textless spoken language understanding with discrete units as intermediate target," *arXiv preprint arXiv:2305.18096*, 2023.

[18] 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 *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[19] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[20] Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han, "Slue: New benchmark tasks for spoken language understanding evaluation on natural speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7927–7931.

[21] Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan S Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe, "SLUE phase-2: A benchmark suite of diverse spoken language understanding tasks," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. July 2023, pp. 8906–8937, Association for Computational Linguistics.

[22] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël

Primet, and Joseph Dureau, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *ArXiv*, vol. abs/1805.10190, 2018.

[23] Tu Nguyen, Maureen de Seyssel, Patricia Roz'e, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux, "The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," *ArXiv*, vol. abs/2011.11588, 2020.

[24] Tu Nguyen, Maureen de Seyssel, Robin Algayres, Patricia Roze, Ewan Dunbar, and Emmanuel Dupoux, "Are word boundaries useful for unsupervised language learning?," *ArXiv*, vol. abs/2210.02956, 2022.