TEXTLESS PHRASE STRUCTURE INDUCTION FROM VISUALLY-GROUNDED SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

We study phrase structure induction from visually-grounded speech without intermediate text or text pre-trained models. The core idea is to first segment the speech waveform into sequences of word segments, then induce phrase structure based on the inferred segment-level continuous representations. To this end, we present the Audio-Visual Neural Syntax Learner (AV-NSL) that learns non-trivial phrase structure by listening to audio and looking at images, without ever reading text. Experiments on SpokenCOCO, the spoken version of MSCOCO with paired images and spoken captions, show that AV-NSL infers meaningful phrase structures similar to those learned from naturally-supervised text parsing, quantitatively and qualitatively. The findings in this paper extend prior work in unsupervised language acquisition from speech and grounded grammar induction, and manifest one possibility of bridging the gap between the two fields.

1 INTRODUCTION

Toddlers learn their first language through listening, talking, and interacting with the world through multi-sensory inputs. Different levels of early language acquisition happen without supervisory feedback (Dupoux, 2018): phonetics, phonology, morphology, syntax, semantics, pragmatics. It is therefore crucial to think about learning language, from identifying lower-level phones or words to inducing high-level linguistic structure like grammar, in natural settings.[1] To this end, there have been two ongoing efforts in parallel:

- Zero-resource speech processing, where speech models are constructed without any textual intermediates, with the goal of mimicking how children learn to speak before learning to read or write. The modeling tasks are constrained to unsupervised learning of subphones, phones, and words (Jansen et al., 2013).
- Grammar induction, which aims to learn latent syntactic structures, including constituency trees and dependency trees, with no annotation of syntactic structures as supervision.

Notably in recent years, multi-modal induction has emerged as a promising and effective objective for both efforts. In speech, Harwath (2018) proposed to leverage parallel image-speech data to acquire associated words (Harwath & Glass, 2017) and phones (Harwath et al., 2020) from raw waveforms. In syntax induction, Shi et al. (2019) proposed to induce phrase-structure grammar from parallel image-text data. The above observations motivated us to build a computational model that leverages the visual modality to acquire low-level words up to high-level phrase-structure from raw speech waveforms, without any intermediate textual forms or any direct supervision.[2]

In this paper, we present the Audio-Visual Neural Syntax Learner (AV-NSL), an approach toward learning phrase structure from raw speech waveforms without relying on any kind of intermediate textual form or text pre-trained models (Figure 1). In a nutshell, AV-NSL trains a visually-grounded syntax learner directly on a sequence of *continuous* speech representations given by an audio-visual word segmentation model. We also introduce a self-training process and an unsupervised decoding method to improve the final output of in AV-NSL. To measure the effectiveness of AV-NSL, we

[1] *Natural settings* here means situations that are similar to human language learning; that is, we are able to access parallel data from different modalities, while the amount of data is limited.

[2] It is worth noting that human language acquisition in similar settings has also attracted the attention from the developmental psychology community (Mason, 1980; Naigles, 1990; Dupoux, 2018, *inter alia*). For instance, from Dupoux (2018): "*Yet . . . there is still a large gap between models that learn from speech, which are limited to the discovery of phonemes and word forms, and models that learn syntax and semantics, which only work from textual input.*"

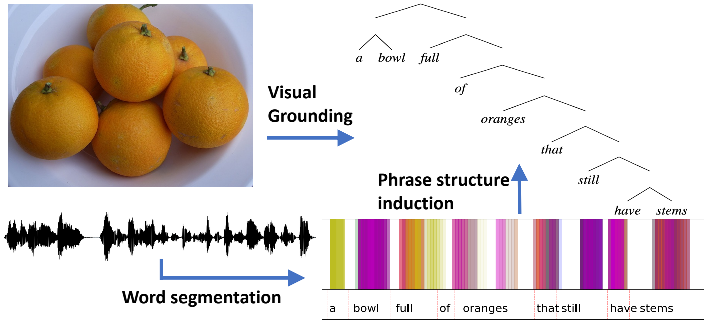


Figure 1: We study the process of inducing phrase structure, in the form of constituency parse tree, on unsupervised inferred word segments from raw speech waveform. No intermediate text tokens or ASR is needed. For illustration purpose, here we show the gold parse tree from the given text caption.

compare it to text-based syntax learner VG-NSL (Shi et al., 2019) and further introduce a novel evaluation metric, SAIoU, that accounts for structure differences when the number of tree nodes are mismatched. To validate our design choice of AV-NSL, we construct several baselines and introduce alternative modeling choices, including acoustic compound-PCFG (Kim et al., 2019a). Qualitatively, we provide constituency recall analyses and the visualizations of the inferred word segmentation and tree structures.

In summary, we present the first study on inducing phrase structure from visually-grounded speech without relying on text, introducing the AV-NSL model (§3) with comprehensive experiments (§4) and analysis (§5). As a by product, we improve over the previous state of the art in unsupervised word segmentation (§4.4).

2 Related Work

2.1 Unsupervised and Distantly Supervised Grammar Induction

Much work has been proposed to induce grammar from different sources of distant supervision, including language modeling (Shen et al., 2018; 2019; Kim et al., 2019a;b), masked language modeling (Drozdov et al., 2019), natural language inference (Li et al., 2019), and, more recently, visual grounding via image-caption matching (Shi et al., 2019; Zhao & Titov, 2020; Hong et al., 2021; Wan et al., 2022, *inter alia*). There has also been extensive study directly targeting unsupervised constituency parsing (Klein & Manning, 2002; 2004; Bod, 2006; Spitkovsky et al., 2013, *inter alia*). To the best of our knowledge, existing work on grammar induction from distant supervision has been based almost exclusively on text input. The most relevant work to ours is MMC-PCFG (Zhang et al., 2021), where speech features are treated as an auxiliary input for video-text grammar induction. However, text data and an off-the-shelf automatic speech recognition (ASR) model are required. In contrast to them, AV-NSL induces constituency parse trees from raw speech bypassing text, with distant supervision from parallel audio-visual data.

2.2 Unsupervised language acquisition from speech

The earliest work (de Sa, 1994; De Marcken, 1996; Roy & Pentland, 2002) on language acquisition from speech required phonetic lexicon/labels in the process. The idea of spoken term discovery, i.e., discovering repetitive patterns or keywords from unannotated speech, was first addressed by Park & Glass (2007). Thereafter, subsequent work improved upon the original (Zhang & Glass, 2009; Jansen & Van Durme, 2011; McInnes & Goldwater, 2011; Zhang, 2013, *inter alia*). Other related work has considered tasks like unsupervised word segmentation and unsupervised ASR, sometimes jointly with spoken term discovery (Lee & Glass, 2012; Lee et al., 2015; Kamper et al., 2015; 2017; Kamper & van Niekerk, 2021; Chorowski et al., 2021; Bhati et al., 2021; Kamper, 2022; Algayres et al., 2022) The discovery of lexical units was applied to text-free language modeling (Nguyen et al., 2020; Peng & Harwath, 2022a) and speech generation (Lakhotia et al., 2021; Polyak et al., 2021; Kharitonov et al., 2022). The ZeroSpeech challenges (Versteegh et al., 2015; Dunbar et al., 2017; 2019; 2020; Nguyen et al., 2020) have been a major driving force in the field.

Harwath (2018) opened up a new direction in visually grounded language acquisition, showing word-like (Harwath & Glass, 2017) and phone-like (Harwath et al., 2020) units are acquired from speech by analyzing audio-visual retrieval models. Numerous works have studied the characteristics of the linguistic information acquired in visually grounded speech models (Havard et al., 2019; Khorrami & Räsänen, 2021; Olaleye & Kamper, 2021; Wang & Hasegawa-Johnson, 2021; Mitja Nikolaus, 2022). Peng & Harwath (2022b) shows that clear word segmentation and identification naturally emerge from a visually grounded, self-supervised speech model named VG-HuBERT, by analyzing the model's self-attention heads. Unlike the above, AV-NSL acquires phrase structure, in the form of constituency parsing on top of unsupervised word segments.

2.3 Speech parsing and its applications

Early work on speech parsing can be traced back to the SParseval toolkit (Roark et al., 2006), for evaluating text parsers given (errorful) ASR output. Tran et al. (2018; 2019); Tran & Ostendorf (2021) explored the use of acoustic-prosodic features for text parsing with auxiliary speech input. Lou et al. (2019) trained a text parser (Kitaev & Klein, 2018) to detect speech disfluencies. In the past, syntax has also been studied in the context of speech prosody (Wagner & Watson, 2010; Köhn et al., 2018). The most relevant work to ours is Pupier et al. (2022), where a text dependency parser is trained from speech jointly with an ASR model. Moreover, text syntax parsing has been applied to prosody modeling in end-to-end text-to-speech (TTS; Guo et al., 2019; Tyagi et al., 2020; Kaiki et al., 2021). This work builds on top of pre-existing text parsing algorithms or pre-existing phrase structures from text, whereas we study phrase structure acquisition in the absence of text.

3 Method

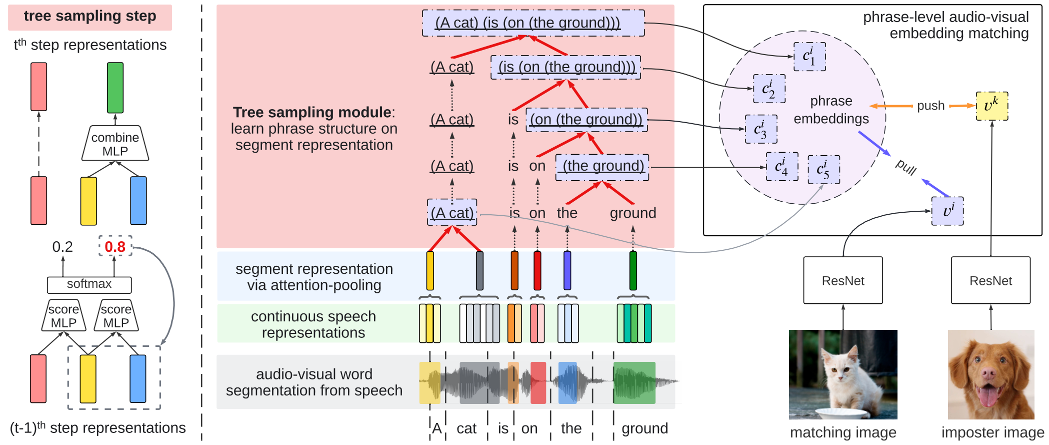

tree sampling step
t^{th} step representations
combine MLP
0.2
0.8
softmax
score MLP
score MLP
$(t-1)^{th}$ step representations
Tree sampling module: learn phrase structure on segment representation
(A cat) (is (on (the ground)))
(A cat)
(is (on (the ground)))
is
(on (the ground))
on
(the ground)
the
ground
segment representation via attention-pooling
continuous speech representations
audio-visual word segmentation from speech
A
cat
is
on
the
ground
phrase-level audio-visual embedding matching
phrase embeddings
c_1^i
c_2^i
c_3^i
c_4^i
c_5^i
push
pull
v^k
v^i
ResNet
ResNet
matching image
imposter image

Figure 2: Illustration of AV-NSL, which extends VG-NSL (Shi et al., 2019) to audio-visual inputs.

Given a set of paired spoken captions and images, the Audio-Visual Neural Syntax Learner (AV-NSL) infers phrase structures from subsequences of raw speech segments without relying on text. The basis of AV-NSL is the Visually-Grounded Neural Syntax Learner (VG-NSL) (Shi et al., 2019). VG-NSL learns constituency parse trees by guiding a sequential tree sampling process with text-image matching. To extend VG-NSL to audio-visual inputs, the central challenge is extracting *semantically-meaningful word segments* from unannotated speech. We break down the problem into a two-step process: (1) obtaining sequences of word segments, and (2) extracting segment-level self-supervised representations. With these simple modifications, AV-NSL learns non-trivial phrase structure without ever reading text, instead by listening to speech and looking at images.

3.1 Background: Visually-Grounded Neural Syntax Learner

VG-NSL (Shi et al., 2019) is composed of a bottom-up text parser and a text-image embedding matching module. The parser consists of an embedding similarity scoring function *score* and an embedding combination function *combine*. Given a text caption, denoted by a sequence of word embeddings $W = \{w_i^0\}_{i=1}^N$ of length N, the parser synthesizes a constituency parse tree by recursively scoring and combining adjacent embeddings at each step. At step t, VG-NSL (1) evaluates all consecutive pairs of embeddings $\langle w_i^t, w_{i+1}^t \rangle$ and assigns a scalar score to each with *score*, (2) selects a pair $\langle w_{i'}^t, w_{i'+1}^t \rangle$ based on the corresponding scores,[3] and (3) combines the selected pair of embeddings via *combine* to form a new phrase embedding for the next step, copying the remaining ones to the next step. In VG-NSL, *score* is parameterized by a 2-layer ReLU-activated MLP, and *combine* is defined by the L2-normalized sum of the input embeddings. The resulting tree is inherently binary and there are $N-1$ combining steps in total, as the tree parser must combine two nodes in each step.

The text-image embedding matching module of VG-NSL is based on the standard hinge-based triplet loss (Kiros et al., 2014), where the sentence-based loss is modified to a phrase-based one. Additionally, the loss function is adapted to estimate the visual *concreteness* of a text span: intuitively, the smaller the loss related to a candidate constituent c, the larger the concreteness of c, and vice versa. The concreteness of a constituent c is defined as

$$\textit{concrete}\left(\mathbf{c};\mathbf{i}\right) = \sum_{\mathbf{c}'} \left[\cos\left(\mathbf{i},\mathbf{c}\right) - \cos\left(\mathbf{i},\mathbf{c}'\right) - \delta\right]_+ + \sum_{\mathbf{i}'} \left[\cos\left(\mathbf{i}',\mathbf{c}\right) - \cos\left(\mathbf{i}',\mathbf{c}\right) - \delta\right]_+ ,$$

where $\mathbf{c}$ is the vector representation of c; $\mathbf{i}$ is the corresponding vector of the parallel image of c; $\mathbf{c}'$ is a candidate constituent from a sentence that is not in parallel with $\mathbf{i}$; $\mathbf{i}'$ is an image that is not in parallel with c; δ is a constant margin. Here, $[\cdot]_+ := \max(\cdot, 0)$. Finally, the estimated concreteness scores are passed back to the parser as rewards to the constituents. VG-NSL jointly optimizes the visual-semantic embedding loss, and trains the parser with REINFORCE (Williams, 1992).

3.2 Audio-Visual Neural Syntax Learner

AV-NSL extends VG-NSL by: (1) incorporating an audio-visual word segmentation model for obtaining sequences of word segments from unannotated speech, (2) jointly optimizing segment-level embeddings along with phrase structure induction, and (3) employing deeper *score* and *combine* function parameterization in the parsing module. We empirically found (3) necessary, mainly because speech embeddings are inherently richer, less clean, and semantically more ambiguous than word embeddings. In AV-NSL, *score* is parameterized by a 4-layer MLP with GELU nonlinearities (Hendrycks & Gimpel, 2016), and *combine* is a 5-layer MLP with GELUs. On the other hand, such parameterization may cause the text-based sampling procedure to favor sampling the visually-salient words (Shi et al., 2019; Kojima et al., 2020). We describe (1) and (2) in detail as follows.

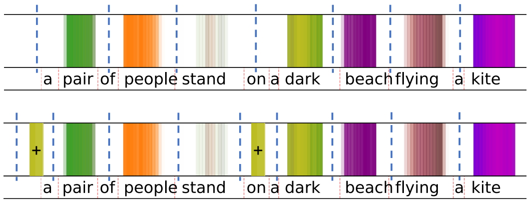


Figure 3: Example of word segmentation from VG-HuBERT (top). We use the midpoints of adjacent attention boundaries (vertical blue dashed lines) as the word boundaries. We observe that function words are ignored by VG-HuBERT; to account for this, we introduce *segment insertion* (bottom): short segments are placed in long enough gaps between existing segments, such that function words are recovered. Inserted segments are marked with "+". Best viewed in color.

Audio-visual word segmentation: AV-NSL leverages VG-HuBERT Peng & Harwath (2022b) for word segmentation (Figure 2; bottom). VG-HuBERT is trained to associate spoken captions with natural images via retrieval training, without any textual supervision. After training, spoken word segmentation emerges via magnitude thresholding the self-attention heads of the model's audio encoder: at layer l, we threshold each CLS token attention weights over each temporal speech frame token to only show top $p\%$ of the magnitude. In Figure 3, we visualize the attention weights that each speech frame receives from the CLS token. Weights from different attention heads are plotted in different colors, and color transparency represents the magnitude of the attention weights.

However, an issue we observed with VG-HuBERT is that they tend to ignore function words such as "a", "the", and "of". While this is less of an issue for word segmentation and identification, it is problematic for our purpose, as the function words are critical for phrase induction. Therefore, we devise a simple heuristic to pick up function words' segments – *segment insertion*. We insert a short word segment whenever there is a sufficiently long enough gap of s seconds, and VG-HuBERT fails to place an attention segment. See bottom of Figure 3. Since this could introduce

[3] In the training stage, the pair is sampled from a distribution where the probability of a pair is proportional to $\exp(\textit{score})$; in the inference stage, the $\arg\max$ is selected.

false positives (inserting segments where there is no word spoken), we apply unsupervised voice activity detection (Tan et al., 2020) to further restrict segment insertion only in voiced regions. The length of the insertion gap s, the VG-HuBERT segmentation layer l, attention magnitude threshold $p\%$, and model training snapshots over different random seeds and training steps, are all determined in an unsupervised fashion with minimal Bayes' risk decoding, introduced in Section 3.4.

Speech segment representations: Given the word segments from the audio-visual segmentation model, segment representations are extracted as inputs for the tree sampling module. Ideally, these segments should be semantically-meaningful and mimic word embeddings $W = \{w_i^0\}_{i=1}^N$. A naive method is speech discretization that converts the inputs into sequences of discrete tokens (Lakhotia et al., 2021). Yet, we are targeting word-level phrase structures, while speech discretization, namely acoustic unit discovery, are sub-phone level, which does not fit into our setup. Different from it, AV-NSL is based on *continuous* segment-level self-supervised representations. Let's denote the frame-level representation sequence as $R = \{r_j\}_{j=1}^T$, where T is the speech sequence length. Audio-visual word segmentation returns an alignment $A(i) = r_{p:q}$ that maps the ith word segment to the pth to qth acoustic frames. The segment-level continuous representation for the ith word is simply,

$$w_i^0 = \sum_{t \in A(i)} a_{it} r_{it}$$

where a_{it} is the attention weights over the segments specified by $A(i)$. By default in AV-NSL, R is the layer representation from VG-HuBERT, and a_{it} is the CLS token attention weights over frames within each segment. In some cases, visual grounding is not available in AV-NSL's word segmentation, e.g. VG-HuBERT is not available. We instead take R as the layer representation from a vanilla HuBERT (Hsu et al., 2021a), and a_{it} is parameterized by a hidden layer that is jointly optimized with the tree sampling module. Despite its simplicity, AV-NSL learns meaningful phrase structures on these segment representation sequences.

3.3 Self-Training

A self-training procedure is introduced for AV-NSL to further improve its parsing capability. Previously, it has been shown that self-training consistently improves the performance of text-based unsupervised constituency parsing. In Shi et al. (2020), the self-training model was based on Benepar (Kitaev & Klein, 2018), a supervised neural constituency parser, which (1) takes a sentence as the input, (2) maps it to word representations, and (3) predicts a score for any constituency parse tree. In the inference stage, the model evaluates all possible tree structures and outputs the highest-scoring one using the CKY algorithm (Kasami, 1966; Younger, 1967; Cocke, 1969).

In this work, we introduce s-Benepar, which is based on the original Benepar, except the model input is the segment-level continuous HuBERT representations mean-pooled over unsupervised word segmentation from VG-HuBERT with segment insertion, and model output is AV-NSL's inferred constituency parse from Section 3.2. We also removed part-of-speech tag prediction as in Benepar, as there is no textual supervision in our setting. To summarize, with paired speech D_A and image D_V data, the training scheme for AV-NSL with self-training is as follows:

1. Train an AV-NSL from audio-visual data (D_A, D_V) and obtain the trained model M_{av}.
2. Generate parse tree T_0 with M_{av} for D_A. Obtain audio-tree pairs (D_A, T_0). Set $T = T_0$.
3. Train an s-Benepar from (D_A, T) and obtain the trained model M_s^i.
4. Generate parse tree T_i with M_s^i for D_A. Obtain audio-tree pairs (D_A, T_i). Set $T = T_i$.
5. Go to Step 3 if we have not reached the desirable number of iterations; return T otherwise.

We find it helpful to iterate s-Benepar training twice ($i = 2$), but the results plateau afterwards.

3.4 Unsupervised Decoding

One key ingredient of AV-NSL is applying minimum Bayes risk (MBR) decoding (Bickel & Li, 1977) as the selection criterion for fully-unsupervised spoken word segmentation and phrase-structure induction.[4] Specifically, this is in contrast to all prior unsupervised word segmentation work, in which ground truth word segments from a development set are required for decoding.

[4] MBR decoding is widely adopted in machine translation (Kumar & Byrne, 2004; Zhang & Gildea, 2008; Shi et al., 2022, *inter alia*).

At a high level, given a loss function $\ell_{MBR}(O_1, O_2)$ between two outputs O_1 and O_2, and a set of k outputs $\mathcal{O} = \{O_1, \ldots, O_k\}$, we select the optimal output

$$\hat{O} = \arg\min_{O' \in \mathcal{O}} \sum_{O'' \in \mathcal{O}} \ell_{MBR}(O', O'').$$

For word segmentation, we define the loss between two segmentation proposals $\mathcal{S}_1$ and $\mathcal{S}_2$ by $\ell_{MBR}(\mathcal{S}_1, \mathcal{S}_2) = -\text{MIoU}(\mathcal{S}_1, \mathcal{S}_2)$, where $\text{MIoU}(\cdot, \cdot)$ denotes the mean intersection over union ratio across all matched pairs of predicted word spans from $\mathcal{S}_1$ and $\mathcal{S}_2$. We match the predicted word spans using the maximum weight matching algorithm (Galil, 1986), where word spans correspond to vertices, and we define edge weights by the temporal overlap between the corresponding spans.

For phrase structure induction, we define the loss function between two parse trees $\mathcal{T}_1$ and $\mathcal{T}_2$ by $\ell_{MBR}(\mathcal{T}_1, \mathcal{T}_2) = 1 - F_1(\mathcal{T}_1, \mathcal{T}_2)$, where $F_1(\cdot, \cdot)$ denotes the F_1 score between two trees.

4 Experiments

4.1 Setting

Dataset: All models are evaluated on SpokenCOCO, the spoken version of MSCOCO (Lin et al., 2014) where the text captions are read out by MTurk users (Hsu et al., 2021b). It contains 83k/5k/5k images for training, validation, and test: each image has 5 corresponding spoken captions. SpokenCOCO totals 740h of read speech from 2.3k speakers, with an average utterance duration of about 4 seconds, covering 29K different word types.

Preprocessing: For oracle word segmentation, we ran an off-the-shelf English ASR from Montreal Force Aligner (McAuliffe et al., 2017) that was pre-trained on Librispeech and adapted to SpokenCOCO. We removed a few utterances that have mismatches in their ASR transcripts and their text captions. Following Shi et al. (2019), we included trivial spans in tree evaluation. Additionally, we ran an off-the-shelf English parser (Kitaev & Klein, 2018) on the ASR transcript (normalized text with punctuation removed) to generate the oracle trees for SpokenCOCO.

4.2 Baselines and Toplines

AV-NSL segments speech waveforms into word segments, then learns phrase structures on top of the learned segments. Both segmentation and structure induction are fully-unsupervised and visually-grounded. To help us examine the role of each component in AV-NSL, we therefore further construct the following baselines and toplines. Their full descriptions are in Appendix A.1.

Trivial tree structures: Following (Shi et al., 2019), we include baselines without linguistic information: random binary trees, left-branching binary trees, and right-branching binary trees.

AV-cPCFG: We train compound probabilistic context free grammar (cPCFG) (Kim et al., 2019a) on word-level discrete speech tokens. Similar to AV-NSL, word segments and segment representations are based on VG-HuBERT. Different from AV-NSL, the segment representations are discretized via kmeans to obtain word-level discrete indices. In short, AV-cPCFG leverages visual cues only for segmentation and segment representations, but not for phrase structure induction.

DPDP-cPCFG: Instead of training cPCFG on audio-visual word segments and audio-visual segment representations, DPDP-cPCFG does not rely on any visual grounding throughout. Instead, DPDP (Kamper, 2022) and vanilla HuBERT representations are used. As in AV-cPCFG, kmeans is used for word-level discretization.

Oracle AV-NSL: To remove the uncertainty of unsupervised word segmentation, we directly train AV-NSL on top of oracle word segmentation via force alignment.

4.3 Evaluation Metric

Word segmentation. We use the standard word boundary prediction metrics (precision, recall and F1), which are calculated by comparing the temporal position between inferred word boundaries and force aligned word boundaries. In particular, following Peng & Harwath (2022b), when an inferred boundary is located within ±20*ms* of a force aligned boundary, we declare a successful prediction.

Parsing. For parsing with oracle word segmentation, we use `EVALB` to calculate the F_1 score between the predicted and ground-truth parse trees.[5] For parsing with inferred word segmentation, due to the mismatch in the number of nodes between the predicted and ground-truth parse trees, we introduce the structured average intersection-over-union ratio (SAIoU) as an additional metric.

SAIoU takes both word segmentation quality and temporal overlap between induced constituents into consideration. Concretely, the input is two constituency parse trees over the same speech utterance $\mathcal{T}_1 = \{c_{1,i} = (\ell_{1,i}, r_{1,i})\}_{i=1}^{n_1}$ and $\mathcal{T}_2 = \{c_{2,j} = (\ell_{2,j}, r_{2,j})\}_{j=1}^{n_2}$, represented by a set of constituency temporal boundaries ℓ and r. We first compute the optimal valid alignment between the constituents in $\mathcal{T}_1$ and $\mathcal{T}_2$, $\hat{\mathcal{A}} = \arg\max_{\textit{valid}\ \mathcal{A}} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathcal{A}_{i,j} \text{IoU}(c_{1,i}, c_{2,j})$, where $\mathcal{A}_{i,j} = 1$ denotes $c_{1,i}$ aligns with $c_{2,j}$, and $\mathcal{A}_{i,j} = 0$ otherwise; $\text{IoU}(\cdot, \cdot)$ denotes the intersection-over-union ratio between two spans. A valid alignment $\mathcal{A}$ is one that satisfies the following conditions:

1. Any constituent may be aligned with up to 1 constituent in the other tree;

2. For any pair of i and j where $\mathcal{A}_{i,j} = 1$,

- Any descendant of $c_{1,i}$, $c_{1,k}$, may either align to a descendant of $c_{2,j}$ or be left unaligned;
- Any ancestor of $c_{1,i}$, $c_{1,k'}$, may either align to a ancestor of $c_{2,j}$ or be left unaligned;
- Any descendant of $c_{2,j}$, $c_{2,p}$, may either align to a descendant of $c_{1,i}$ or be left unaligned;
- Any ancestor of $c_{2,j}$, $c_{2,p'}$, may either align to a ancestor of $c_{1,i}$ or be left unaligned.

Given the optimal alignment $\hat{\mathcal{A}}$, we calculate the structured average IoU between $\mathcal{T}_1$ and $\mathcal{T}_2$ by

$$\text{SAIoU}(\mathcal{T}_1, \mathcal{T}_2) = \frac{2}{n_1 + n_2} \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \hat{\mathcal{A}}_{i,j} \text{IoU}(c_{1,i}, c_{2,j}) \right).$$

4.4 Unsupervised Word Segmentation

We validate our decision of adopting VG-HuBERT to extract word-like units from raw speech waveforms for later phrase structure parsing. In particular, we investigate two questions: (1) How does segment insertion affect word segmentation performance? (2) how does MBR-based VG-HuBERT compare to supervised selected VG-HuBERT?

In Table 1, in addition to VG-HuBERT, we also list a speech-only word segmentation algorithm DPDP (Kamper, 2022). Note that audio-visual model VG-HuBERT significantly outperform DPDP. For question (1), by comparing the third row and the fourth row, as expected we see that performing segment insertion improves recall and hurts precision, and slightly improves F1. For question (2), by comparing the fourth row and the fifth row (second to last row), we see that MBR selection actually leads to better performance than supervised selection. The final MBR selection we adopted is based on the last row, where we first performed MBR selection on SpokenCOCO val set on all 405 candidates, and subsequently chose the 10 most selected combinations to perform another round of MBR decoding. Getting the top 10 most selected combinations does not require knowing the performance on segmentation, and therefore this process is still completely unsupervised. The reason for doing 2 iterations of MBR is because performing MBR on 405 candidates on SpokenCOCO training set is estimated to take 2 months, and MBR on 10 candidates can be done in 5 days. Comparing the last two rows, we observe that two iterations of MBR does not lead to worse results.

4.5 Unsupervised Phrase Structure Induction

We quantitatively show that AV-NSL learns meaningful phrase structure given word segments. First, Table 2 is the main result of the fully-unsupervised AV-NSL on SpokenCOCO, evaluated with SAIoU. The best performing AV-NSL is based on our improved VG-HuBERT with MBR top 10 selection for word segmentation, attention-weighted mean-pool over VG-HuBERT layers as the segment representations, and another MBR decoding over all phrase structure induction hyperparameters. Comparing AV-NSL against AV-cPCFG and AV-cPCFG against DPDP-cPCFG, we empirically show the necessity of training AV-NSL on *continuous* segment representation instead of discretized speech tokens, and the effectiveness of visual-grounding in our overall model design.

[5] `https://nlp.cs.nyu.edu/evalb/`

Method	Insertion	Out. Sel.	#Sel. Cand.	Precision	Recall	F_1
DPDP (Kamper, 2022)		supervised		17.37	9.00	11.85
VG-HuBERT (Peng & Harwath, 2022b)		supervised		36.19	27.22	31.07
Improved VG-HuBERT (Ours)	✓	supervised		34.34	29.85	31.94
	✓	MBR	405	33.83	34.37	34.10
	✓	MBR (2iter)	405 →10	33.31	34.90	34.09

Table 1: Word Segmentation Performance on SpokenCOCO validation set. Out. Sel. denotes output selection methods, and #Sel. Cand. denotes the number of candidate models to be selected. MBR (2iter) means we first run MBR on all 405 candidates, and then run MBR again on the 10 most selected candidates. Our improved VG-HuBERT with MBR achieves the best boundary F_1.

Model			Output	SAIoU
Syntax Induction	Segmentation	Seg. Representation (continuous/discrete)	Selection	
Right-Branching	VG-HuBERT+MBR$_{10}$			**0.546**
Right-Branching	DPDP			0.478
AV-NSL	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$ (continuous)	MBR	0.516
AV-NSL	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10,11,12}$ (continuous)	MBR	**0.521**
AV-cPCFG	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$+4k km (discrete)	last ckpt.	0.499
AV-cPCFG	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$+8k km (discrete)	last ckpt.	0.481
DPDP-cPCFG	DPDP	HuBERT$_2$+2k km (discrete)	last ckpt.	0.465
DPDP-cPCFG	DPDP	HuBERT$_{10}$+2k km (discrete)	last ckpt.	0.426

Table 2: Fully-unsupervised phrase structure induction results on SpokenCOCO. The best overall number and the best number produced by neural models are in **boldface**. Full table in Appendix 7.

Model	Segmentation	Seg. Representation	tree target			Output	SAIoU
			train	val	test	Selection	
s-Benepar	VG-HuBERT+MBR$_{10}$	HuBERT$_2$	AV-NSL	AV-NSL	oracle	last ckpt.	**0.538**
s-Benepar	VG-HuBERT+MBR$_{10}$	HuBERT$_6$	AV-NSL	AV-NSL	oracle	last ckpt.	**0.538**
s-Benepar	VG-HuBERT+MBR$_{10}$	HuBERT$_{2,4,6,8,10,12}$	AV-NSL	AV-NSL	oracle	MBR	0.536

Table 3: Single round self-training in Section 3.3 improves the best AV-NSL from Table 2. We train s-Benepar on the trees from fully-unsupervised AV-NSL. Full table in Appendix 8.

Secondly, Table 3 shows that our proposed self-training with s-Benepar complements AV-NSL. Generally, a single round of self-training improves the SAIoU, and our best s-Benepar improves the best AV-NSL from 0.521 to 0.538. Thirdly, Table 4 isolates phrase structure induction from word segmentation quality with oracle AV-NSL. Different from Table 2, since there is no mismatch in the number of tree nodes, we can adopt F_1 evaluation. With proper segment-level representations, unsupervised oracle AV-NSL matches or out-performs text-based VG-NSL. Similar to Tabel 3, self-training with s-Benepar on oracle AV-NSL trees further improves the syntax induction results, almost matching that of right-branching tree. Last but not least, perhaps surprisingly, right-branching trees (RBT) on the given word segmentation reach the best SAIoU and F_1 scores. We note that the right-branching approach highly aligns with the head-initial property of English (Baker, 2001), especially in our setting where all punctuation marks were removed; thus, it is nontrivial for AV-NSL to reach the performance on par with RBT without inductive biases favoring any specific type of trees.

5 Analyses

Unsupervised Constituent Recall: Following Shi et al. (2019), we show the recall of specific types of constituents (Table 5). While VG-NSL benefits from the head-initial (HI) bias, where abstract words are encouraged to appear in the beginning of a constituent, it is worth noting that AV-NSL outperforms all variations of VG-NSL, without inductive biases favoring any specific types of trees.

Model		Output	F_1
Syntax Induction	Seg. Representation	Selection	
Random			32.77
Left-Branching			24.56
Right-Branching			**57.39**
VG-NSL	word embeddings	Supervised	53.11
oracle AV-NSL	log-Mel spectrogram	Supervised	42.01
oracle AV-NSL	HuBERT$_2$	Supervised	55.51
oracle AV-NSL	HuBERT$_2$	MBR	54.99
oracle AV-NSL	HuBERT$_{2,4,6,8,10,12,24}$	MBR	**55.96**
oracle AV-NSL → s-Benepar	HuBERT$_2$	MBR	57.24
oracle AV-NSL → s-Benepar	HuBERT$_{12}$	MBR	**57.33**

Table 4: Phrase structure induction with oracle segmentation given. Full table in Appendix 9.

Ablation Study: We present two ablations to examine the effectiveness of high-quality word segmentation and visual representation (Table 6). We train AV-NSL with the following modifications:

1. Fix the visual representations, but replace oracle segmentation with naive uniform word segmentation, where the number of words in each caption is given (uniform AV-NSL).
2. Fix the oracle word segmentation, but replace visual embeddings with random images, where each pixel is independently sampled from a uniform distribution.

We observe that there are significant performance drops in both settings, comparing to the AV-NSL trained with oracle segmentation and high-quality visual representation. This set of results complement Table 2, stressing that precise word segmentation and high-quality visual representations are both necessary for phrase structure induction from speech. Furthermore, we provide tree structure and word segmentation visualizations for qualitative analysis in the Appendix.

Table 5: Recall of specific typed phrases, including noun phrases (NP), verb phrases (VP), prepositional phrases (PP) and adjective phrases (ADJP), and overall F_1 score, evaluated on the SpokenCOCO test split. The VG-NSL numbers are taken from (Shi et al., 2019). AV-NSL here are trained on oracle segmentation with vanilla HuBERT as the layer representations.

Model	F_1	Constituent Recall			
		NP	VP	PP	ADJP
VG-NSL (Shi et al., 2019)	50.4	**79.6**	26.2	42.0	22.0
VG-NSL + HI	53.3	74.6	32.5	66.5	21.7
VG-NSL + HI + FastText	54.4	78.8	24.4	65.6	22.0
oracle AV-NSL	**55.6**	55.5	**68.1**	**66.6**	**22.1**

Table 6: Top rows: performance of AV-NSL with word segmentation in various quality and high-quality visual embeddings. Bottom rows: performance of AV-NSL with visual embeddings in various quality and high-quality word segmentation. DINO: a self-supervised model that produces high-quality visual representations (Caron et al., 2021).

Model		Visual	F_1
Syntax Induction	Seg. Repre.		
oracle AV-NSL	HuBERT$_{10}$	ResNet101	50.50
uniform AV-NSL	HuBERT$_{10}$	ResNet101	36.62
oracle AV-NSL	HuBERT$_2$	DINO	55.71
oracle AV-NSL	HuBERT$_2$	random	31.23

6 Conclusion

In recent years, there have been fruitful progresses in multi-modal induction for zero-resource speech processing and grammar induction respectively. The idea of leveraging the visual modality to learn language competence, either lexicon units from speech or syntactic structure from text, is an attractive approach for modeling human language acquisition. Our study contributes to both lines of research, by presenting an unifying framework that learns phrase structure from visually-grounded speech, without any text. We show that our proposed model, AV-NSL, infers meaningful constituency parse trees on top of continuous word segment representations, both quantitatively and qualitatively. To justify our modeling design choices, we construct several baselines and introduce a novel evaluation metric. We envision our research as the first of many in textless structure learning.

Ethics Statement

This work is scientific at its core, as the goal is to study the process of grammar induction from speech with visual grounding. The data used in this work is also publicly available. One potential concern is that the data and experiments are based on English, which does not represent the global human population. However, we believe that our proposed method is general enough to be applied to other spoken languages when the data is available, because we do not use any language specific speech processing techniques, and we do not have any built-in bias within the models.

Reproducibility Statement

AV-NSL code, s-Benepar code, and SAIoU evaluation code will be made publicly available. AV-NSL code is based on the VG-NSL codebase. s-Benepar code is based on the Benepar codebase. SpokenCOCO is publicly available to download. All models are trained on a single GPU. We also included as many experimental details as we can in the main content of the paper and in Appendix A.2.

References

Robin Algayres, Tristan Ricoul, Julien Karadayi, Hugo Laurençon, Salah Zaiem, Abdelrahman Mohamed, Benoît Sagot, and Emmanuel Dupoux. Dp-parse: Finding word boundaries from raw speech with an instance lexicon. *arXiv preprint arXiv:2206.11332*, 2022.

Mark C Baker. *The atoms of language.* Basic Books, 2001.

Saurabhchand Bhati, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velazquez, and Najim Dehak. Segmental contrastive predictive coding for unsupervised word segmentation. *Interspeech*, 2021.

Peter J Bickel and Bo Li. Mathematical statistics. In *Test*. Citeseer, 1977.

Rens Bod. An all-subtrees approach to unsupervised parsing. *COLING-ACL*, 2006.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Jan Chorowski, Grzegorz Ciesielski, Jarosław Dzikowski, Adrian Łańcucki, Ricard Marxer, Mateusz Opala, Piotr Pusz, Paweł Rychlikowski, and Michał Stypułkowski. Aligned contrastive predictive coding. *Interspeech*, 2021.

John Cocke. *Programming languages and their compilers: Preliminary notes*. New York University, 1969.

Carl De Marcken. *Unsupervised language acquisition*. PhD thesis, Massachusetts Institute of Technology, 1996.

Virginia R de Sa. Learning classification with unlabeled data. *NeurIPS*, 1994.

Andrew Drozdov, Pat Verga, Mohit Yadav, Mohit Iyyer, and Andrew McCallum. Unsupervised latent tree induction with deep inside-outside recursive autoencoders. *NAACL-HLT*, 2019.

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. The zero resource speech challenge 2017. *ASRU*, 2017.

Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W Black, et al. The zero resource speech challenge 2019: Tts without t. *Interspeech*, 2019.

Ewan Dunbar, Julien Karadayi, Mathieu Bernard, Xuan-Nga Cao, Robin Algayres, Lucas Ondel, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. The zero resource speech challenge 2020: Discovering discrete subword and word units. *Interspeech*, 2020.

Emmanuel Dupoux. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59, 2018.

Zvi Galil. Efficient algorithms for finding maximum matching in graphs. *ACM Comput. Surv.*, 18 (1):23–38, mar 1986. ISSN 0360-0300. doi: 10.1145/6462.6502. URL `https://doi.org/10.1145/6462.6502`.

Haohan Guo, Frank K Soong, Lei He, and Lei Xie. Exploiting syntactic features in a parsed tree to improve end-to-end tts. *Interspeech*, 2019.

David Harwath and James R Glass. Learning word-like units from joint audio-visual analysis. *ACL*, 2017.

David Harwath, Wei-Ning Hsu, and James Glass. Learning hierarchical discrete linguistic units from visually-grounded speech. *ICLR*, 2020.

David Frank Harwath. *Learning spoken language through vision*. PhD thesis, Massachusetts Institute of Technology, 2018.

William N. Havard, Jean-Pierre Chevrot, and Laurent Besacier. Word recognition, competition, and activation in a model of visually grounded speech. In *CoNLL*, 2019.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Yining Hong, Qing Li, Song-Chun Zhu, and Siyuan Huang. Vlgrammar: Grounded grammar induction of vision and language. *ICCV*, 2021.

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*, 29:3451–3460, 2021a.

Wei-Ning Hsu, David Harwath, Christopher Song, and James Glass. Text-free image-to-speech synthesis using learned segmental units. *ACL*, 2021b.

Aren Jansen and Benjamin Van Durme. Efficient spoken term discovery using randomized algorithms. *ASRU*, 2011.

Aren Jansen, Emmanuel Dupoux, Sharon Goldwater, Mark Johnson, Sanjeev Khudanpur, Kenneth Church, Naomi Feldman, Hynek Hermansky, Florian Metze, Richard Rose, et al. A summary of the 2012 jhu clsp workshop on zero resource speech technologies and models of early language acquisition. *ICASSP*, 2013.

Nobuyoshi Kaiki, Sakriani Sakti, and Satoshi Nakamura. Using local phrase dependency structure information in neural sequence-to-sequence speech synthesis. In *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pp. 206–211. IEEE, 2021.

Herman Kamper. Word segmentation on discovered phone units with dynamic programming and self-supervised scoring. *arXiv preprint arXiv:2202.11929*, 2022.

Herman Kamper and Benjamin van Niekerk. Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks. *Interspeech*, 2021.

Herman Kamper, Aren Jansen, and Sharon Goldwater. Fully unsupervised small-vocabulary speech recognition using a segmental bayesian model. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Herman Kamper, Aren Jansen, and Sharon Goldwater. A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*, 46:154–174, 2017.

Tadao Kasami. An efficient recognition and syntax-analysis algorithm for context-free languages. *Coordinated Science Laboratory Report no. R-257*, 1966.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, et al. Text-free prosody-aware generative spoken language modeling. *ACL*, 2022.

Khazar Khorrami and Okko Johannes Räsänen. Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? - a computational investigation. *ArXiv*, abs/2109.14200, 2021.

Yoon Kim, Chris Dyer, and Alexander M Rush. Compound probabilistic context-free grammars for grammar induction. *ACL*, 2019a.

Yoon Kim, Alexander M Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. Unsupervised recurrent neural network grammars. *NAACL-HLT*, 2019b.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In *ACL*, 2018.

Dan Klein and Christopher D Manning. A generative constituent-context model for improved grammar induction. *ACL*, 2002.

Dan Klein and Christopher D Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. *ACL*, 2004.

Arne Köhn, Timo Baumann, and Oskar Dörfler. An empirical analysis of the correlation of syntax and prosody. *Interspeech*, 2018.

Noriyuki Kojima, Hadar Averbuch-Elor, Alexander M Rush, and Yoav Artzi. What is learned in visually grounded neural syntax acquisition. *ACL*, 2020.

Shankar Kumar and William Byrne. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 169–176, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL `https://aclanthology.org/N04-1022`.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021.

Chia-ying Lee and James Glass. A nonparametric bayesian approach to acoustic model discovery. *ACL*, 2012.

Chia-ying Lee, Timothy J O'donnell, and James Glass. Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, 3:389–403, 2015.

Bowen Li, Lili Mou, and Frank Keller. An imitation learning approach to unsupervised parsing. *ACL*, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Paria Jamshid Lou, Yufei Wang, and Mark Johnson. Neural constituency parsing of speech transcripts. *NAACL-HLT*, 2019.

Jana M Mason. When do children begin to read: An exploration of four year old children's letter and word reading competencies. *Reading Research Quarterly*, pp. 203–227, 1980.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. *Interspeech*, 2017.

Fergus McInnes and Sharon Goldwater. Unsupervised extraction of recurring words from infant-directed speech. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

Grzegorz Chrupała Mitja Nikolaus, Afra Alishahi. Learning english with peppa pig. *TACL*, 2022.

Letitia Naigles. Children use syntax to learn verb meanings. *Journal of child language*, 17(2): 357–374, 1990.

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. *Self-Supervised Learning for Speech and Audio Processing NeurIPS Workshop*, 2020.

Kayode Olaleye and Herman Kamper. Attention-based keyword localisation in speech using visual grounding. In *Interspeech*, 2021.

Alex S Park and James R Glass. Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):186–197, 2007.

Puyuan Peng and David Harwath. Self-supervised representation learning for speech using visual grounding and masked language modeling. *Self-Supervised Learning for Speech and Audio Processing Workshop at AAAI*, 2022a.

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

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech resynthesis from discrete disentangled self-supervised representations. *Interspeech*, 2021.

Adrien Pupier, Maximin Coavoux, Benjamin Lecouteux, and Jérôme Goulian. End-to-end dependency parsing of spoken french. In *Interspeech*, 2022.

Brian Roark, Mary Harper, Eugene Charniak, Bonnie Dorr, Mark Johnson, Jeremy G Kahn, Yang Liu, Mari Ostendorf, John Hale, Anna Krasnyanskaya, et al. Sparseval: Evaluation metrics for parsing speech. *LREC*, 2006.

Deb K Roy and Alex P Pentland. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146, 2002.

Yikang Shen, Zhouhan Lin, Chin-Wei Huang, and Aaron Courville. Neural language modeling by jointly learning syntax and lexicon. *ICLR*, 2018.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. Ordered neurons: Integrating tree structures into recurrent neural networks. *ICLR*, 2019.

Freda Shi, Daniel Fried, Marjan Ghazvininejad, Luke Zettlemoyer, and Sida I Wang. Natural language to code translation with execution. *arXiv preprint arXiv:2204.11454*, 2022.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. Visually grounded neural syntax acquisition. *ACL*, 2019.

Haoyue Shi, Karen Livescu, and Kevin Gimpel. On the role of supervision in unsupervised constituency parsing. In *EMNLP*. Association for Computational Linguistics, 2020.

Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. Breaking out of local optima with count transforms and model recombination: A study in grammar induction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1983–1995, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `https://aclanthology.org/D13-1204`.

Zheng-Hua Tan, Najim Dehak, et al. rvad: An unsupervised segment-based robust voice activity detection method. *Computer speech & language*, 59:1–21, 2020.

Trang Tran and Mari Ostendorf. Assessing the use of prosody in constituency parsing of imperfect transcripts. *Interspeech*, 2021.

Trang Tran, Shubham Toshniwal, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Mari Ostendorf. Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information. *NAACL-HLT*, 2018.

Trang Tran, Jiahong Yuan, Yang Liu, and Mari Ostendorf. On the role of style in parsing speech with neural models. *Interspeech*, 2019.

Shubhi Tyagi, Marco Nicolis, Jonas Rohnke, Thomas Drugman, and Jaime Lorenzo-Trueba. Dynamic prosody generation for speech synthesis using linguistics-driven acoustic embedding selection. *Interspeech*, 2020.

Maarten Versteegh, Roland Thiolliere, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. The zero resource speech challenge 2015. *ISCA*, 2015.

Michael Wagner and Duane G Watson. Experimental and theoretical advances in prosody: A review. *Language and cognitive processes*, 25(7-9):905–945, 2010.

Bo Wan, Wenjuan Han, Zilong Zheng, and Tinne Tuytelaars. Unsupervised vision-language grammar induction with shared structure modeling. *ICLR*, 2022.

Liming Wang and Mark Hasegawa-Johnson. A translation framework for visually grounded spoken unit discovery. In *ACSSC*, 2021.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992. URL `https://link.springer.com/content/pdf/10.1007/BF00992696.pdf`.

Daniel H Younger. Recognition and parsing of context-free languages in time n3. *Information and control*, 10(2):189–208, 1967.

Hao Zhang and Daniel Gildea. Efficient multi-pass decoding for synchronous context free grammars. In *Proceedings of ACL-08: HLT*, pp. 209–217, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `https://aclanthology.org/P08-1025`.

Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, and Jiebo Luo. Video-aided unsupervised grammar induction. *NAACL-HLT*, 2021.

Yaodong Zhang. *Unsupervised speech processing with applications to query-by-example spoken term detection*. PhD thesis, Massachusetts Institute of Technology, 2013.

Yaodong Zhang and James R Glass. Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams. *ASRU*, 2009.

Yanpeng Zhao and Ivan Titov. Visually grounded compound pcfgs. *EMNLP*, 2020.

A Appendix

A.1 Baselines

AV-cPCFG: We train compound probabilistic context free grammar (cPCFG) (Kim et al., 2019a) on word-level discrete speech tokens. Similar to AV-NSL, word segments are obtained from VG-HuBERT with segment insertion, and segment representations are extracted from VG-Hubert layer 10 with CLS attention weighted mean-pool. Different from AV-NSL, the segment representations are discretized via kmeans to obtain word-level discrete indices. Because the discretization is word-level instead of phone-level, we swept the number of kmeans cluster over $\{1k, 2k, 4k, 8k, 12k, 16k, 20k\}$, which corresponds to the dictionary size in cPCFG. In summary, AV-cPCFG leverages visual cues only for segmentation and segment representations, but not for phrase structure induction.

DPDP-cPCFG: Instead of training cPCFG on audio-visual word segments and audio-visual segment representations, DPDP-cPCFG does not rely on any visual grounding throughout. Instead, DPDP (Kamper, 2022), a recent speech-only word segmentation algorithm, and vanilla HuBERT representations mean-pooled over DPDP segments are used. We swept through HuBERT layer $\{2, 4, 6, 8, 10, 12\}$. As in AV-cPCFG, kmeans is used for word-level discretization.

oracle AV-NSL: To remove the uncertainty of unsupervised word segmentation, we directly train AV-NSL on top of oracle word segmentation via force alignment. The segment representations are based on learnable attention pooling over vanilla HuBERT layer $\{2, 4, 6, 8, 10, 12\}$ representations. We also tried log Mel spectrograms and HuBERT-L 300M to examine the effectiveness of different input representations. One note is that simpler *score* and *combine* parametrization suffices here[6].

A.2 Hyperparameters

For VG-HuBERT, we run MBR selection on the combination of insertion gap $\{0.1,0.2,0.3\}$ seconds, segmentation layer $\{9,10,11\}$, attention magnitude threshold at top $\{30\%,20\%,10\%\}$, three training random seeds, and model snapshots at training step 20k, 30k, 40k, 50k, 60k. This gives 405 combinations in total.

A.3 Full results table

A.4 Word Segmentation Viz

We show more examples of word segmentation generated by our improved VG-HuBERT in Figure 4. Segments marked with "+" are inserted segments, and vertical blue dotted lines are inferred word boundaries.

A.5 Visualization of Induced Trees

We visualize the induced trees in Figure 5.

[6] We found that for oracle AV-NSL, the original *score* and *combine* parametrization in VG-NSL works better.

Model			Output	SAIoU
Syntax Induction	Segmentation	Seg. Representation (continuous/discrete)	Selection	
Right-Branching	VG-HuBERT+MBR$_{10}$			**0.546**
Right-Branching	DPDP			0.478
AV-NSL	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$ (continuous)	MBR	0.516
AV-NSL	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{11}$ (continuous)	MBR	0.498
AV-NSL	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{12}$ (continuous)	MBR	0.492
AV-NSL	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10,11,12}$ (continuous)	MBR	**0.521**
AV-cPCFG	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$+1k km (discrete)	last ckpt.	0.454
AV-cPCFG	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$+2k km (discrete)	last ckpt.	0.444
AV-cPCFG	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$+4k km (discrete)	last ckpt.	0.499
AV-cPCFG	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$+8k km (discrete)	last ckpt.	0.481
AV-cPCFG	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$+12k km (discrete)	last ckpt.	0.473
AV-cPCFG	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$+16k km (discrete)	last ckpt.	0.471
AV-cPCFG	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$+20k km (discrete)	last ckpt.	0.454
DPDP-cPCFG	DPDP	HuBERT$_2$+1k km (discrete)	last ckpt.	0.434
DPDP-cPCFG	DPDP	HuBERT$_2$+2k km (discrete)	last ckpt.	0.465
DPDP-cPCFG	DPDP	HuBERT$_2$+4k km (discrete)	last ckpt.	0.444
DPDP-cPCFG	DPDP	HuBERT$_2$+8k km (discrete)	last ckpt.	0.387
DPDP-cPCFG	DPDP	HuBERT$_2$+12k km (discrete)	last ckpt.	0.447
DPDP-cPCFG	DPDP	HuBERT$_2$+16k km (discrete)	last ckpt.	0.360
DPDP-cPCFG	DPDP	HuBERT$_{10}$+1k km (discrete)	last ckpt.	0.403
DPDP-cPCFG	DPDP	HuBERT$_{10}$+2k km (discrete)	last ckpt.	0.426
DPDP-cPCFG	DPDP	HuBERT$_{10}$+4k km (discrete)	last ckpt.	0.415
DPDP-cPCFG	DPDP	HuBERT$_{10}$+8k km (discrete)	last ckpt.	0.367
DPDP-cPCFG	DPDP	HuBERT$_{10}$+12k km (discrete)	last ckpt.	0.415
DPDP-cPCFG	DPDP	HuBERT$_{10}$+16k km (discrete)	last ckpt.	0.414

Table 7: Fully-unsupervised phrase structure induction results evaluated with SAIoU.

Model	Segmentation	Seg. Representation	tree target			Output	SAIoU
			train	val	test	Selection	
s-Benepar	VG-HuBERT+MBR$_{10}$	HuBERT$_2$	AV-NSL	AV-NSL	oracle	last ckpt.	**0.538**
s-Benepar	VG-HuBERT+MBR$_{10}$	HuBERT$_4$	AV-NSL	AV-NSL	oracle	last ckpt.	0.536
s-Benepar	VG-HuBERT+MBR$_{10}$	HuBERT$_6$	AV-NSL	AV-NSL	oracle	last ckpt.	**0.538**
s-Benepar	VG-HuBERT+MBR$_{10}$	HuBERT$_8$	AV-NSL	AV-NSL	oracle	last ckpt.	0.532
s-Benepar	VG-HuBERT+MBR$_{10}$	HuBERT$_{10}$	AV-NSL	AV-NSL	oracle	last ckpt.	0.537
s-Benepar	VG-HuBERT+MBR$_{10}$	HuBERT$_{12}$	AV-NSL	AV-NSL	oracle	last ckpt.	0.536
s-Benepar	VG-HuBERT+MBR$_{10}$	HuBERT$_{2,4,6,8,10,12}$	AV-NSL	AV-NSL	oracle	MBR	0.536

Table 8: Self-training results evaluated with SAIoU.

Model			Output	F_1
Syntax Induction	Segmentation	Seg. Representation	Selection	
Random	oracle			32.77
Left-Branching	oracle			24.56
Right-Branching	oracle			**57.39**
VG-NSL		word embeddings	Supervised	53.11
AV-NSL	oracle	log-Mel spectrogram	Supervised	42.01
AV-NSL	oracle	HuBERT$_2$	Supervised	55.51
AV-NSL	oracle	HuBERT-L$_{24}$	Supervised	54.63
AV-NSL	oracle	HuBERT$_2$	MBR	54.99
AV-NSL	oracle	HuBERT$_4$	MBR	53.25
AV-NSL	oracle	HuBERT$_6$	MBR	53.46
AV-NSL	oracle	HuBERT$_8$	MBR	53.14
AV-NSL	oracle	HuBERT$_{10}$	MBR	36.67
AV-NSL	oracle	HuBERT$_{12}$	MBR	48.51
AV-NSL	oracle	HuBERT-L$_{24}$	MBR	54.39
AV-NSL	oracle	HuBERT$_{2,4,6,8,10,12}$	MBR	55.56
AV-NSL	oracle	HuBERT$_{2,4,6,8,10,12,24}$	MBR	**55.96**
AV-NSL → s-Benepar	oracle	HuBERT$_2$	MBR	57.24
AV-NSL → s-Benepar	oracle	HuBERT$_4$	MBR	57.08
AV-NSL → s-Benepar	oracle	HuBERT$_6$	MBR	56.81
AV-NSL → s-Benepar	oracle	HuBERT$_8$	MBR	56.94
AV-NSL → s-Benepar	oracle	HuBERT$_{10}$	MBR	57.16
AV-NSL → s-Benepar	oracle	HuBERT$_{12}$	MBR	**57.33**

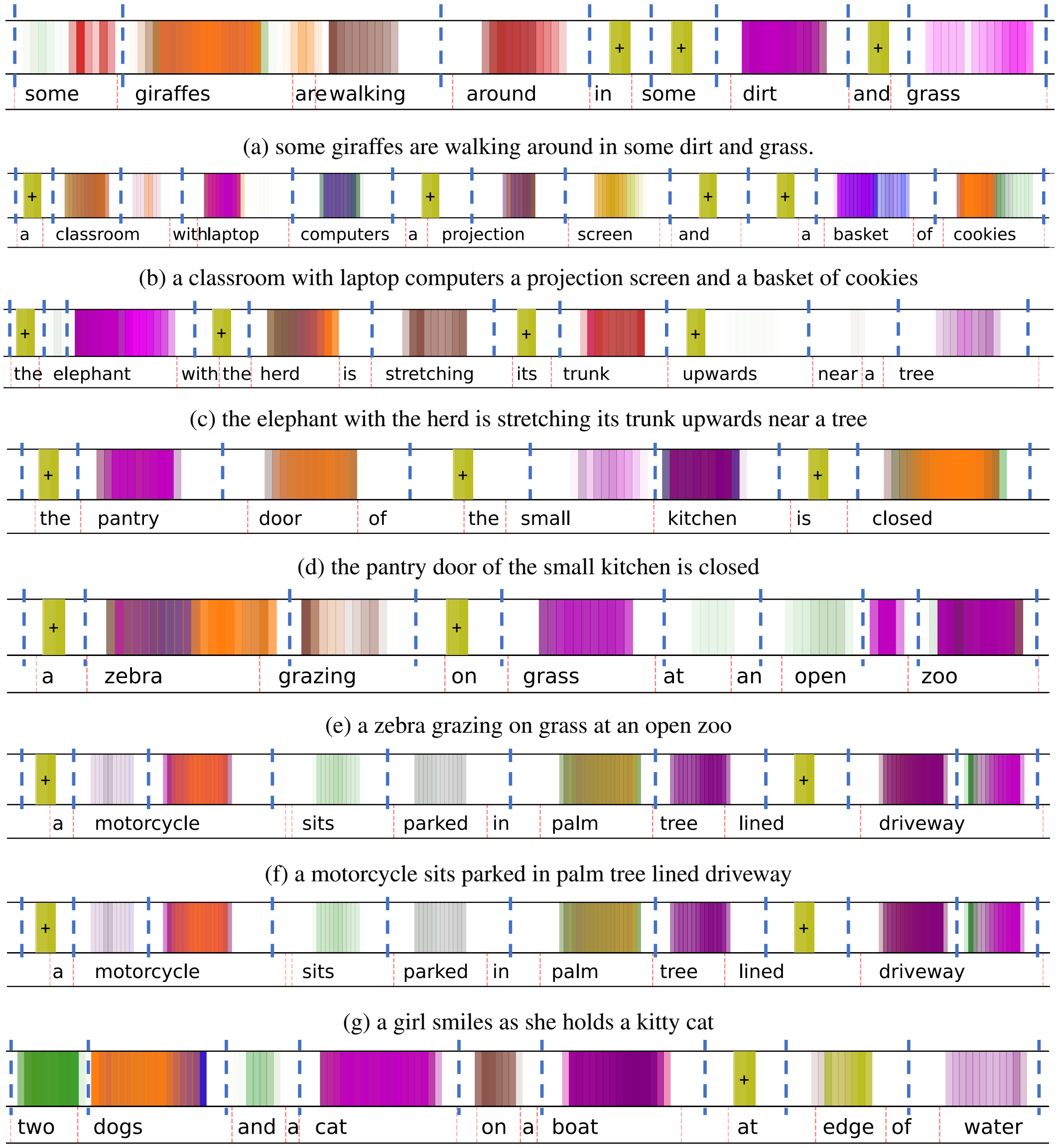
Table 9: Phrase structure induction with oracle segmentation given results evaluated with F_1.

Model	F_1	Constituent Recall			
		NP	VP	PP	ADJP
VG-NSL (Shi et al., 2019)	50.4	**79.6**	26.2	42.0	22.0
VG-NSL + HI	53.3	74.6	32.5	66.5	21.7
VG-NSL + HI + FastText	54.4	78.8	24.4	65.6	22.0
AV-NSL (oracle seg. + HuBERT$_2$)	**55.6**	55.5	**68.1**	**66.6**	22.1
AV-NSL (oracle seg. + HuBERT$_4$)	53.7	57.4	56.8	61.3	21.3
AV-NSL (oracle seg. + HuBERT$_6$)	53.9	59.4	55.4	59.3	21.2
AV-NSL (oracle seg. + HuBERT$_8$)	53.9	56.0	58.0	64.9	**22.5**
AV-NSL (oracle seg. + HuBERT$_{10}$)	50.6	55.8	48.1	57.0	20.5
AV-NSL (oracle seg. + HuBERT$_{12}$)	49.0	62.5	34.4	45.0	17.4

Table 10: Recall of specific typed phrases, and overall F_1 score, evaluated on the SpokenCOCO test split. VG-NSL numbers are taken directly from (Shi et al., 2019). AV-NSL here are trained on oracle segmentation with vanilla HuBERT as the layer representations.

Model			Visual	F_1
Syntax Induction	Segmentation	Seg. Representation	Embedding	
AV-NSL	oracle	HuBERT$_2$	ResNet101	55.51
AV-NSL	uniform	HuBERT$_2$	ResNet101	48.97
AV-NSL	oracle	HuBERT$_{10}$	ResNet101	50.50
AV-NSL	uniform	HuBERT$_{10}$	ResNet101	36.62
AV-NSL	oracle	HuBERT$_2$	DINO	55.71
AV-NSL	oracle	HuBERT$_2$	random	31.23

Table 11: Top rows: Impact of segmentation quality for AV-NSL with number of words segments known in advance. Bottom rows: Impact of visual embedding for AV-NSL



(a) some giraffes are walking around in some dirt and grass.

(b) a classroom with laptop computers a projection screen and a basket of cookies

(c) the elephant with the herd is stretching its trunk upwards near a tree

(d) the pantry door of the small kitchen is closed

(e) a zebra grazing on grass at an open zoo

(f) a motorcycle sits parked in palm tree lined driveway

(g) a girl smiles as she holds a kitty cat

(h) two dogs and a cat on a boat at edge of water

Figure 4: Examples of attention segments generated by VG-HuBERT. Inserted segments are marked with “+”. Vertical blue dotted lines are inferred word boundaries.

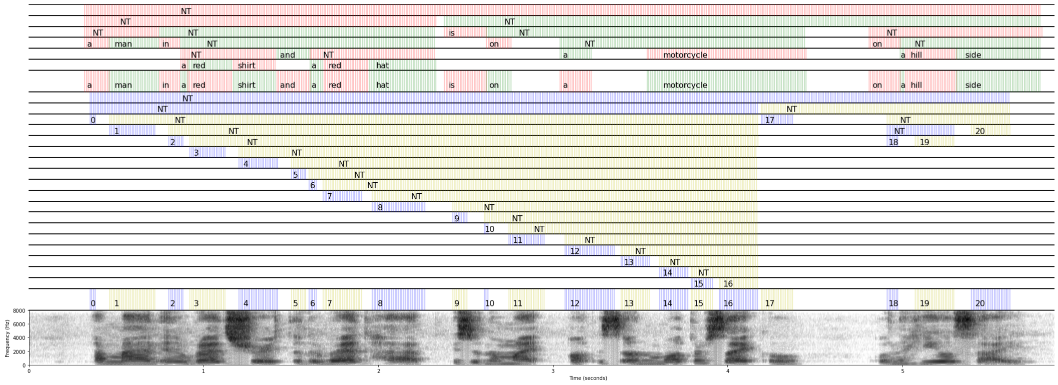


Figure 5: Visualization of an example produced by AV-NSL (best viewed in color). Top (red and green): the ground-truth parse tree; bottom (blue and yellow): the generated parse tree. In each tree, a parent segment adjacently covers its two children segments.