INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Syllable Discovery and Cross-Lingual Generalization in a Visually Grounded, Self-Supervised Speech Model

Puyuan Peng[1], *Shang-Wen Li*[2], *Okko Räsänen*[3], *Abdelrahman Mohamed*[4], *David Harwath*[1]

[1]Department of Computer Science, The University of Texas at Austin, USA
[2]Meta AI, USA, [3]Unit of Computing Sciences, Tampere University, Finland, [4]Rembrand, USA
pyp@utexas.edu

Abstract

In this paper, we show that representations capturing syllabic units emerge when training a self-supervised speech model with a visually-grounded training objective. We demonstrate that a nearly identical model architecture (HuBERT) trained with a masked language modeling loss does not exhibit this same ability, suggesting that the visual grounding objective is responsible for the emergence of this phenomenon. We propose the use of a minimum cut algorithm to automatically predict syllable boundaries in speech, followed by a 2-stage clustering method to group identical syllables together. We show that our model not only outperforms a state-of-the-art syllabic segmentation method on the language it was trained on (English), but also generalizes in a zero-shot fashion to Estonian. Finally, we show that the same model is capable of zero-shot generalization for a word segmentation task on 4 other languages from the Zerospeech Challenge, in some cases beating the previous state-of-the-art.[1]

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

1. Introduction

Traditionally, automatic speech recognition, speech synthesis, and spoken language understanding tasks have relied on supervised learning and the assumption that ground-truth text transcriptions of the training speech are available. Such transcriptions are costly to collect and represent a major hurdle in developing speech recognition and related technologies that can serve the thousands of languages around the world.

Recently the speech community has made tremendous progress developing self-supervised models that can learn powerful representations of the speech signal by being pre-trained on untranscribed speech data. After pre-training the models can be fine-tuned on a small amount of transcribed data to achieve impressive performance on a variety of tasks [1, 2, 3, 4, 5]. Furthermore, the representations learned by these models can be clustered into discrete speech units that have been shown to be strongly correlated with words and phones [6, 7]. These units can be used to tokenize speech into a pseudo-text sequence, which can be used as a drop-in replacement for a text transcription in a wide variety of downstream tasks, giving rise to a new genre of "textless" speech processing research [8, 9, 10, 11].

Because of the emergent nature of these units, it is not yet understood how to control what type of linguistic structure (e.g. phones, syllables, words) they will capture. It has been shown that the representations of self-supervised speech models tend to correlate with lower-level structure such as phones at lower model layers, and higher-level structure such as words at higher model layers [6, 12]. However, it has also been demonstrated that the model's training objective strongly influences the nature of these representations. Training the model to perform cross-modal grounding of speech to contextually-relevant visual images has been shown to dramatically increase the model's word learning capability over a masked language modeling objective, even when the model architecture is held nearly constant [7].

In this paper, we build on [7] and demonstrate that multimodal self-supervision simultaneously results in the emergence of word-like and syllable-like representations within the same model. While [7] showed that word-like units are encoded by the Transformer's attention heads, we show that syllabic structure emerges within the embeddings of the token sequence itself. We propose the use of a minimum cut segmentation algorithm to derive syllable boundaries from these features, outperforming a state-of-the-art method for unsupervised syllabic segmentation. We then show that these segments can be clustered across a speech corpus to perform syllable discovery, enabling tokenization of the speech signal at the level of syllable-like units. Finally, we also show surprising results where our model trained only on English speech is able to perform zero-shot segmentation of syllables on another language (Estonian) and words in multiple non-English languages, in several cases outperforming the state-of-the-art models on the Zerospeech challenge [13].

2. Related Work

Besides the aforementioned work on self-supervised and textless speech processing, our work is also related to spoken term discovery and visually grounded speech processing.

Spoken term discovery - inferring the temporal boundary and identity of words and short phrases from untranscribed speech audio data - has been an important research direction in Zero-resource speech processing [13]. The earliest work that tackles spoken term discovery date back to at least the segmental dynamic programming algorithm proposed by Park and Glass [14]. Since then, numerous other approaches have been proposed. [15, 16] developed Bayesian models for hierarchical phoneme and word discovery. Based on the fact that syllables are organized around particularly sonorous speech sounds, [17] developed sonority fluctuation-based method for syllabic segmentation. Other works model word directly either via an iterative segmentating-clustering approach [18], or reinforcement learning [19]. Self-supervised learning has also been considered for end-to-end phoneme and word segmentation [20, 21]. Mostly recently, Algayres et al. [22] identified the key issues in applying text-based models for speech segmentation, and proposed the DP-Parse algorithm which uses instance lexicon to mitigate clustering error. Herman [23] applied vector quantization for phoneme-like unit discovery, and then ran a dynamic program-

[1] Code & Model: https://github.com/jasonppy/syllable-discovery.

10.21437/Interspeech.2023-2044

ming algorithm on the discovered units for word segmentation.

Visually grounded speech (VGS) processing [24] generalizes the idea of self-supervised learning to multimodal (visual) data and learns speech representations by associating speech audio with contextually-relevant visual input. VGS usually leverages image-speech [25, 26] or video-speech [27, 28] paired data. In practice, besides speech-image retrieval and alignment [29, 30, 31, 32, 33, 34], VGS models has also be shown to achieves competitive performance keyword spotting [35], query-by-example research [36], and varies tasks in the SUPERB benchmark [37, 38]. The study of linguistic information learned in VGS models has been attracting increasing attention. In particular, researchers has measured the phonetic, syllabic, and lexical information in VGS models [39, 40, 6, 41, 42, 7, 43]. In addition to [7] which we build our work on, [43] is the most relevant to ours where they studied the emergence of phonetic, syllabic, and lexical information in different layers of CNN-based VGS models. Our work is different from their in that none of the modules of our model receives textual supervision, while their image encoder is pre-trained on Imagenet classification [44]. In addition, we show the emergence of hierarchical linguistic information in the non-hierarchical Transformer model, while they use hierarchical CNN models.

3. Technical Approach

VG-HuBERT [7] is a self-supervised dual-encoder model trained using a contrastive loss to match speech waveforms with the images they describe. Although VG-HuBERT is not trained with any textual supervision, the model has been shown to exhibit strong word discovery capabilities [7]. Specifically, its CLS token places concentrated chunks of attention weight on word segments in input utterances (see lower left subfigure of figure 1 for an example). Our motivating hypothesis is that VG-HuBERT's word discovery ability is predicated on its ability to also discover sub-word units at earlier layers. To probe this we first extract a sequence of frame embeddings from some layer of the model given an input waveform, $\mathbf{C} \in \mathbb{R}^{T \times D}$, ($T$ is number of speech frames, D is the feature dimension). Next, we then calculate the feature self-similarity matrix as featSSM $:= \mathbf{C}\mathbf{C}^{\intercal}$. We normalize featSSM by subtracting smallest element of the matrix from all elements to insure that all frame-pair similarity scores are non-negative. Figure 1 shows an example of featSSM, where green color denotes high similarity and blue denotes low similarity. We see a clear block diagonal structure in VG-HuBERT's featSSM, where each block corresponds to a syllable. In HuBERT's featSSM, however, the block structure hardly exists. Based on the different patterns we see between the feature self-similarity matrix and the CLS attention, we hypothesize that visually grounded training leads to the emergence of syllable identity being encoded in VG-HuBERT's features, and the CLS token attending to these features to infer the presence of words. To quantitatively study the syllable discovery phenomenon, we adopt the normalized minimum cut algorithm [45, 46, 47] to automatically segment the blocks in featSSM, and use the block boundaries to predict syllable boundaries.

A min-cut segmentation algorithm for featSSM. We define a fully-connected, undirected graph $G(V, E)$ for every speech utterance. Set V consists of all speech frames as nodes; Set E consists of edges, where the edge weight $w(u, v)$ is defined as the similarity score corresponding to nodes u and v. Segmenting the blocks in featSSM means partitioning the corresponding graph $G(V, E)$ into disjoint sets $A_1, A_2, \cdots, A_k$ such that similarity among nodes (i.e. frames) within each set are maximized, and while minimizing the similarities of nodes between sets. To achieve this, [45] proposed the following objective:

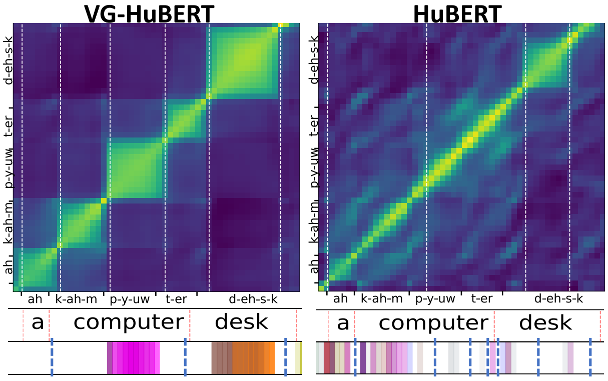

Figure 1: *Visualization of feature self-similarity matrix (upper) and the attention (lower) in VG-HuBERT and HuBERT. The vertical white dotted lines are generated by minCutMerge, and vertical blue dotted lines are generated by taking the midpoint of boundaries of adjacent attention segments*

$$\text{Ncut}_k(V) = \frac{cut(A_1, V - A_1)}{vol(A_1)} + \cdots + \frac{cut(A_k, V - A_k)}{vol(A_k)}$$

where $cut(A, B) := \sum_{u \in A, v \in B} w(u, v)$, and $vol(A) := \sum_{u \in A, v \in V} w(u, v)$. For sequential data, the above minimization problem can be solved using a dynamic programming algorithm [46] in $O(KN^2)$ time. Here K is the number of partitions (estimated number of syllables in the utterance in our case), and N is the number of nodes (speech frames). K needs to be set up-front for every utterance, and we use a hyperparameter second-per-syllable (secPerSyllable) to decide K based on the duration of the utterance. In practice, we use the variant introduced in [47], where we first oversegment featSSM, and then iteratively merge temporally adjacent partitions if the cosine similarity of the averaged features belonging to the two partitions falls below some threshold (denoted as mergeThres). We found that this variant always outperformed the original algorithm proposed in [46].

Clustering. With hypothesized syllabic segment boundaries produced by the min-cut algorithm, we further use a 2-step clustering approach to categorize the segments. Average features within each segment are used as the embedding of the segment. We initially cluster the segment embeddings using KMeans to produce a large number of clusters, and then run agglomerate clustering to merge similar clusters. We found our 2-step clustering approach to work better compared to just using Kmeans, given the same number of final clusters. Since our work and [7] are both based on VG-HuBERT, we denote [7]'s segmentation approach as **VG-HuBERT$_{\text{cls}}$**, where the CLS attention is used to segment speech, and denote our approach as **VG-HuBERT$_{\text{featSSM}}$**, where the min-cut algorithm is used on featSSM for segmentation. Both approaches used the 2-step clustering method for segment categorization.

4. Experiments

4.1. Datasets

Following [7], the training dataset is SpokenCOCO [48], an image-English spoken caption dataset built on top of the MSCOCO image-text caption dataset [49]. For evaluation on English, we use the test set of SpokenCOCO. Since SpokenCOCO does not have syllable alignment, we first use the Montreal

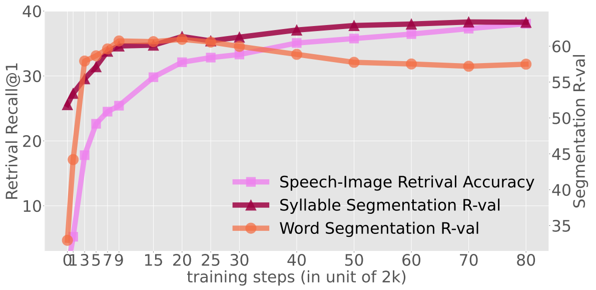

Figure 2: *The performance of speech-image retrieval, and syllable and word segmentation of VG-HuBERT as training progress.*

Forced Aligner[2] to generate phonetic and word alignment, and then derive the corresponding syllable alignment utilizing a rule-based syllabification script[3]. For cross-lingual generalization experiments, we follow [17] and evaluate our approaches on Estonian syllabic segmentation using the Phonetic Corpus of Estonian Spontaneous Speech [50], which contains conversational speech between two test subjects recorded with near-field microphones. The corpus comes with manually verified syllable transcription and alignment. We also evaluate our approach on the Zerospeech word segmentation task, which contains five languages: Mandarin, English, French, German, and Wolof.

4.2. Implementation details

Model training. We use the official open-sourced codebase and training recipe released by Peng and Harwath [7] and train a VG-HuBERT on SpokenCOCO. Model snapshots are saved during training for syllable and word discovery analysis.

Evaluation. To evaluate segmentation performance, we use precision, recall, F1 and R-value [51, 23]. For the calculation of above metrics, we use a tolerance window of 50ms for SpokenCOCO and Estonian following [17], and 30ms for the Zerospeech Challenge [13]. To evaluate the quality of our syllable clustering, we first match hypothesized syllable segments with the ground truth segments for each utterance. To do so, we use a Hungarian matching algorithm where each segment is a node and edge weights are defined by temporal intersection-over-union between each hypothesized segment and ground truth segment (unmatched segments are assigned to a dummy segment). Then, we follow [7] and use cluster purity and number of detected syllables (DS). A syllable is defined as being detected if it achieves an F1 score greater than 0.5 for some cluster [7]. To avoid conflating word detection and syllable detection, we only evaluate on multisyllabic words.

Hyperparameter tuning. For SpokenCOCO, we tune the mergeThres to maximize the segmentation R-value on the SpokenCOCO validation set. The number of clusters in Kmeans and agglomerative clustering are fixed at 16384 and 4096. For syllabic segmentation on Estonian, we tune the hyperparameters on a validation set created following the procedure introduced in [17], using a subset of the original Estonain corpus [50]. For cross-lingual word segmentation on the Zerospeech challenge, we use the hyperparameters selected from the SpokenCOCO validation set.

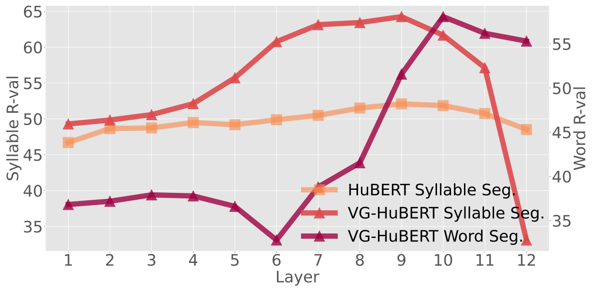

Figure 3: *Layer-wise performance of VG-HuBERT on syllable and word segmentation, and HuBERT on syllabic segmentation on SpokenCOCO val set. HuBERT word segmentation gives very poor results [7] and therefore is not shown.*

4.3. When do syllables and words emerge during training?

We first investigate when syllable and word information emerges during the training of VG-HuBERT. In Figure 2, we show the syllable and word segmentation performance of VG-HuBERT as a function of training iteration, along with speech-image retrieval accuracy on the SpokenCOCO validation set. Since the contrastive training loss is a direct approximation of the retrieval metric, speech-image retrieval accuracy keeps improving throughout the course of training as expected. For syllabic segmentation, VG-HuBERT reaches the first peak at 20*2k steps, and the performance keeps improving shortly afterwards, with a trend similar to retrieval performance. Interestingly, VG-HuBERT peaks at 20*2k steps for word segmentation, and the performance slightly decreases before levelling off. Anecdotally, by manually examining some examples we found that VG-HuBERT's CLS token tends to ignore more words in the later stages of training. This might be because the model is starting to ignore non-salient words in order to produce semantic representations that are more discriminative in terms of retrieval performance. Notably, as we can see in Figure 1, syllabic information for the entire utterance tends to persist in the model's representations even when some segments are ignored by the CLS token's attention.

4.4. Where in the model do syllables and words emerge?

We next perform a layer-wise study to show how visual grounding helps the emergence of syllables and words, and the interplay between the discovery of different linguistic units. Figure 3 compares VG-HuBERT to HuBERT for syllabic segmentation, and also shows VG-HuBERT's word segmentation on the SpokenCOCO validation set. HuBERT performs quite evenly across all layers, while syllabic segmentation is best in VG-HuBERT's mid to late layers, and VG-HuBERT's word segmentation ability is concentrated in the final few layers. We also fine-tuned HuBERT on the SpokenCOCO utterances using its original self-supervised loss to mitigate the potential domain gap, but did not see any improvement in syllabic segmentation (see first two rows in Table 1). We see a 'division of labor' between different layers in VG-HuBERT with middle layers performing best in syllabic segmentation, while the last three layers specialize in word segmentation. In addition, we note that the best syllabic segmentation layer (layer 9) is right before the best word segmentation layer (layer 10), indicating that the attention heads may be learning to string syllables together into words. We leave a more in-depth investigation of this phenomenon for future work.

[2] https://montreal-forced-aligner.readthedocs.io/en/latest/
[3] https://github.com/kylebgorman/syllabify

Table 1: *Syllabic segmentation performance of different models on SpokenCOCO test set. DS denotes detected syllables.*

Model	Prec.	Rec.	F1	R-val.	Purity	DS
HuBERT ft. [2]	43.8	49.4	46.4	51.5	29.0	519
HuBERT [2]	43.8	46.5	45.1	52.0	30.1	522
VG-HuBERT$_{\text{cls}}$ [7]	58.7	37.1	45.5	54.3	66.1	751
Oscillator [17]	52.0	64.6	57.6	57.4	-	-
VG-HuBERT$_{\text{featSSM}}$	57.4	63.6	**60.3**	**64.3**	45.8	**902**

4.5. Syllable discovery on English

Table 1 compares VG-HuBERT with other models for syllable discovery on the SpokenCOCO test set. We see that HuBERT performs the worst on this dataset, no matter whether it is fine-tuned on SpokenCOCO or not. VG-HuBERT$_{\text{cls}}$ denotes the CLS token's attention-based segmentation, a method that has been shown to achieve SotA on word segmentation [7], gives high precision and low recall on this syllabic segmentation task as expected. In terms of syllable detection, we see that VG-HuBERT$_{\text{cls}}$ can detect more than 700 syllables with a high cluster purity. Considering the high cluster purity and low boundary recall of VG-HuBERT$_{\text{cls}}$, we conclude that this approach is able to discover a smaller number of syllables, but is highly confident of the ones that it does discover. Oscillator [17] is a signal processing-based syllabic segmentation algorithm that achieves SotA for unsupervised syllabic segmentation on multiple languages, including English. Oscillator performs reasonably well on this dataset, only lagging behind our approach on segmentation. Our VG-HuBERT$_{\text{featSSM}}$ model achieves the best performance in both syllabic segmentation (best F1 and R-val) and clustering (best DS).

4.6. Zero-shot syllabic segmentation on Estonian

Syllables are strongly correlated with speech intensity and voicing, and are organized around sonorant speech sounds [17]. This suggests that a syllable detection model trained on one language may able to generalize to other languages. We thus evaluate our English-trained models on a non-English language, namely Estonian. We use the same five-hour subset and evaluation pipeline as [17]. Table 2 lists the results. We see that compared to other methods including the Oscillator, our VG-HuBERT performs the best in both F1 and R-val metrics, indicating that its syllabic segmentation ability is at least somewhat language-agnostic.

Table 2: *Syllabic segmentation on the Estonian corpus.*

Approach	Prec.	Rec.	F1	R-val.
VG-HuBERT$_{\text{cls}}$ [7]	56	77	65	57
HuBERT [2]	64	75	69	70
WN [17]	77	62	69	72
EnvMin [52]	67	71	69	73
Vseg [53]	82	63	71	73
Oscillator [17]	71	78	74	77
Oscillator (our reprod.)	72	78	75	78
VG-HuBERT$_{\text{featSSM}}$	77	80	**79**	**82**

4.7. Zero-shot word segmentation on unseen languages

Lastly, we ask the question: if VG-HuBERT's CLS token detects words in English, what does it do for a language it has not seen during training? To investigate CLS token's behavior on languages unseen during training, we first visualize the CLS attention for Estonian and Mandarin utterances in figure 4. We see that anecdotally, the CLS attention appears to be performing syllabic segmentation, but it sometimes also connect adjacent syllables together. In some cases, the connections give invalid words - in figure 4, for Estonian (the upper figure), 'h_ve' and 'i' are connected, but the result is not a valid word; for Mandarin, '必须分' is connected (in the middle figure), and the result is also not a valid word. However, in some other cases, the connections happen to give valid words - in the two Mandarin examples in figure 4, '历史' and '不知' got connected, and they are valid words.

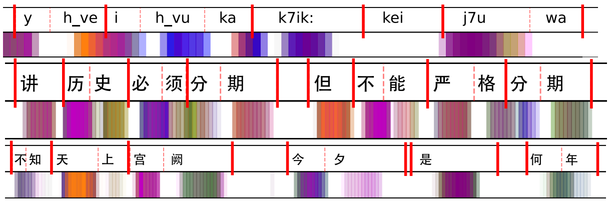

Figure 4: *Visualizations of VG-HuBERT's CLS attention on unseen languages - Estonian and Mandarin. Thin dashed lines denote syllable boundaries, thick vertical line denotes word boundaries. Word boundaries are also syllable boundaries.*

Based on the observation that the CLS token produces a mixture of monosyllablic and multisyllabic segmentation, we test VG-HuBERT$_{\text{cls}}$ for word segmentation on the Zerospeech challenge. In table 3, we see that VG-HuBERT achieves SotA performance on three out of five languages, despite only being trained on English. Interestingly, VG-HuBERT performs very differently on Mandarin and Wolof. While this could be due to hyperparameter settings (we use the same hyperparameters for all languages), we are not able to verify because the Wolof transcripts are not publicly available.

Table 3: *Word segmentation performance on the Zerospeech Challenge. Token F1 is a stricter metric than boundary F1 where a word is considered a hit only when both it's start and end boundaries are successfully predicted.*

Approach	Mand.	French	Engl.	German	Wolof
PDTW [54]	4.4	5.1	4.1	2.9	4.2
ES-KMeans [18]	8.1	6.3	19.2	14.5	10.9
SEA [55]	12.1	6.3	6.6	6.3	12.6
DP-Parse [22]	16.0	15.3	21.9	13.4	**17.5**
DPDP [23]	**26.3**	12.2	19.2	9.0	15.0
VG-HuBERT$_{\text{cls}}$	19.5	**15.5**	**26.6**	**15.8**	7.1

5. Concluding Discussion

In this paper, we demonstrated that the VG-HuBERT visually-grounded speech model exhibits emergent syllable recognition behavior. We proposed the use of a minimum cut algorithm to automatically extract syllable boundaries from the model's learned representations, and showed that this segmentation ability could transfer to Estonian speech even though the model was only trained on English. Furthermore, we demonstrated that the emergent word discovery ability that is also present in the model could be applied in a zero-shot transfer fashion to segment words in non-English languages, achieving state-of-the-art segmentation performance for several languages in the Zerospeech Challenge benchmark. In our future work, we plan to apply our syllable discovery method to tokenize speech waveforms and use these tokenizations in various textless speech processing tasks such as spoken language modeling and speech-to-speech translation, as well as unsupervised speech recognition.

6. References

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

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

[3] Y.-A. Chung *et al.*, "w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," *ASRU*, 2021.

[4] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *JSTSP*, 2021.

[5] A. Mohamed *et al.*, "Self-supervised speech representation learning: A review," *JSTSP*, 2022.

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

[7] P. Peng and D. F. Harwath, "Word discovery in visually grounded, self-supervised speech models," in *Interspeech*, 2022.

[8] K. Lakhotia *et al.*, "On generative spoken language modeling from raw audio," *TACL*, 2021.

[9] X. Li, Y. Jia, and C.-C. Chiu, "Textless direct speech-to-speech translation with discrete speech representation," *ArXiv preprint*, 2022.

[10] T. Nguyen *et al.*, "Generative spoken dialogue language modeling," *ArXiv*, 2022.

[11] G.-T. Lin *et al.*, "Dual: Textless spoken question answering with speech discrete unit adaptive learning," *ArXiv preprint*, 2022.

[12] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," *ArXiv preprint*, 2022.

[13] E. Dunbar, N. Hamilakis, and E. Dupoux, "Self-supervised language learning from raw audio: Lessons from the zero resource speech challenge," *JSTSP*, 2022.

[14] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *TASLP*, no. 1, 2008.

[15] C.-y. Lee, T. J. O'Donnell, and J. Glass, "Unsupervised lexicon discovery from acoustic input," *TACL*, 2015.

[16] T. Taniguchi, S. Nagasaka, and R. Nakashima, "Nonparametric bayesian double articulation analyzer for direct language acquisition from continuous speech signals," *TCDS*, 2015.

[17] O. Räsänen, G. Doyle, and M. C. Frank, "Pre-linguistic segmentation of speech into syllable-like units," *Cognition*, 2018.

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

[19] Y. Wang, H. Lee, and L. Lee, "Segmental audio word2vec: Representing utterances as sequences of vectors with applications in spoken term detection," in *ICASSP*, 2018.

[20] S. Bhati *et al.*, "Segmental contrastive predictive coding for unsupervised word segmentation," in *Interspeech*, 2021.

[21] S. Cuervo *et al.*, "Contrastive prediction strategies for unsupervised segmentation and categorization of phonemes and words," *ICASSP*, 2022.

[22] R. Algayres *et al.*, "Dp-parse: Finding word boundaries from raw speech with an instance lexicon," *TACL*, 2022.

[23] H. Kamper, "Word segmentation on discovered phone units with dynamic programming and self-supervised scoring," *TASLP*, 2022.

[24] G. Chrupała, "Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques," *J. Artif. Intell. Res.*, 2021.

[25] S. Gabriel, V. Maarten, and E. Dupoux, "Learning words from images and speech," in *NeurIPS Workshop on Learning Semantics*, 2014.

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

[27] A. Rouditchenko *et al.*, "Avlnet: Learning audio-visual language representations from instructional videos," in *Interspeech*, 2021.

[28] M. Nikolaus, A. Alishahi, and G. Chrupała, "Learning english with peppa pig," *TACL*, vol. 10, pp. 922–936, 2022.

[29] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech," *TASLP*, vol. 27, pp. 89–98, 2017.

[30] R. Sanabria, A. Waters, and J. Baldridge, "Talk, don't write: A study of direct speech-based image retrieval," in *Interspeech*, 2021.

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

[32] Y.-J. Shih *et al.*, "Speechclip: Integrating speech with pre-trained vision and language model," *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 715–722, 2022.

[33] K. Khorrami and O. J. Räsänen, "Evaluation of audio-visual alignments in visually grounded speech models," in *Interspeech*, 2021.

[34] D. F. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. R. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," *IJCV*, vol. 128, pp. 620–641, 2018.

[35] K. Olaleye, "Visually grounded keyword detection and localisation for low-resource languages," *ArXiv*, vol. abs/2302.00765, 2023.

[36] H. Kamper, A. Anastassiou, and K. Livescu, "Semantic query-by-example speech search using visual grounding," *ICASSP*, pp. 7120–7124, 2019.

[37] S. Yang *et al.*, "SUPERB: speech processing universal performance benchmark," in *Interspeech*, 2021.

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

[39] A. Alishahi, M. Barking, and G. Chrupała, "Encoding of phonology in a recurrent neural model of grounded speech," *ArXiv*, vol. abs/1706.03815, 2017.

[40] O. J. Räsänen and K. Khorrami, "A computational model of early language acquisition from audiovisual experiences of young infants," in *Interspeech*, 2019.

[41] W. N. Havard, J.-P. Chevrot, and L. Besacier, "Models of visually grounded speech signal pay attention to nouns: A bilingual experiment on english and japanese," *ICASSP*, pp. 8618–8622, 2019.

[42] ——, "Word recognition, competition, and activation in a model of visually grounded speech," in *Conference on Computational Natural Language Learning*, 2019.

[43] K. Khorrami and O. J. Räsänen, "Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? - a computational investigation," *Language Dev. Research*, 2021.

[44] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, pp. 211–252, 2014.

[45] J. Shi and J. Malik, "Normalized cuts and image segmentation," *CVPR*, 1997.

[46] I. Malioutov and R. Barzilay, "Minimum cut model for spoken lecture segmentation," in *ACL*, 2006.

[47] D. F. Harwath and T. J. Hazen, "Topic identification based extrinsic evaluation of summarization techniques applied to conversational speech," *ICASSP*, 2012.

[48] W.-N. Hsu *et al.*, "Text-free image-to-speech synthesis using learned segmental units," in *ACL*, 2021.

[49] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[50] P. Lippus *et al.*, "Phonetic corpus of estonian spontaneous speech," *Institute of Estonian and General Linguistics, University of Tartu. DOI: https://doi. org/10.15155/TY. D*, 2013.

[51] O. Räsänen, U. K. Laine, and T. Altosaar, "An improved speech segmentation quality measure: the r-value," in *Interspeech*, 2009.

[52] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," *TASLP*, 2007.

[53] R. Villing, J. Timoney, and T. E. Ward, "Automatic blind syllable segmentation for continuous speech," 2004.

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

[55] S. Bhati *et al.*, "Self-expressing autoencoders for unsupervised spoken term discovery," in *Interspeech*, 2020.