Interspeech 2022
18-22 September 2022, Incheon, Korea

Word Discovery in Visually Grounded, Self-Supervised Speech Models

Puyuan Peng, David Harwath

Department of Computer Science, The University of Texas at Austin

pyp@utexas.edu, harwath@utexas.edu

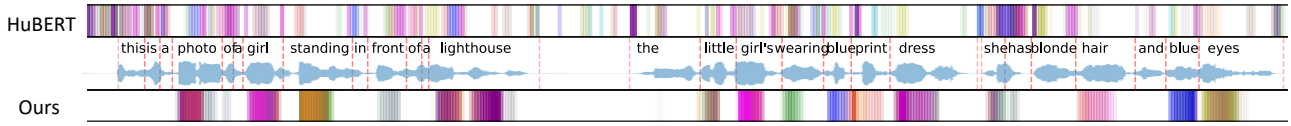

Figure 1: *HuBERT: sum of attention weights each frame receives from other frames. Ours (VG-HuBERT$_3$): attention weights each frame receives from the* `[CLS_A]` *token. Attention weights from different attention heads are coded with different colors.*

Abstract

We present a method for visually-grounded spoken term discovery. After training either a HuBERT or wav2vec2.0 model to associate spoken captions with natural images, we show that powerful word segmentation and clustering capability emerges within the model's self-attention heads. Our experiments reveal that this ability is not present to nearly the same extent in the base HuBERT and wav2vec2.0 models, suggesting that the visual grounding task is a crucial component of the word discovery capability we observe. We also evaluate our method on the Buckeye word segmentation and ZeroSpeech spoken term discovery tasks, where we outperform all currently published methods on several metrics. Code and model weights are available at https://github.com/jasonppy/word-discovery

Index Terms: visually-grounded speech, spoken term discovery, speech segmentation, self-supervised speech processing

1. Introduction and Related Work

The task of spoken term discovery involves starting with a corpus of untranscribed, unsegmented speech audio data and developing an algorithm or computational model that is capable of 1) inferring the temporal location of each word boundary and 2) clustering the resulting speech segments into categories that reflect word identity. The motivation behind this task is twofold. First, a model capable of solving this problem would be immensely useful for building speech understanding systems using a limited amount of transcribed data that could be applied to low-resource languages and domains. Second, the model could be used by cognitive science researchers as a model for the process by which human children acquire spoken language [1].

The spoken term discovery task has a long history dating back at least to the seminal work of Park and Glass [2], who proposed a segmental dynamic time warping algorithm to solve both the segmentation and clustering sub-tasks. In the time since then, numerous other approaches have been proposed, based on techniques such as refinements to the S-DTW algorithm [3], Bayesian probabilistic models [4, 5], dynamic programming algorithms [6, 7, 8] and contrastive predictive coding with neural models [9, 10, 11]. The ZeroSpeech [12] challenge was also introduced to provide a standard benchmark to the community for the word discovery task, in addition to other related tasks in zero-resource speech processing.

A separate line of research has investigated training speech models using a visual grounding (VG) objective [13, 14, 15, 16]. Given untranscribed, unsegmented speech that semantically describes an auxiliary input from the visual modality, a model is trained to associate the inputs together. Numerous works have demonstrated that these models' learned representations capture linguistic abstractions such as phonemes and words [17, 18, 19, 20, 21, 22, 23, 24, 25], but so far none of them have explicitly used a VG model on the spoken term discovery task.

In this paper, we present a method for *visually-grounded* spoken term discovery using a Transformer [26] model. Our method is simple: it simply involves applying a binary threshold to the self-attention maps produced by the model, and extracting contiguous temporal regions of the speech signal with attention scores above the threshold. We find that these segments accurately predict both the boundaries and identity of the underlying words, which we verify with a suite of experiments on the SpokenCOCO dataset. Additionally, we evaluate our proposed method on the Buckeye word segmentation [27] and ZeroSpeech [12] English spoken term discovery tasks, and find in both cases it significantly outperforms the current state-of-the-art, non-visually-grounded models from the literature.

2. Models

Our model has a simple Transformer-based dual-encoder architecture consisting of an audio encoder and an image encoder. The model takes as input a raw speech waveform A and an image I, and outputs a similarity score that should be large when the speech faithfully describes the content of the image, and small otherwise. The audio encoder has the same architecture as the wav2vec2.0 (W2V2) and HuBERT Base models. The input speech utterance is represented as waveform samples which are processed by a convolutional block to obtain features of a higher temporal resolution. Convolutional positional embeddings are added to the output features. We then follow [28] and augment the features with a `[CLS_A]` token before feeding them to the stack of Transformer layers. The image encoder is a Vision Transformer (ViT) [29, 30]. It takes an RGB image and splits it into a grid of non-overlapping patches, then uses a linear layer to embed the patches into a sequence of feature vectors. Trainable positional embeddings are added to the features, which are then concatenated with a `[CLS_I]` token before being fed to the Transformer layers. The similarity score between image I and audio A is defined as the dot product of `[CLS_I]` and `[CLS_A]`. Given parallel speech-image data, we can train the model using the InfoNCE loss[9, 31], where positive samples are matched pairs and negative samples unmatched pairs. The loss encourages high similarity scores between matched image-speech pairs and low similarity scores otherwise.

Unlike most models used in vision and language research, none of the modules in our model are pretrained on any supervised tasks such as image classification, object detection or au-

Copyright © 2022 ISCA

10.21437/Interspeech.2022-10652

tomatic speech recognition. Both our image and audio encoders are pretrained in a completely self-supervised fashion. We will refer to our visually-grounded variants of HuBERT and W2V2 as **VG-HuBERT** and **VG-W2V2**. As will be shown in section 3, similar to [32], we found that reinitializing the last few Transformer layers of the pretrained HuBERT or W2V2 with random weights can significantly improve the word discovery performance. We name models that reinitialize weights of the last x Transformer layers **VG-HuBERT**$_x$ and **VG-W2V2**$_x$.

3. Experiments

Datasets. We train our models on the visual grounding task with the Karpathy training split [33] of the SpokenCOCO dataset [34]. SpokenCOCO contains 123k images, each with 5 spoken captions produced by humans speaking aloud the text captions in MSCOCO [35]. We evaluate our model using both speech-image retrieval and word discovery tasks on the SpokenCOCO Karpathy test set. We additionally evaluate our model on the spoken term discovery task using a 6 hour test set [8] of the Buckeye corpus [27], as well as the ZeroSpeech 2020 [12] spoken term discovery evaluation split.

Implementation Details. Our image encoder inherits the architecture and is initialized from the DINO [30] self-supervised Vision Transformer. In particular, we initialize the image encoder weights from DINO-ViT small with patch size 8x8. The model consists of a linear embedding layer, a learnable positional embedding, and a 12-layer transformer with hidden dimension of 384 and 6 attention heads. The audio encoder inherits its architecture and initial weights from HuBERT (or W2V2) Base, which contains a 7-layer convolutional block, and a 12-layer Transformer with a hidden dimension of 786 and 12 attention heads. We project the output `[CLS_I]` and `[CLS_A]` separately using 2-layer MLPs to a 2048 dimensional space before taking their dot product. Models are trained on the visual grounding task with batch size 100 for 30 epochs. We optimize with BertAdam [36] using a learning rate linearly ramped up $5e-5$ over the first 10% of training, and then linearly decayed to 0. Except for the convolutional block in the audio encoder that is frozen, all other layers are trainable.

Retrieval Performance. As an initial sanity check, we first quantify how well our models learn to associate spoken descriptions with images by evaluating them on an image-speech retrieval task. We report Recall@K for $K = 1, 5, 10$ in Table 1. Although VG-HuBERT and VG-HuBERT$_3$ do not use any supervised vision modules, they still perform relatively well on this benchmark compared to FaST-VGS$_{\text{CO}}$, which is the current SoTA dual-encoder retrieval model that uses a pretrained Faster RCNN for image feature extraction.

Table 1: *Comparison of VG-HuBERT with other models on the 1k retrieval task on the SpokenCOCO test set.*

	Speech → Image			Image → Speech		
Model	R@1	R@5	R@10	R@1	R@5	R@10
ResDAVEnet [37]	17.3	41.9	55.0	22.0	50.6	65.2
FaST-VGS$_{\text{CO}}$ [28]	31.8	62.5	75.0	42.5	73.7	84.9
FaST-VGS$_{\text{CTF}}$ [28]	**35.9**	**66.3**	**77.9**	**48.8**	**78.2**	**87.0**
VG-HuBERT	30.4	60.7	72.7	42.7	73.6	84.1
VG-HuBERT$_3$	30.6	60.8	72.8	42.8	73.5	83.9

Word Discovery. The main focus of this paper is studying the emergence of word segmentation in visually grounded speech models (the image branch is not used in any word discovery experiment). We first discovered this phenomenon by plotting the `[CLS_A]` token's attention weight across each temporal frame of the speech at layer y of the audio encoder;

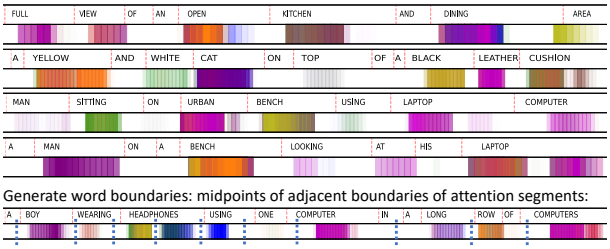

Figure 2: *Examples of CLS attention segments from SpokenCOCO, the bottom one also show how word boundaries are generated from attention segments*

Figure 2 displays several examples using $y = 12$ from the VG-HuBERT$_3$ model. We threshold the weights to retain only the top $z\%$ of the total attention weight for each attention head ($z = 10$). The ground truth transcripts and force aligned word boundaries (red dashed lines) are also plotted for reference. Figure 1 is generated the same way with $y = 9, z = 10$. Although we do not impose any constraints on the attention weights except for the thresholding described above, the visualization shows that the our model exhibits word localization, segmentation, and identification as it puts chunks of attention associated with different heads (indicated by different colors) under different words. In the following paragraphs, we pose several questions and metrics to quantitatively investigate these phenomena.

1. Does the model's attention map indicate the presence of words? We define two metrics to study this question: **word coverage (WC)**, defined as $\frac{\#\text{hits}}{\#\text{words}}$, where $\#$hits is the number of times the model assigns one or more attention segments to a word, and $\#$words is the count of total words in the corpus. We define an attention segment to be a consecutive sequence of frames with nonzero attention weights (after thresholding has been applied), and consider a segment to be assigned to a particular word if more than half of the segment's length falls between the ground-truth boundaries of that word. We also define **Temporal intersection over union (tIoU)** as $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\frac{s\cap w_s}{s\cup w_s}$, where $\mathcal{S}$ is the set of all attention segments, w_s is the word that attention segment s is assigned to, and the output of the intersection (or union) operation is the duration in frames of the intersection (or union) between the segment spanned by w_s and the segment spanned by s. WC and tIoU are complementary to each other — WC reflects how good the model is at detecting all words within an utterance, and tIoU shows how good the model is at localizing the span of each word it detects. A model could achieve a high WC if it assigns an attention segment every few frames (similar to HuBERT's attention map shown in Figure 1), but this will lead to a low tIoU; on the other hand, a model might achieve a good tIoU score if it ignores most words, and places attention segments only under the words that it is very confident about (e.g. salient visual words). To balance these metrics, we introduce a summary metric named **area score (A-score)**, defined as the harmonic mean of WC and tIoU, i.e. A-score $= \frac{2\text{WC}\cdot\text{tIoU}}{\text{WC}+\text{tIoU}}$. In addition, we also show **center distance (CD)** which is defined as the average distance in milliseconds between the center of each attention segment and center of the word that the segment is assigned to. We group the all above metrics under the name "Area".

2. Can we use the model to identify word boundaries? Here, we are interested in seeing whether we can use the model to precisely predict the onset and offset of each word in an utterance. For evaluation, we use the standard boundary **precision**, **recall**, $\boldsymbol{F_1}$, **over-segmentation (OS)**, and $\boldsymbol{R}$**-value** metrics [8].

Table 2: *Different models evaluated on their word discovery ability on the on SpokenCOCO test. ± indicates the standard deviation across 5 clustering runs with different random seeds; all models used 4096 k-means clusters. HuBERT$_{FT}$ and W2V2$_{FT}$ indicate models that were further fine-tuned on the SpokenCOCO audio captions.*

	Area				Boundary					Word	
Model	WC	tIoU	CD (ms)	A-score	Precision	Recall	F_1	OS	R-val	Purity	WD
ResDAVEnet-VQ [38]	98.39	23.78	119.4	38.30	10.42	50.96	17.30	38.88	-250.77	31.6 ± 0	262 ± 0
W2V2 [39]	77.68	38.56	111.9	51.54	11.52	24.33	15.63	11.12	-33.34	41.2 ± 0.3	407 ± 13
W2V2$_{FT}$ [39]	73.96	39.87	116.1	51.81	11.88	24.79	16.06	10.87	-31.10	43.0 ± 0.1	397 ± 11
HuBERT [40]	72.63	37.72	117.9	49.66	12.18	24.97	16.37	10.51	-28.26	43.1 ± 0.2	404 ± 12
HuBERT$_{FT}$ [40]	73.78	39.31	116.0	51.30	11.90	25.81	16.29	11.68	-36.72	45.3 ± 0.3	500 ± 9
FaST-VGS [28]	75.92	54.86	67.8	63.69	28.99	26.17	27.51	-9.72	40.10	70.9 ± 0.5	1011 ± 12
FaST-VGS+ [41]	77.40	51.46	80.5	61.82	22.66	27.86	24.99	22.93	28.54	72.3 ± 0.3	1026 ± 12
VG-W2V2	65.94	45.56	76.8	53.89	18.47	19.78	19.10	7.09	28.86	62.5 ± 0.4	743 ± 14
VG-W2V2$_4$	71.28	54.06	82.0	61.49	28.15	22.90	25.26	-18.64	39.67	73.0 ± 0.3	1182 ± 20
VG-W2V2$_5$	73.32	53.09	72.2	61.58	28.70	25.45	26.98	-11.32	39.94	75.4 ± 0.1	1149 ± 16
VG-HuBERT	73.16	44.02	89.0	54.97	18.31	18.90	18.60	3.26	29.60	63.2 ± 0.4	823 ± 20
VG-HuBERT$_3$	70.94	61.29	64.6	**65.77**	35.90	27.03	**30.84**	-24.72	**44.42**	75.3 ± 0.2	1167 ± 26
VG-HuBERT$_4$	73.97	56.35	78.9	63.97	28.39	25.64	26.94	-9.70	39.64	75.2 ± 0.2	**1230** ± 18

Table 3: *Performance of VG-HuBERT$_x$ on SpokenCOCO validation set.*

	Area			Boundary	Word
reinit last x	WC	tIoU	A-score	F_1	WD
x=0	73.21	44.16	55.09	18.46	817
x=1	69.29	51.88	59.33	22.17	1003
x=2	69.63	57.73	63.12	28.52	1036
x=3	70.95	61.51	**65.91**	**31.07**	1169
x=4	73.90	56.32	63.92	27.02	**1229**
x=5	74.92	54.44	63.05	25.52	1044
x=6	73.43	51.38	60.45	23.30	879

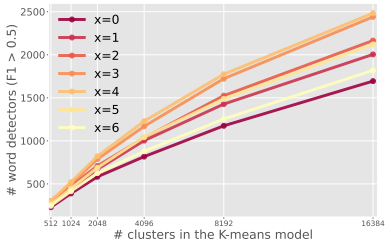

Figure 3: *#WD given by VG-HuBERT$_x$ with varied #K-means clusters.*

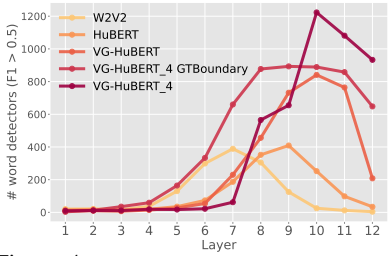

Figure 4: *#WD given by different models across all layers.*

These metrics are quite strict and only assume a very small tolerance window (20ms for SpokenCOCO and Buckeye [8] and 30ms for ZeroSpeech [12]). However, we note that visually, the boundaries of our thresholded attention segments are typically narrower than the words they are assigned to. To address this, we instead infer word boundaries to be the midpoints between the edges of adjacent attention segments (example shown at the bottom in Figure 2). We group the metrics used for studying question 2 under the name "Boundary".

3. Can the model determine word identity? Here, we use the word detection evaluation introduced by [38]. To do so, we first average pool the representations belonging to each attention segment to obtain a single feature vector representing that segment. We apply this procedure to all utterances in the SpokenCOCO test set, and then use a K-means clustering algorithm to group all of the segments across the test set. Next, following [38], we consider each cluster as a potential "detector" for every different vocabulary word appearing in the ground-truth transcripts. We compute precision, recall, and F1 scores for each unique cluster-word pair, and consider a cluster to be a "word detector" if it achieves an F1 score greater than or equal to 0.5 for some vocabulary word. Finally, we count the number of word-detecting clusters and report that as the **word detection (WD)** score. We additionally report cluster **purity** (in terms of the identity of each attention segment's assigned word) as a way of measuring the clustering performance. We group the metrics used for studying word detection under the name "Word".

Hyperparameter tuning. Performance on all of the above metrics depends on the model layer being examined, as well as the threshold applied to the attention maps. We tune these hyperparameters on the SpokenCOCO dev set separately for each metric group (Area, Boundary, and Word), respectively optimizing for A-score, boundary F1, and the number of word detectors found. During this tuning, we consider all 12 Transformer layers. For the attention threshold we sweep the threshold between 0.5 and 0.99. After tuning, we report all final results on the SpokenCOCO test set.

Results on SpokenCOCO. We summarize our results across the Area, Boundary, and Word metrics on the SpokenCOCO test dataset in Table 2. We compare against several visually-grounded baseline models from the literature, including ResDAVEnet-VQ [19], FaST-VGS [28] and FaST-VGS+ [41]. For ResDAVEnet-VQ, we use the layer 3 representations from the $\{3\} \rightarrow \{2, 3\}$ model from [19] distributed by the authors as it achieved the highest word detection performance in [19]. We first fine-tune ResDAVEnet-VQ on the SpokenCOCO data and then derive segments from consecutive frames assigned to the same VQ3 codebook entry. For the FaST-VGS models, we consider all layers of the `trm1` module and report results for layer achieving the best results on the dev set. Because `trm1` is also a transformer stack augmented with a `[CLS_A]` token, we derive attention segments from these models in the same manner as for VG-HuBERT. Finally, we compare against the Librispeech-pretrained HuBERT and wav2vec2.0 Base models, and these models fine-tuned on audio of SpokenCOCO. Although these are non-visually grounded models, they share the same architecture as VG-HuBERT and VG-W2V2 so making this comparison allows us to probe whether the same architecture trained solely with a masked language modeling objective will still exhibit emergent word discovery ability in its self-attention maps. Because W2V2 and HuBERT don't use a `[CLS_A]` token, we use the sum of attention weights each frame receives from other frames (separate summation for each attention head) to get the self-attention map, and then segments are derived in the same manner as for VG-HuBERT.

In Table 2, we first note that the highest overall performance for the Area and Boundary metrics is achieved by the VG-HuBERT$_3$ model, while the highest performance for the Word metrics is achieved by the VG-HuBERT$_4$ model. For both VG-HuBERT and VG-W2V2, we find that reinitializing the last few layers before training on the visual grounding task is highly beneficial, which is in line with the results found by [32]. A more extensive comparison of layer reinitialization is shown in Table 3. ResDAVEnet-VQ, W2V2, and HuBERT all have a low A-score, mainly due to a low tIoU, which indicates their relative inability to segment words; this is further confirmed by their low boundary F1 scores. In Figure 1, we visually see that the attention segments produced by HuBERT are highly fragmented and do not clearly correspond to words. Despite this, both HuBERT and W2V2 achieve nontrivial word detection results, outperforming the ResDAVEnet-VQ model but significantly underperforming VG-HuBERT, VG-W2V2, and FaST-VGS, all of which learn more than twice as many word detectors as HuBERT. This indicates that the masked language modeling objective used to train HuBERT and W2V2 does result in representations strong enough to capture word identity, but the visual grounding objective appears to be significantly more powerful in this respect. Additionally, the visual grounding objective leads to the emergence of word segmentation ability, whereas the masked language modeling objective does not. Finally, in comparing to FaST-VGS and FaST-VGS+ we note that these models use a strongly supervised vision module (Faster RCNN), whereas VG-HuBERT and VG-W2V2 are fully self-supervised and yet still outperform FaST-VGS and FaST-VGS+ across all metrics.

In Figures 4 and 3, we examine the word detection performance of various models across different Transformer layers and as we vary the number of K-means clusters. Although increasing the number of clusters does lead to a larger number of detectors (Figure 3), we see significant diminishing returns and thus fix $K = 4096$ in all of our experiments, unless stated otherwise. In Figure 4, we see that for all models, word detection performance is best in the middle and upper half of the model, which is also consistent with the analysis of [32] Finally, we compare the word detection performance of VG-HuBERT when using oracle word boundaries to determine segments (rather than the model's thresholded self-attention), which we counterintuitively find hurts the model's word detection ability. Combined with the observation that the attention segments tend to concentrate at the nucleus of words, we hypothesize that the model's contextualization is pushing word identity information towards the temporal center of each word.

Table 4: *Results on ZeroSpeech 2020 spoken term discovery.*

	Words	NED	Cov	M-score	Prec.	Rec.	F_1
J.V. [42]	18821	**32.4**	7.9	14.1	32.1	3.2	5.9
Granada et al.	92544	72.3	76.8	40.7	27.8	45.5	34.5
Räsänen et al.	**321603**	52.5	28.7	35.8	25.8	29.9	27.7
ES-KMeans [6]	42473	72.3	**100.0**	43.4	39.6	61.4	48.2
SEA [43]	240033	89.5	99.5	19.0	27.3	75.9	40.1
PDTW [44]	85425	48.2	85.4	64.5	26.5	**88.2**	40.8
VG-HuBERT$_3$ (Ours)	104696	42.5	95.4	**71.8**	**44.7**	54.2	**49.0**

Results on Buckeye and ZeroSpeech 2020. Finally, we test our best performing model (VG-HuBERT$_3$) [1], on two additional benchmarks: the ZeroSpeech 2020 spoken term discovery track (English) and the Buckeye word segmentation task[2].

[1] we use model's best word segmentation hyperparams, max-pooling instead of mean-pooling for segment features, and 16384 clusters for K-means.

[2] we make use of https://github.com/kamperh/vqwordseg

Table 5: *Unsupervised word segmentation on Buckeye test set.*

	Boundary				Token
Model	Prec.	Rec.	F_1	R-val.	F_1
Adaptor gram. [45]	15.9	**57.7**	25.0	-139.9	4.4
SylSeg [46]	27.7	28.9	28.3	37.7	19.3
ES-KMeans [6]	30.3	16.6	21.4	39.1	19.2
BES-GMM [5]	31.5	12.4	17.8	37.2	18.6
SCPC [47]	36.9	29.9	33.0	45.6	-
mACPC [10]	42.1	30.3	35.1	47.4	-
DPDP [8]	35.3	37.7	36.4	44.3	25.0
VG-HuBERT$_3$ (Ours)	**47.6**	42.3	**44.8**	**54.2**	**31.0**

Table 4 shows model performance on ZeroSpeech, where following the official challenge leaderboard and [44], we report number of discovered words (Word), phoneme level normalized edit distance (NED) of the discovered segment pairs, coverage of the discovered segments (Cov) w.r.t. all utterances in the given corpus, and also word segmentation performance (precision, recall, F-score). We also report M-score [44], which is the harmonic mean of (100 - NED) and Cov that summarizes the trade-off between purity and coverage (similar to A-score we used for previous evaluation). We see that VG-HuBERT$_3$ achieves the highest M-score and boundary F_1.

Table 5 shows word segmentation performance on Buckeye test set, where we again report boundary precision, recall, and F_1. Token F_1 is harsher metric where both left and right boundaries of a word need to be successfully predicted. We see that our VG-HuBERT$_3$ significantly improves over all previous approaches on Boundary F_1, R-value, and Token F_1 by 23%, 14%, and 24% relative respectively.

Importantly, we note that all other models in Table 4 and 5 are unimodal models, i.e. they are only trained on the audio modality, while our model uses the visual modality during training (visual modality is not needed during testing). On the other hand, these other models are explicitly designed for word discovery and segmentation via specific mechanisms or learning objectives, while our model is instead trained with a cross-modal grounding objective. Our experiments demonstrate that this objective, combined with the Transformer architecture pretrained on a masked language modeling task, facilitates *emergent* word discovery ability from raw speech signals.

4. Conclusion and Future Directions

In this paper, we demonstrated that training the popular HuBERT and wav2vec2.0 models on a visual grounding task resulted in them learning to automatically discover (localize, segment, and identify) words in the speech signal. We verified that this same ability did not exist in the models prior to them undergoing training on the visual grounding task, suggesting that a combination of multiple self-supervised objectives may be essential in order for a computational model to learn the structure of spoken language from untranscriped speech audio.

In our future work, we plan to use the word-level representations learned by our model to anchor the learning of sub-word structure (e.g. automatically learning a pronunciation lexicon in terms of discovered speech units) as well as the higher level syntactic structure of spoken language. We also plan to investigate whether a similar method to the one we propose could facilitate visual object discovery, and also explore applications of our method to low-resource ASR.

5. Acknowledgements

We thank Abdelrahman Mohamed, Karen Livescu, Haoyue Freda Shi, Cheng-I Jeff Lai, and Bowen Shi for their insights.

6. References

[1] E. Dupoux, "Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner," in *Cognition*, 2018.

[2] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *TASLP*, vol. 16, no. 1, pp. 186–197, 2008.

[3] A. Jansen, K. W. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *INTERSPEECH*, 2010.

[4] C.-y. Lee, T. J. O'Donnell, and J. Glass, "Unsupervised lexicon discovery from acoustic input," *TACL*, vol. 3, pp. 389–403, 2015.

[5] H. Kamper, A. Jansen, and S. Goldwater, "A segmental framework for fully-unsupervised large-vocabulary speech recognition," *Computer Speech & Language*, 2017.

[6] H. Kamper, K. Livescu, and S. Goldwater, "An embedded segmental k-means model for unsupervised segmentation and clustering of speech," in *ASRU*, 2017.

[7] H. Kamper and B. van Niekerk, "Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks," in *Interspeech*, 2021.

[8] H. Kamper, "Word segmentation on discovered phone units with dynamic programming and self-supervised scoring," *ArXiv*, vol. abs/2202.11929, 2022.

[9] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748.

[10] S. Cuervo, M. Grabias, J. Chorowski, G. Ciesielski, A. La'ncucki, P. Rychlikowski, and R. Marxer, "Contrastive prediction strategies for unsupervised segmentation and categorization of phonemes and words," *ArXiv*, vol. abs/2110.15909, 2021.

[11] S. Bhati, J. Villalba, P. Żelasko, L. Moro-Velázquez, and N. Dehak, "Segmental contrastive predictive coding for unsupervised word segmentation," in *Interspeech*, 2021.

[12] M. Versteegh, R. Thiollière, T. Schatz, X.-N. Cao, X. A. Miró, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2017," *ASRU*, 2017.

[13] G. Synnaeve, M. Versteegh, and E. Dupoux, "Learning words from images and speech," 2014.

[14] D. F. Harwath and J. R. Glass, "Deep multimodal semantic embeddings for speech and images," *ASRU*, 2015.

[15] D. F. Harwath, A. Torralba, and J. R. Glass, "Unsupervised learning of spoken language with visual context," in *NIPS*, 2016.

[16] G. Chrupała, "Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques," *JAIR*, vol. 73, pp. 673–707, 2022.

[17] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," *ECCV*, 2018.

[18] W. N. Havard, J.-P. Chevrot, and L. Besacier, "Word recognition, competition, and activation in a model of visually grounded speech," in *CoNLL*, 2019.

[19] D. F. Harwath, W.-N. Hsu, and J. R. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," in *ICLR*, 2020.

[20] K. Khorrami and O. J. Räsänen, "Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? - a computational investigation," *ArXiv*, vol. abs/2109.14200, 2021.

[21] G. C. Mitja Nikolaus, Afra Alishahi, "Learning english with peppa pig," *ArXiv*, vol. abs/2202.12917, 2022.

[22] S. Scholten, D. Merkx, and O. Scharenborg, "Learning to recognise words using visually grounded speech," *ISCAS*, 2021.

[23] K. Olaleye and H. Kamper, "Attention-based keyword localisation in speech using visual grounding," in *Interspeech*, 2021.

[24] L. Wang and M. Hasegawa-Johnson, "A translation framework for visually grounded spoken unit discovery," in *ACSSC*, 2021.

[25] A. H. Liu, S. Jin, C.-I. Lai, A. Rouditchenko, A. Oliva, and J. R. Glass, "Cross-modal discrete representation learning," *ArXiv*, vol. abs/2106.05438, 2021.

[26] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[27] M. Pitt, K. Johnson, E. Hume, S. F. Kiesling, and W. Raymond, "The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability," *Speech Commun.*, 2005.

[28] P. Peng and D. Harwath, "Fast-slow transformer for visually grounding speech," in *ICASSP*, 2022.

[29] Alexey Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[30] M. Caron, H. Touvron, I. Misra, H. J'egou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021.

[31] G. Ilharco, Y. Zhang, and J. Baldridge, "Large-scale representation learning from visually grounded untranscribed speech," in *CoNLL*, 2019.

[32] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *ASRU*, 2021.

[33] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, 2017.

[34] W.-N. Hsu, D. F. Harwath, C. Song, and J. Glass, "Text-free image-to-speech synthesis using learned segmental units," in *ACL*, 2021.

[35] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[37] W.-N. Hsu, D. F. Harwath, and J. Glass, "Transfer learning from audio-visual grounding to speech recognition," in *INTERSPEECH*, 2019.

[38] D. Harwath, W.-N. Hsu, and J. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," in *ICLR*, 2020.

[39] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[40] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[41] P. Peng and D. Harwath, "Self-supervised representation learning for speech using visual grounding and masked language modeling," in *SAS@AAAI*, 2022.

[42] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *ASRU*, 2011.

[43] S. Bhati, J. Villalba, P. Żelasko, and N. Dehak, "Self-expressing autoencoders for unsupervised spoken term discovery," in *Interspeech*, 2020.

[44] O. J. Räsänen and M. A. C. Blandón, "Unsupervised discovery of recurring speech patterns using probabilistic adaptive metrics," in *Interspeech*, 2020.

[45] A. J. et al., "A summary of the 2012 jhu clsp workshop on zero resource speech technologies and models of early language acquisition," in *ICASSP*, 2013.

[46] O. Räsänen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Interspeech*, 2015.

[47] S. Bhati, J. Villalba, P. Żelasko, L. Moro-Velázquez, and N. Dehak, "Segmental contrastive predictive coding for unsupervised word segmentation," in *Interspeech*, 2021.