2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) | 979-8-3503-0689-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASRU57964.2023.10389619

AUDIO-VISUAL NEURAL SYNTAX ACQUISITION

Cheng-I Jeff Lai[1], Freda Shi[2*], Puyuan Peng[3*]*,
Yoon Kim[1], Kevin Gimpel[2], Shiyu Chang[4], Yung-Sung Chuang[1], Saurabhchand Bhati[1],
David Cox[5], David Harwath[3], Yang Zhang[5], Karen Livescu[2], James Glass[1]

[1]MIT [2]TTIC [3]UT Austin [4]UC Santa Barbara [5]MIT-IBM Watson AI Lab
`https://github.com/jefflai108/AV-NSL`

ABSTRACT

We study phrase structure induction from visually-grounded speech. The core idea is to first segment the speech waveform into sequences of word segments, and subsequently induce phrase structure using the inferred segment-level continuous representations. We present the Audio-Visual Neural Syntax Learner (AV-NSL) that learns phrase structure by listening to audio and looking at images, without ever being exposed to text. By training on paired images and spoken captions, AV-NSL exhibits the capability to infer meaningful phrase structures that are comparable to those derived by naturally-supervised text parsers, for both English and German. Our findings extend prior work in unsupervised language acquisition from speech and grounded grammar induction, and present one approach to bridge the gap between the two topics.

Index Terms— multi-modal learning, unsupervised learning, grammar induction, speech parsing

1. INTRODUCTION

Multiple levels of early language acquisition happen without supervisory feedback [1]; it is therefore interesting to consider whether automatic learning of language, from identifying lower-level phones or words to inducing high-level linguistic structure like grammar, can also be done in *natural settings*. In these settings, we have access to parallel data from different modalities, while the amount of data is limited. To this end, two concurrent lines of effort have been pursued:

- Zero-resource speech processing, exemplified by the unsupervised discovery of sub-phones, phones, and words [2], involves constructing speech models without relying on textual intermediates, and models how children naturally learn to speak prior to acquiring reading or writing skills.
- Grammar induction is a process that learns latent syntactic structures, such as constituency [3] and dependency trees [4], without relying on annotated structures as supervision.

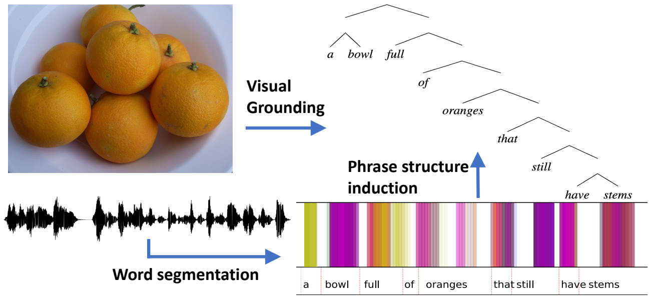

Fig. 1: We study the process of inducing constituency parse trees on unsupervised inferred word segments from raw speech waveforms. No intermediate text tokens or automatic speech recognition (ASR) is needed. For illustration, here we show the gold parse tree from the given text caption.

In recent years, multi-modal learning has emerged as a promising and effective objective in various domains: in speech processing, [5] proposes leveraging parallel image-speech data to acquire associated words [6] and phones [7]; in syntax induction, [8] proposes to induce constituency parses from captioned images. These successes, coupled with insights from developmental psychology [1], motivate us to develop a computational model that utilizes the visual modality to acquire both low-level words and high-level phrase structures directly from speech waveforms, without relying on intermediate text or any form of direct supervision.

In this paper, we present the Audio-Visual Neural Syntax Learner (AV-NSL; Fig. 1), which induces the syntactic structure of visually grounded speech utterances. The speech utterances are represented by sequences of *continuous* speech segment representations, which are derived from a pretrained model that simultaneously discovers word-like units and learns segment representations [9]. AV-NSL (1) learns to map the representations of speech segments and images into a shared embedding space, resulting in higher similarity scores for segments and images that convey similar meanings, (2) estimates the visual *concreteness* of speech segments using the learned embedding space, and (3) outputs speech segments with higher concreteness as the constituents.

To assess the effectiveness of AV-NSL, we compare it with both the ground truth and the grounded text parser VG-NSL [8], as well as several alternative modeling choices such as compound-PCFGs [10] over acoustic units. An ablation study supports the reasonability of our approach. As a by-product, we improve over the previous state of the art in unsupervised word segmentation.

[*] First three authors contributed equally. Correspond to `clai24@mit.edu` and `freda@ttic.edu`

979-8-3503-0689-7/23/$31.00 ©2023 IEEE

2. RELATED WORK

Grounded grammar induction. Since the proposal of the visually grounded grammar induction task [8], there has been subsequent research on the topic [11, 12, 13, *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 [12], where speech features are treated as auxiliary input for video-text grammar induction; that is, [12] still requires text data and an off-the-shelf automatic speech recognition model. In contrast to existing approaches, AV-NSL employs raw speech data and bypasses text to induce constituency parse trees, utilizing distant supervision from parallel audio-visual data.

Spoken word discovery. Following the pioneering work in spoken term discovery [14], a line of work has been done to discover repetitive patterns or keywords from unannotated speech [15, 16, 17, *inter alia*]. Other related work has considered tasks such as unsupervised word segmentation and spoken term discovery [18, 19, 20, 21, *inter alia*], and the ZeroSpeech challenges [22] have been a major driving force in the field. In a new line of work, [6, *inter alia*] show that word-like and phone-like units can be acquired from speech by analyzing audio-visual retrieval models. [9] shows that word discovery naturally emerges from a visually grounded, self-supervised speech model, by analyzing the model's self-attention heads. In contrast, AV-NSL attempts to induce phrase structure, in the form of constituency parsing on top of unsupervised word segments.

Speech parsing and its applications. Early work on speech parsing can be traced back to SParseval [23], a toolkit that evaluates text parsers given potentially errorful ASR output. In the past, syntax has also been studied in the context of speech prosody [24, 25], and [26, 27, 28] incorporate acoustic-prosodic features for text parsing with auxiliary speech input. [29] trains a text parser [30] to detect speech disfluencies, and [31] trains a text dependency parser from speech jointly with an ASR model. There is concurrent work [32] that extends DIORA [33] to unsupervised speech parsing. On the application side, syntactic parses of text have been applied to prosody modeling in end-to-end text-to-speech [34, 35, 36]. While this work builds upon pre-existing text parsing algorithms, we focus on phrase structure induction in the absence of text.

3. METHOD

Given a set of paired spoken captions and images, the Audio-Visual Neural Syntax Learner (AV-NSL) infers phrase structures from speech utterances without relying on text. The basis of AV-NSL is the Visually-Grounded Neural Syntax Learner (VG-NSL) [8, §3.1], which learns to induce constituency parse trees by guiding a sequential sampling process with text-image matching. We break down the problem into two steps: (1) obtaining sequences of word segments, and (2) extracting segment-level self-supervised representations. With these simple extensions to VG-NSL, AV-NSL induces phrase structure without reading text, but rather by listening to speech and looking at images.

3.1. Background: VG-NSL

VG-NSL [8] consists 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*$_\Theta$, (2) selects a pair $\langle w_{i'}^t, w_{i'+1}^t\rangle$ based on the corresponding scores,[1] 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 L_2-normalized vector addition of the input embeddings. The resulting tree is inherently binary and there are $N-1$ combining steps in total, as the parser must combine two nodes in each step.

VG-NSL trains the word embeddings W and a text-image embedding matching module (parameterized with Φ) jointly by minimizing the phrase-level hinge-based triplet loss:

$$\mathcal{L}_{\Phi,W} = \sum_{\mathbf{c}_W,\mathbf{i}_\Phi,\mathbf{c}'_W} [\cos(\mathbf{i}_\Phi,\mathbf{c}'_W) - \cos(\mathbf{i}_\Phi,\mathbf{c}_W) + \delta]_+ + \sum_{\mathbf{c}_W,\mathbf{i}_\Phi,\mathbf{i}'_\Phi} [\cos(\mathbf{i}'_\Phi,\mathbf{c}_W) - \cos(\mathbf{i}_\Phi,\mathbf{c}_W) + \delta]_+,$$

where $\mathbf{c}$, $\mathbf{i}$ are the corresponding vector representations to a pair of parallel text constituent and image; $\mathbf{c}'$ is the representation of an imposter constituent that is not paired with $\mathbf{i}$; $\mathbf{i}'$ is an imposter image representation that is not in parallel with c; δ is a constant margin; $[\cdot]_+ := \max(\cdot, 0)$. By minimizing the above loss function, the embedding space brings semantically similar image and text span representations closer to each other, while pushing apart those that are semantically different. Additionally, the loss function can be 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. Taking the additive inverse of values inside both $[\cdot]_+$ operators, the concreteness of a constituent c is defined as

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

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 [37].

[1] 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.

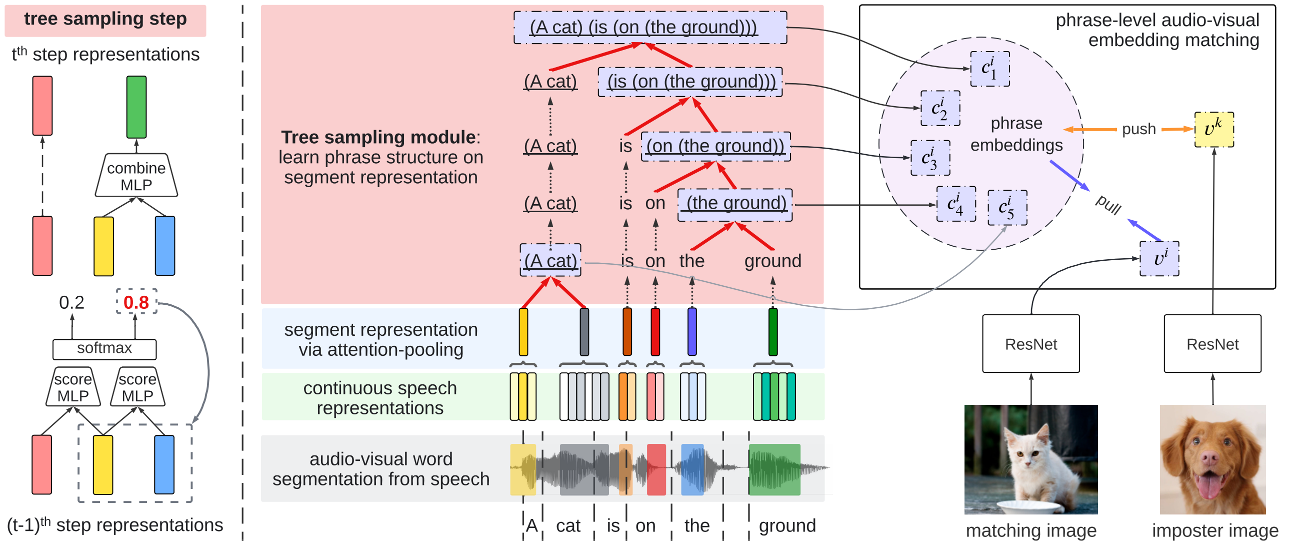
tree sampling step
t^th step representations
combine MLP
0.2
0.8
softmax
score MLP
score MLP
(t-1)^th step representations
Tree sampling module: learn phrase structure on segment representation
(A cat) (is (on (the ground)))
(A cat)
(is (on (the ground)))
(A cat)
is
(on (the ground))
(A cat)
is
on
(the ground)
(A cat)
is
on
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
c_1^i
c_2^i
c_3^i
c_4^i
c_5^i
phrase embeddings
push
pull
v^k
v^i
ResNet
ResNet
matching image
imposter image

Fig. 2: Illustration of AV-NSL, which extends VG-NSL [8] to audio-visual inputs. Taking a pair of speech utterance and its corresponding image as the input, AV-NSL encodes spans of speech utterances and images into a joint embedding space. We train AV-NSL by encouraging it to output more visually concrete spans as constituents. Note that no text is used throughout.

3.2. Audio-Visual Neural Syntax Learner

AV-NSL extends VG-NSL by: (1) incorporating audio-visual word segmentation to obtain sequences of word segments from unannotated speech, (2) jointly optimizing segment-level embeddings and phrase structure induction, and (3) employing deeper parameterization for the *score* and *combine* functions in the parser to handle the noisier speech representations. In AV-NSL, *score* and *combine* are parameterized by GELU-activated [38] multi-layer perceptrons (MLPs). Below we describe (1) and (2) in detail.

Audio-visual word segmentation: For word segmentation, AV-NSL leverages VG-HuBERT [9] (Fig. 2; bottom), a model trained to associate spoken captions with natural images via retrieval. After training, spoken word segmentation emerges via magnitude thresholding of the self-attention heads of the audio encoder: at layer l, we (1) sort the attention weights from the `[CLS]` token to other tokens in descending order, and (2) apply a threshold p to retain the top $p\%$ of the overall attention magnitude (Fig. 3, top).

Empirically, however, the VG-HuBERT word segmenter tends to ignore function words such as *a* and *of*. Therefore, we devise a simple heuristic to pick up function word segments by inserting a short word segment wherever there is a gap of more than s seconds that VG-HuBERT fails to place a segment (Fig. 3). We additionally apply unsupervised voice activity detection [39] to restrict segment insertion to only voiced regions. The length of the insertion gap s, the VG-HuBERT segmentation layer l, attention magnitude threshold $p\%$, and model training snapshots across random seeds and training steps, are all chosen in an unsupervised fashion using minimal Bayes' risk decoding (§3.4).

Speech segment representations: We use the word segments output by VG-HuBERT to calculate the representations. Let $R = \{r_j\}_{j=1}^{T}$ denote the frame-level representation sequence, where T is the speech sequence length. Audio-visual word segmentation returns an alignment $A(i) = r_{p:q}$ that maps the i^{th} word segment to the p^{th} to q^{th} acoustic frames. The segment-level continuous representation for the i^{th} word is $w_i^0 = \sum_{t \in A(i)} a_{i,t} r_{i,t}$, where $a_{i,t}$ is the attention weights over the segments specified by $A(i)$. In AV-NSL, R is the layer representation from a pretrained speech model (e.g., VG-HuBERT), and $a_{i,t}$ is the `[CLS]` token attention weights over frames within each segment.

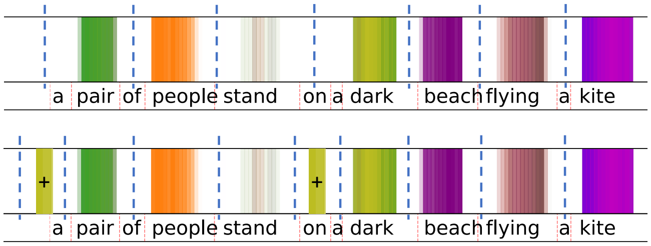

Fig. 3: Example of VG-HuBERT word segmentation (top). Different colors denote different attention heads, and color transparency represents the magnitude of the attention weights. Adjacent attention boundaries (vertical dashed lines) are used as the word boundaries. *Segment insertion* (bottom): short segments (marked with "+") are placed in long enough gaps between existing segments to recover function words. Best viewed in color.

3.3. Self-Training with s-Benepar

[40] has shown that self-training can usually improve parsing performance: the approach involves training an additional parser to fit the output generated by a pre-existing learned parser. Con-

cretely, [40] uses Benepar [30], a supervised neural constituency parser, as the base model for self-training, where it (1) takes a sentence as the input, (2) maps it to word representations, and (3) predicts a score for all spans of being in the constituency parse tree. For inference, the model evaluates all possible tree structures and outputs the highest-scoring one.

Following [40], we apply self-training to improve AV-NSL. We extend Benepar to the speech domain and introduce s-Benepar, which takes segment-level continuous mean-pooling HuBERT representations, instead of words, as the input, and outputs the constituency parse trees.

3.4. Unsupervised Decoding

Another key ingredient of AV-NSL is applying consistency-based decoding [8], which is similar in spirit to minimum Bayes risk (MBR) decoding, for both spoken word segmentation and phrase-structure induction. Given a loss function $\ell_{MBR}(O_1,O_2)$ between two outputs O_1 and O_2, and a set of k outputs $\mathcal{O}=\{O_1,...,O_k\}$, 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 $\mathcal{S}_2$ as $\ell_{MBR}(\mathcal{S}_1,\mathcal{S}_2)=-\text{MIOU}(\mathcal{S}_1,\mathcal{S}_2)$, where $\text{MIOU}(\cdot,\cdot)$ denotes the mean intersection over union ratio across all matched pairs of predicted word spans. We match the predicted word spans using the maximum weight matching algorithm [41], where word spans correspond to vertices, and we define edge weights by the temporal overlap between the corresponding spans.

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

4. EXPERIMENTS

4.1. Setup

Datasets. We first evaluate models on SpokenCOCO [42], the spoken version of MSCOCO [43] where the text captions in English are read verbally by humans. It contains 83k/5k/5k images for training, validation and testing, respectively. Each image has five corresponding captions.

We also extend our experiments to German, where we synthesize German speech from the Multi30K captions [44].[2] It contains 29k/1k/1k images for training, validation and testing, respectively. Each image has one corresponding caption. Following [8], we use pretrained Benepar [30], an off-the-shelf parser, to generate the oracle parse trees for captions.

Preprocessing. For oracle word segmentation, we use the Montreal Forced Aligner [45] trained on the specific language (i.e., English or German). We remove utterances that have mismatches between ASR transcripts and text captions.

4.2. Baselines and Toplines

We consider the following baselines and modeling alternatives to examine each component of AV-NSL:

Trivial tree structures. Following [8], 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 grammars (cPCFG) [10] on word-level discrete speech tokens given by VG-HuBERT. Unlike in AV-NSL, the segment representations are discretized via k-Means to obtain word-level indices; that is, AV-cPCFG leverages visual cues only for segmentation and segment representations, and not for phrase structure induction.

DPDP-cPCFG. In contrast to AV-cPCFG, DPDP-cPCFG does not rely on any visual grounding throughout. We use DPDP [46] and pre-trained HuBERT [47] followed by k-Means to obtain discrete word indices.[3]

Oracle AV-NSL (topline). To remove the uncertainty of unsupervised word segmentation, we directly train AV-NSL on top of oracle word segmentation via forced alignment. Due to the absence of VG-HuBERT, the frame-level representations R are obtained from pre-trained HuBERT while the attention weights $a_{i,t}$ are parameterized by a 1-layer MLP, jointly trained with the tree sampling module instead.

4.3. Evaluation Metrics

Word segmentation. We use the standard word boundary prediction metrics (precision, recall and F_1), which are calculated by comparing the temporal position between inferred word boundaries and forced aligned word boundaries. An inferred boundary located within ±20*ms* of a forced aligned boundary is considered a successful prediction.

Parsing. For parsing with oracle word segmentation, we use PARSEVAL [48] to calculate the F_1 score between the predicted and reference parse trees. For parsing with inferred word segmentation, due to the mismatch in the number of nodes between the predicted and reference parse trees, we use the structured average intersection-over-union ratio (SAIOU [49]) 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, $\mathcal{T}_1=\{a_i\}_{i=1}^n$ and $\mathcal{T}_2=\{b_j\}_{j=1}^m$, where a_i and b_j are time spans. Suppose a_i from $\mathcal{T}_1$ is aligned to b_j from $\mathcal{T}_2$. In a valid alignment, the following conditions must be satisfied: (1) any descendant of a_i may either align to a descendant of b_j or be left unaligned; (2) any ancestor of a_i may either align to an ancestor of b_j or be left unaligned; (3) any descendant of b_j, may either

[2] Synthesized with pre-trained German Tacotron2 from https://github.com/thorstenMueller/Thorsten-Voice.

[3] We sweep the number of word clusters over $\{1\text{k},2\text{k},4\text{k},8\text{k},12\text{k},16\text{k}\}$.

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-cPCFG	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$+4k km (discrete)	last ckpt. (supervised)	0.499
AV-cPCFG	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$+8k km (discrete)	last ckpt. (supervised)	0.481
DPDP-cPCFG	DPDP	HuBERT$_{2}$+2k km (discrete)	last ckpt. (supervised)	0.465
AV-NSL	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10}$ (continuous)	MBR over 10th layer	0.516
AV-NSL	VG-HuBERT+MBR$_{10}$	VG-HuBERT$_{10,11,12}$ (continuous)	MBR over {10th,11th,12th} layer	0.521

Table 1: Fully-unsupervised English phrase structure induction results on SpokenCOCO. Subscripts denote layer number, e.g. HuBERT$_{10}$ denotes the 10th layer representation from HuBERT. We list the best-performing hyperparameters for each modeling choice.

Method	Decoding	Precision	Recall	F_1
DPDP [46]	supervised	17.37	9.00	11.85
VG-HuBERT [9]	supervised	**36.19**	27.22	31.07
VG-HuBERT	supervised	34.34	29.85	31.94
w/ seg. ins. (ours)	MBR	33.31	**34.90**	**34.09**

Table 2: English word segmentation results on the SpokenCOCO validation set. Supervised decoding methods require an annotated development set to choose the best hyperparameters. The best number in each column is in boldface. VG-HuBERT with segment insertion and MBR decoding achieves the best boundary F_1.

Segment Representation	Output Selection	SAIoU
HuBERT	last ckpt.	**0.538**
HuBERT$_{2,4,6,8,10,12}$	MBR	0.536

Table 3: Results of self-training with s-Benepar, trained on outputs from the best AV-NSL model (SAIoU 0.521) from Table 1. Inputs to s-Benepar are segment-level HuBERT representations instead of VG-HuBERT representations.

align to a descendant of a_i or be left unaligned; (4) any ancestor of b_j, may either align to an ancestor of a_i or be left unaligned.

Given a Boolean matrix $\boldsymbol{A}$, where $A_{i,j} = 1$ denotes that a_i aligns to b_j, we compute the structured average IoU between $\mathcal{T}_1$ and $\mathcal{T}_2$ over $\boldsymbol{A}$ by

$$\text{SAIoU}(\mathcal{T}_1,\mathcal{T}_2;\boldsymbol{A}) = \frac{2}{n+m}\left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} A_{i,j}\text{IoU}(a_i,b_j)\right),$$

and the final evaluation result is obtained by maximizing the SAIoU score across all valid alignments. The calculation of the optimal SAIoU score can be done within $\mathcal{O}(n^2m^2)$ time by dynamic programming.

4.4. Unsupervised Word Segmentation

We validate the effectiveness of our unsupervised word segmentation approach. We first compare our improved VG-HuBERT with segment insertion to the original VG-HuBERT [9] and DPDP [46], a speech-only word segmentation method (Table 2). We find that segment insertion improves recall and hurts precision, and achieves the highest F_1 score.

Next, we compare MBR-based and supervised decoding. For efficiency in practice, we implement MBR-based decoding as follows: we first run a pilot hyperparameter selection, performing word segmentation on all candidates in the SpokenCOCO validation set, and subsequently choose the 10 most selected sets of hyperparameters to perform another round of MBR selection on the training set.

For German word segmentation, we employ identical models and settings as those used for English, as [50] has shown that the word segmentation capability of English VG-HuBERT demonstrates cross-lingual generalization without any adaptation. On German Multi30K, our method achieves an F_1 score of 37.46 with MBR, which outperforms that of supervised hyperparameter tuning (36.45).

4.5. Unsupervised Phrase Structure Induction

We quantitatively show that AV-NSL learns meaningful phrase structure given word segments (Table 1). The best performing AV-NSL is based on our improved VG-HuBERT with MBR top 10 selection for word segmentation, VG-HuBERT layers as the segment representations, and another MBR decoding over phrase structure induction hyperparameters, including training checkpoints and segment representation layers. 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.

Next, we compare the performance of AV-NSL with and without self-training (Table 3), and find that self-training with an s-Benepar backbone improves the best AV-NSL performance from 0.521 (Table 1) to 0.538.

Thirdly, Table 4 isolates phrase structure induction from word segmentation quality with oracle AV-NSL. Unlike in Table 1, we can adopt ParsEval F_1 score [48] for evaluation since there is no

Model		Output	F_1
Syntax Induction	Seg. Representation	Selection	
Right-Branching	N/A	N/A	**57.39**
VG-NSL	word embeddings	Supervised	53.11
oracle AV-NSL	HuBERT$_2$	Supervised	55.51
oracle AV-NSL → s-Benepar	HuBERT$_2$	MBR	57.24

Table 4: PARSEVAL F_1 scores given oracle segmentation. The best number is in boldface.

Model		Output	SAIoU
Induction	Segmentation	Selection	
Right-Branching	VG-HuBERT+MBR$_{10}$	N/A	0.456
Left-Branching	VG-HuBERT+MBR$_{10}$	N/A	0.461
AV-NSL	VG-HuBERT+MBR$_{10}$	MBR	**0.487**

Table 5: Phrase structure induction results on the German Multi30K test set. The best number is in boldface.

Model	F_1	Constituent Recall			
		NP	VP	PP	ADJP
VG-NSL [8]	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	**55.5**	55.5	**68.1**	**66.6**	**22.1**

Table 6: Recall of specific typed phrases, incl. noun phrases (NP), verb phrases (VP), prepositional phrases (PP) and adjective phrases (ADJP), and overall F_1 score, evaluated on SpokenCOCO test set. VG-NSL numbers are taken from [8].

Model		Visual	F_1
Word Segmentation	Seg. Repre.		
MFA	HuBERT$_2$	ResNet 101	55.51
Uniform	HuBERT$_2$	ResNet 101	48.97
MFA	HuBERT$_2$	random	31.23
MFA	logMel spec	ResNet 101	42.01

Table 7: PARSEVAL F_1 scores for ablations over word segmentation, visual representation, and speech representation.

mismatch in the number of tree nodes. With proper segment-level representations, unsupervised oracle AV-NSL matches or outperforms text-based VG-NSL. Similarly to Table 3, self-training with s-Benepar on oracle AV-NSL trees further improves the syntax induction results, almost matching that of right-branching trees.

Perhaps surprisingly, right-branching trees (RBT) with oracle and VG-HuBERT word segmentation reach the best English SAIoU and F_1 scores on SpokenCOCO, respectively. We note that the RBTs highly align with the head-initiality of English [51], especially in our setting where all punctuation marks were removed. In contrast, our experiments on German show that AV-NSL out-performs both RBTs and left-branching trees in terms of SAIoU (Table 5).[4]

4.6. Analyses

Unsupervised Constituent Recall: Following [8], we show the recall of specific types of constituents (Table 6). While VG-NSL benefits from the head-initial (HI) bias, where abstract words are encouraged to appear in the beginning of a constituent, AV-NSL outperforms all variations of VG-NSL on all constituent categories except NP.

Ablation Study: We introduce three ablations to evaluate the efficacy of high-quality word segmentation, visual representation, and speech representation (Table 7). Concretely, we train AV-NSL with the following modifications:

1. Given the number of words n, we divide the speech utterances uniformly into n chunks to get the word segmentation, and use the same visual representations as AV-NSL.
2. We replace visual representations with random vectors, where each pixel is independently sampled from a uniform distribution, and use the oracle word segmentation.
3. We replace the self-supervised speech representations (HuBERT) with log-Mel spectrograms.

We observe significant performance drops in all settings, compared to oracle AV-NSL. This set of results complements Table 1, stressing that precise word segmentation and both high-quality visual and speech representations are all necessary for phrase structure induction from speech.

5. CONCLUSION AND DISCUSSION

Previous research has achieved notable progress in zero-resource speech processing and grammar induction by employing multimodal techniques. In our study, we propose an approach to model human language acquisition that leverages the visual modality to acquire language competence. Our approach, AV-NSL, encompasses the extraction of word-level representations from speech and the derivation of syntactic structures from those representations, thereby eliminating the reliance on text. Through quantitative and qualitative analyses, we demonstrate on both English and German that our proposed model successfully infers meaningful constituency parse trees based on continuous word segment representations. Our work represents the initial step in grammar induction within textless settings, paving the way for future research endeavors, which include but are not limited to (1) building end-to-end models that take spoken utterances and produce their syntactic analysis, (2) understanding the contribution of various grounding signals to grammar induction, and (3) modeling human language acquisition in grounded environments.

[4] For German grammar induction with oracle segmentation, oracle AV-NSL attains 33.94 F_1 while LBT/RBT attain 26.70/25.30 F_1 respectively.

6. REFERENCES

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

[2] 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," in *ICASSP*, 2013.

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

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

[5] David Frank Harwath, *Learning spoken language through vision*, Ph.D. thesis, Massachusetts Institute of Technology, 2018.

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

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

[8] Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu, "Visually grounded neural syntax acquisition," in *ACL*, 2019.

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

[10] Yoon Kim, Chris Dyer, and Alexander M Rush, "Compound probabilistic context-free grammars for grammar induction," in *ACL*, 2019.

[11] Yanpeng Zhao and Ivan Titov, "Visually grounded compound PCFGs," in *EMNLP*, 2020.

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

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

[14] Alex S Park and James R Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2007.

[15] Aren Jansen and Benjamin Van Durme, "Efficient spoken term discovery using randomized algorithms," in *ASRU*, 2011.

[16] Fergus McInnes and Sharon Goldwater, "Unsupervised extraction of recurring words from infant-directed speech," in *CogSci*, 2011.

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

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

[19] Herman Kamper, Aren Jansen, and Sharon Goldwater, "A segmental framework for fully-unsupervised large-vocabulary speech recognition," *Computer Speech & Language*, 2017.

[20] Jan Chorowski et al., "Aligned contrastive predictive coding," in *Interspeech*, 2021.

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

[22] Ewan Dunbar, Nicolas Hamilakis, and Emmanuel Dupoux, "Self-supervised language learning from raw audio: Lessons from the zero resource speech challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1211–1226, 2022.

[23] 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," in *LREC*, 2006.

[24] Michael Wagner and Duane G Watson, "Experimental and theoretical advances in prosody: A review," *Language and cognitive processes*, vol. 25, no. 7-9, pp. 905–945, 2010.

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

[26] 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," in *NAACL-HLT*, 2018.

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

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

[29] Paria Jamshid Lou, Yufei Wang, and Mark Johnson, “Neural constituency parsing of speech transcripts,” in *NAACL-HLT*, 2019.

[30] Nikita Kitaev and Dan Klein, “Constituency parsing with a self-attentive encoder,” in *ACL*, 2018.

[31] Adrien Pupier, Maximin Coavoux, Benjamin Lecouteux, and Jérôme Goulian, “End-to-end dependency parsing of spoken french,” in *Interspeech*, 2022.

[32] Yuan Tseng, Cheng-I Jeff Lai, and Hung-yi Lee, “Cascading and direct approaches to unsupervised constituency parsing on spoken sentences,” in *ICASSP*, 2023.

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

[34] Haohan Guo, Frank K Soong, Lei He, and Lei Xie, “Exploiting syntactic features in a parsed tree to improve end-to-end TTS,” in *Interspeech*, 2019.

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

[36] Nobuyoshi Kaiki, Sakriani Sakti, and Satoshi Nakamura, “Using local phrase dependency structure information in neural sequence-to-sequence speech synthesis,” in *O-COCOSDA*. IEEE, 2021.

[37] Ronald J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine Learning*, 1992.

[38] Dan Hendrycks and Kevin Gimpel, “Gaussian error linear units (GELUs),” *arXiv preprint arXiv:1606.08415*, 2016.

[39] Zheng-Hua Tan, Najim Dehak, et al., “rvad: An unsupervised segment-based robust voice activity detection method,” *Computer speech & language*, 2020.

[40] Haoyue Shi, Karen Livescu, and Kevin Gimpel, “On the role of supervision in unsupervised constituency parsing,” in *EMNLP*, 2020.

[41] Zvi Galil, “Efficient algorithms for finding maximum matching in graphs,” *ACM Comput. Surv.*, vol. 18, no. 1, pp. 23–38, mar 1986.

[42] Wei-Ning Hsu, David Harwath, Tyler Miller, Christopher Song, and James Glass, “Text-free image-to-speech synthesis using learned segmental units,” in *ACL*, 2021.

[43] 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 *ECCV*. Springer, 2014.

[44] Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia, “Multi30k: Multilingual english-german image descriptions,” in *Proceedings of the 5th Workshop on Vision and Language*, 2016.

[45] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using Kaldi,” in *Interspeech*, 2017.

[46] Herman Kamper, “Word segmentation on discovered phone units with dynamic programming and self-supervised scoring,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 684–694, 2022.

[47] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[48] E. Black et al., “A procedure for quantitatively comparing the syntactic coverage of English grammars,” in *Speech and Natural Language: Proceedings of a Workshop*, 1991.

[49] Freda Shi, Kevin Gimpel, and Karen Livescu, “Structured tree alignment for evaluation of constituency parsing,” Unpublished manuscript, 2023.

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

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