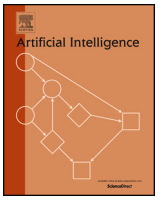
Artificial Intelligence 341 (2025) 104306

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



Grammar induction from visual, speech and text

Yu Zhao [a], Hao Fei [b,*], Shengqiong Wu [b], Meishan Zhang [a], Min Zhang [a], Tat-seng Chua [b]

[a] Harbin Institute of Technology (Shenzhen), Shenzhen, 518055, China
[b] National University of Singapore, Singapore, 118404, Singapore

ARTICLE INFO

Keywords:
Grammar induction
Multimodal learning
Structure modeling

ABSTRACT

Grammar Induction (GI) seeks to uncover the underlying grammatical rules and linguistic patterns of a language, positioning it as a pivotal research topic within Artificial Intelligence (AI). Although extensive research in GI has predominantly focused on text or other singular modalities, we reveal that GI could significantly benefit from rich heterogeneous signals, such as text, vision, and acoustics. In the process, features from distinct modalities essentially serve complementary roles to each other. With such intuition, this work introduces a novel *unsupervised visual-audio-text grammar induction* task (named **VAT-GI**), to induce the constituent grammar trees from parallel images, text, and speech inputs. Inspired by the fact that language grammar natively exists beyond the texts, we argue that the text has not to be the predominant modality in grammar induction. Thus we further introduce a *textless* setting of VAT-GI, wherein the task solely relies on visual and auditory inputs. To approach the task, we propose a visual-audio-text inside-outside recursive autoencoder (**VaTiora**) framework, which leverages rich modal-specific and complementary features for effective grammar parsing. Besides, a more challenging benchmark data is constructed to assess the generalization ability of VAT-GI system. Experiments on two benchmark datasets demonstrate that our proposed VaTiora system is more effective in incorporating the various multimodal signals, and also presents new state-of-the-art performance of VAT-GI. Further in-depth analyses are shown to gain a deep understanding of the VAT-GI task and how our VaTiora system advances. Our code and data: https://github.com/LLLogen/VAT-GI/.

1. Introduction

Within the field of AI, human language acquisition is one of the important research topics. Human language knowledge involves various aspects, such as the vocabulary, phonetics, morphology, syntax, semantics and pragmatics of languages [5,40,2,13,9–12]. Among these, inferring the underlying grammatical rules and linguistic patterns of a language plays a significant role in language learning. The process is also tasked as grammar induction [28,60,49,17], aiming to uncover the latent structure of language constituents from natural inputs in an unsupervised manner. In the community, the mainstream explorations on GI pay an extensive focus on the textual signals [8,27,53,4,21,26,63,50], as texts are the major medium of language. However, it is important to note that human always acquire language knowledge with multimodal signals, i.e., text, vision and acoustic. Knowing this fact, researchers

* Corresponding author.
E-mail addresses: zhaoyu9067@lve.cn (Y. Zhao), haofei37@nus.edu.sg (H. Fei), swu@u.nus.edu (S. Wu), zhangmeishan@hit.edu.cn (M. Zhang), zhangmin2021@hit.edu.cn (M. Zhang), dcscts@nus.edu.sg (T.-s. Chua).

https://doi.org/10.1016/j.artint.2025.104306
Received 15 October 2024; Received in revised form 26 January 2025; Accepted 8 February 2025

Available online 12 February 2025

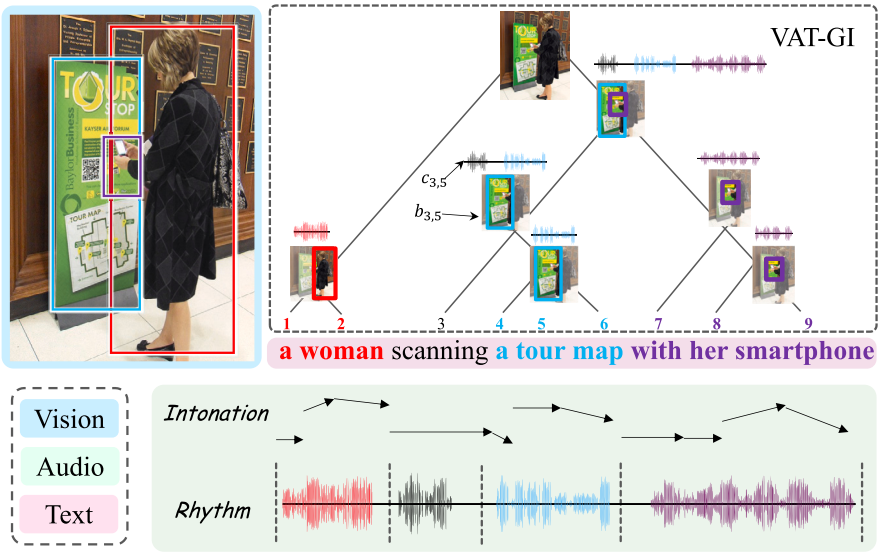


Fig. 1. Unsupervised grammar induction with vision, audio and text modality sources, each of which contributes complementarily to the task.

carry out GI using different modalities of information, such as vision-language grammar learning [63,19,58,51,52], speech-based textless GI [34] and video-aided GI [62,61,42,45].

Unfortunately, most existing researches on GI tend to overlook the mutually complementary contributions of the information in broad-covering modalities. Actually, information and signals from all different modalities together can play crucial complementary roles in the acquisition of language throughout phylogenetic development. This can be exemplified by the cases in Fig. 1: the pixel regions in a visual image are often associated with the noun phrase structure, such as '*a woman*' and '*a tour map*', while speech intonation and rhythm help segment sentences from a constituent-level perspective, such as '*scanning a tour map*' and '*with her smartphone*'. This thus motivates us to introduce a novel task of unsupervised *visual-audio-text grammar induction* (VAT-GI), with the aim of exploring the language learning of NLP-related AI systems in realistic scenarios. Particularly, VAT-GI operates over parallel images, audios and texts, and extracts the shared latent constituency structures that encompass all these modalities (see § 3 for task definition). As depicted in Fig. 1, the terminal nodes of the tree are the individual words, while non-terminal nodes are notional constituent spans, where nodes also associate with visual regions and speech clips. Different sources of modalities together form an integral constituency grammar structure.

To effectively facilitate the multimodal grammar induction, a VAT-GI system should take into full consideration the distinct yet complementary characteristics of each modality:

- **Text:** Language intrinsically encompasses structural information by virtue of its inherent compositionality [5,38,48], thereby providing a straightforward basis for grammar induction.
- **Vision:** The smaller spatial regions of pixels combine to form larger regions with higher-level visual semantics [19,63,58], which correspond to the hierarchical structure observed among textual constituent spans.
- **Audio:** Speech conveys structural information through various intonation and rhythm patterns, naturally depicting the discontinuity and continuity of phonemes [44,54,3,39,18].

With the aforementioned observations, in this work we present a nichetargeting framework to approach VAT-GI. Built upon the deep inside-outside recursive autoencoder (Diora) model [8] for effective text-based GI, we devise a novel *visual-audio-text Diora* (namely **VaTiora**), which extends the Diora with the further capability of capturing specific structural characteristics of three modalities of sources, such as the compositionality of texts, the hierarchical spatial regions of images and the rhythm patterns of speech, and integrating them properly. As shown in Fig. 3, first, rich modal-specific and complementary features, such as text embedding, visual regions and voice pitch, etc., are constructed from the input text, image and audio sources, respectively, all of which are effectively integrated into the inside-outside recursive autoencoding. Then, in the feature interaction phase, text embeddings are first fused with audio features via cross-attention to obtain the token-wise text-audio representations for constructing the span vectors; these span vectors are further mapped with fine-grained visual region features by an attention-aware fusion, thereby enhancing the span representations. Moreover, the composition probabilities of visual regions are considered when calculating the span scores, during which the intonation pitch frequency features and voice activity features (representing the rhythm pattern) are also incorporated. Through such cohesive integration, the cross-modal signals mutually reinforce each other for more accurate constituency parsing in the inside-outside pass of VaTiora. We follow [8] to train the VaTiora with structure reconstruction and contrastive learning. In addition, we further introduce a cosine similarity learning objective across three types of features for representation alignment, bridging the heterogeneity among modalities.

Additionally, we introduce a *textless* setting for VAT-GI, wherein the task solely relies on visual and auditory inputs, and the constituent tree is no longer structured around individual words but instead the segmented speech clips. The motivation for setting the textless VAT-GI comes from the consideration of the languages that lack written forms, e.g., the languages of minority ethnic groups in certain regions. We contend that text is not a fundamental modality for VAT-GI, as supported by the physical fact that *language grammar natively exists beyond the texts*. Through investigating the feasibility of textless VAT-GI, we uncover the potential of modeling syntactic structures using non-textual modalities. Accordingly, we specially propose the aligned span-clip F1 metric (**SCF1**) to measure the textless VAT-GI, due to the inaccessibility of golden word segments of speeches in this setting.

VAT-GI can be evaluated on the public SpokenCOCO dataset [20] with well-aligned images, texts and audio speeches. However, SpokenCOCO has inherent limitations in terms of monotonous styles of visual scenes and speech tones, as well as short captions, which largely weakens the model generalization ability to complex signal inputs. To combat that, we present a more challenging test set for VAT-GI, namely **SpokenStory**, which advances in richer visual scenes, timbre-rich audio and deeper constituent structures. We collect 1,000 aligned image-text from Localized Narratives [47], and further record speech for each sentence by human speakers. The experimental results on the datasets suggest that leveraging multiple modality features helps better grammar induction. Further analyses demonstrate that VaTiora is effective in incorporating the various multimodal signals, and yielding new state-of-the-art induction performance of VAT-GI task.

In summary, this paper contributes in five key aspects.

- We for the first time introduce an important yet challenging task, unsupervised visual-audio-text grammar induction, to better emulate the human-level phylogenetic language acquisition;
- We present a novel VAT-GI model, VaTiora, to properly navigate multimodal input sources;
- We newly contribute a challenging benchmark data, SpokenStory, for VAT-GI; and also propose a new metric for the textless VAT-GI evaluation;
- Our proposed method sets a strong-performing benchmark result for follow-up research. Our resources will be open to facilitate the community.

The rest of the paper is organized as follows. The next section contains the background and related work of textual grammar induction and multi-modal grammar induction. After that, in section 3 we present the definition visual-audio-text grammar induction task and how to evaluate it. In section 4, we elaborate on the details of the proposed VaTiora framework, which incorporates multiple cues from three modalities into an inside-outside algorithm. Following that, in the experiment section, we compare our methods with strong grammar induction baselines, and explore how our method advances via in-depth quantitative and qualitative analysis. Finally, we shed the light on the future work of this research, and then conclude the paper.

2. Related work

In the field of AI, research draws inspiration from language to enhance the development of more intelligent systems and models. Among all topics, the task of grammar induction aims to infer the underlying grammatical rules and linguistic patterns of a language, thereby gaining a deeper understanding of the principles governing human intelligence. GI unsupervisedly determines the constituent syntax of language, in the format of a phrase-structure tree, which intuitively depicts the language compositionality, i.e., how the small components (e.g., words or phrases) combine to larger constituents (e.g., clauses or sentences). Such a process essentially emulates the human language acquisition. Thus, GI [36] has long been a fundamental topic in natural language processing (NLP) [7,53,31,21,26,63,19,58,51].

According to the categories of information modalities, all existing GI research can be categorized into text-based GI, vision-language-based GI, and speech-based GI. Below, we survey relevant studies within each of these groups.

Textual Grammar Induction. The majority of previous GI studies pay the focus on the language domain with textual corpora [7,36,53,31,21,26]. Generally, there are two types of frameworks for text-only grammar induction. The first is PCFG-based methods, which assume the context-free grammar rules exist with a probability and estimates the likelihood of each latent rule [27,32,59]. One way of inducing this probabilistic grammar is to fix the structure of a grammar and then find the set of optimal parameters such that the resulting grammar best explains the language, which is usually approximated by a training corpus. Early works induce PCFG via statistic-based methods, such as EM algorithm [36]. Following efforts focus on improving it by carefully-crafted auxiliary objectives [31], priors or non-parametric models [59], and manually-engineered features [15]. Kim et al. [27] first propose to parameterize the PCFG's rule probability with a trainable neural network. Jin et al. [24] applies depth bounds within a chart-based Bayesian PCFG inducer, limiting the search space of an unsupervised PCFG inducer. The other type of framework is autoencoder-based methods, which estimate span composition likelihood without assuming context-free grammar. Drozdov et al. [8] first propose the Diora model, which incorporates the inside-outside algorithm into a latent tree chart parser.

Visual-Language Grammar Induction. The visually-grounded GI has received increasing research attention [63,19,58,51]. Since visual regions (e.g., objects of interest) intuitively encompass the correspondence of the textual spans (e.g., noun phrases), the visions are imported as a type of enhanced signal for better constituency tree parsing. Shi et al. [51] first propose to import visual features into grammar induction, mapping visual and text via visual-semantic embedding. The following works extend text-only GI models with extra visual information, such as VC-PCFG [63], VG-NSL [51] and Cliora [58]. The visually grounded compound PCFGs (VC-PCFG) extends the compound PCFG model (C-PCFG) by including a matching model between images and text. Similar performance gain

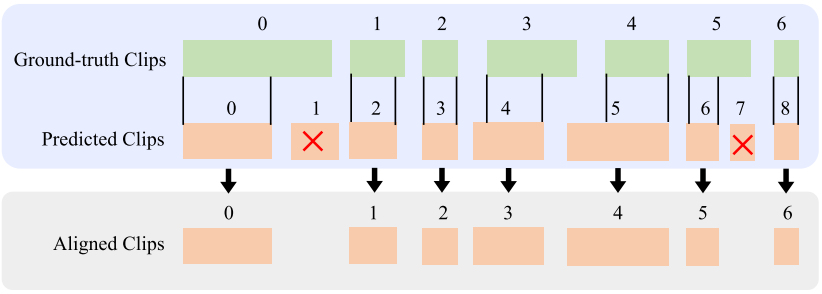


Fig. 2. Illustration of clip alignment in SCF1.

is also observed with VG-NSL. Instead of fusing information of the entire image into the phrase representations, Cliora uses region-based fine-grained alignment to have a thorough understanding of the image. There are also video-aided GI, where, compared with a single image, video frames can further facilitate the understanding of motion dynamics for the correspondence of textual predicates [62,61,42,45]. Zhang et al. [62,61] first propose video-aided GI models that prove videos can provide even richer information than images for grammar induction, including not only static objects but also actions and state changes useful for inducing verb phrases.

Speech-oriented Grammar Induction. On the other hand, acquiring languages from speech has also gained consistent research interests [30,22,23,16,17], where the acoustic-prosodic features (e.g., phonetics, phonology) can offer important clues for the syntax induction from different perspectives than the visions and texts [3,39,18]. Lai et al. [35] first attempt incorporate audio signals for grammar induction and present the novel AV-NSL learner. They segment the speech waveform into sequences of word segments, and subsequently induce phrase structure using the inferred segment-level continuous representations. In this work, we take a combined holistic viewpoint, and investigate the GI under multimodal information sources, i.e., introducing a novel visual-audio-text grammar induction task. This can be quite intuitive, as we humans always perceive the world with varied sensory inputs that can partially share the same common structures and meanwhile preserve distinctly complementary features, which together help achieve more effective GI.

Despite the significant achievements and extensive attention garnered by prior GI research, current approaches remain confined to single modalities, such as text or others, often overlooking the mutually complementary contributions of information from diverse modalities. In reality, information and signals from various sources—including text, vision, and acoustics—collectively and complementarily play crucial roles in language acquisition throughout phylogenetic development. Leveraging rich heterogeneous signals could greatly benefit GI. To address this, we propose a novel task, unsupervised visual-audio-text grammar induction (VAT-GI), to emulate real human-level phylogenetic language acquisition.

3. Task definition

Formulation. Given a sentence $X = \{x_1, x_2, ..., x_n\}$ with n words, an associated image I and a piece of speech S for X, the goal of unsupervised VAT-GI is to induce a constituency tree structure T from X, I and S without any supervision of constituent structure annotations for training. As shown in Fig. 1, the output tree structure T is formed in a phrase constituent tree by Chomsky Normal Form (CNF) [6], where each non-terminal node in the tree has exactly two children. Each node in T contains a text span $(i, j), 1 \le i \le j \le n$, a region box $b_{i,j} \in \mathbb{R}^4$ in I and a speech span $c_{i,j} \in \mathbb{R}^2$, where the text span (i, j) is grounded to the region box $b_{i,j}$ and aligned to the speech span $c_{i,j}$. For the textless setting, the inputs of VAT-GI are only I and S. Thus the node set of output T contains only $c_{i,j}$ and $b_{i,j}$.

Evaluation. We adopt two widely-used metrics, *averaged corpus-level F1 (Corpus-F1)* and *averaged sentence-level F1 (Sent-F1)*, following [27]. Corpus-level F1 calculates precision/recall at the corpus level to obtain F1, while sentence-level F1 calculates F1 for each sentence and averages across the corpus. In this paper, we follow [27] discard trivial spans (length < 2) and evaluate on sentence-level F1 per recent work. In textless settings, where there is no text script for speech-to-text alignment, we introduce a new metric, *aligned span-clip F1 (SCF1)*, to measure the quality of the textless constituency tree. Suppose the output of the textless VAT-GI contains a sequence of speech clips $\{clip_m^p\}$. The predicted constituency tree is denoted as the pairs of speech clips $\{c^p\} = \{(clip_h^p, clip_t^p)\}$, where each clip represents a token. SCF1 first aligns predicted speech clips to ground-truth speech clips via tIoU (Temporal Intersection over Union) [41], which represents the overlap between two time intervals. It is a metric to measure the overlap ratio on the temporal dimension. We use a greedy mapping from start to end, aligning predicted clip and ground-truth segmented clip if their tIoU is over a predefined threshold p. In details, we traverse all the ground-truth clips, denoting $\{clip_n^g\}$, to search the aligned clips. For each $clip_n^g$, we calculate the tIoU between $clip_n^g$ with all the predicted clips $\{clip_m^p\}$. A predicted clip $clip_n^g$ is considered as the aligned one with $clip_m^p$ only when $\text{tIoU}(clip_n^g, clip_m^p) > p$. If a ground-truth clip has multiple mapped predictions, we keep the one with the largest tIoU. After alignment, we then calculate the F1 score based on TP, FP and FN. We say the two spans are "*matched*" when both their head and tail clips are aligned respectively. Let $\{c^g\} = \{(clip_h^g, clip_t^g)\}$ be the ground-truth spans, then we have:

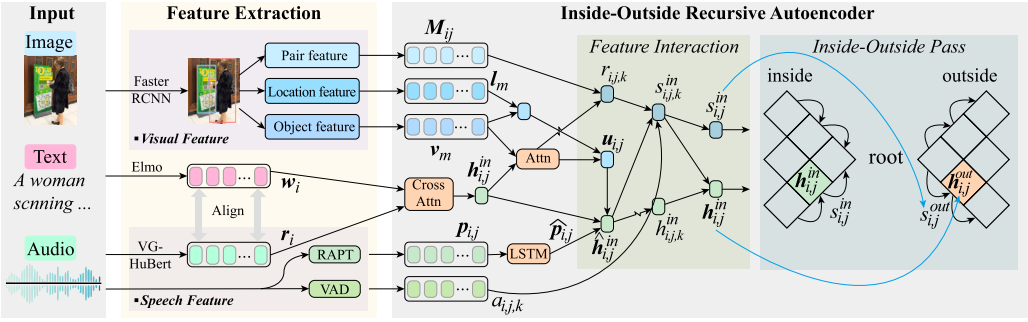


Fig. 3. In our VaTiora framework, **first** the feature extraction module constructs rich modal-specific features from the input image, text and speech. RAPT: robust algorithm for pitch tracking; VAD: voice activity detection. **Then** the inside-outside recursive autoencoder fuses various features and performs grammar induction.

Table 1
Summary of all the features used in VaTiora.

Text feature	$\boldsymbol{w}_i$	Embedding of word i
Object feature	$\boldsymbol{v}_m$	RoI feature of an object m
Location feature	$\boldsymbol{l}_m$	Embedding of bounding box coordinates of object m
Pair feature	$\boldsymbol{M}_{mn}$	Frequency of objecet m and n occuring in one image
Clip feature	$\boldsymbol{r}_i$	Representation of the i-th clip
Pitch feature	$\boldsymbol{p}_{i,j}$	Embedding of average $f0$ values of the i-th clip
Voice activity feature	at_i/nt_i	Voice activity time/non-voice time of the i-th clip

$$\begin{aligned}
&\text{TP} = \text{Count}(pred \text{ and } gt \text{ are matched}, pred \in \{c^p\} \text{ and } gt \in \{c^g\}),\\
&\text{FP} = \text{Count}(pred \in \{c^p\} \text{ that no } gt \in \{c^g\} \text{ could match}),\\
&\text{FN} = \text{Count}(gt \in \{c^g\} \text{ that no } pred \in \{c^p\} \text{ could match}),
\end{aligned} \tag{1}$$

where p is a threshold and Count(f) denotes the number of samples that satisfies the condition f. Fig. 2 gives the illustration of this mapping process.

4. Framework of VaTiora

We propose a novel **v**isual-**a**udio-**t**ext **i**nside-**o**utside **r**ecursive **a**uto-encoder (dubbed VaTiora) framework based on the Diora model [8]. As shown in Fig. 3, VaTiora performs grammar induction with two key modules: the multimodal feature extraction and the inside-outside recursive autoencoding.

4.1. Feature extraction

In the first step, we extract the modal-preserving features of each input modal from various perspectives, which are summarized in Table 1.

Textual Features. For text inputs, we follow the conventional GI methods, taking pre-trained word embeddings ELMo [43] to obtain the textual representations $\boldsymbol{W} = \{\boldsymbol{w}_1, ..., \boldsymbol{w}_n\}$ from the input X. In practice, we can also adopt other pre-trained word embeddings, such as Glove, or use randomly initialized embedding.

Visual Features. For visual feature extraction, given the input image I, we follow [58] to adopt an external object detector (Faster-RCNN [14]) to extract a sequence of the object $\{o_1, ..., o_M\}$ along with their object RoI features $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_M\}$. Besides, each object o_m has a bounding box b_m, which is encoded by an embedding layer to obtain the location features $\boldsymbol{l}_m$. Denote b_m as:

$$b_m = (x_{min}, y_{min}, x_{max}, y_{max}). \tag{2}$$

With the top-left as the origin point, we normalize it to:

$$b'_m = (x_{min}/w, y_{min}/h, x_{max}/w, y_{max}/h), \tag{3}$$

where w, h is the width and length of the image. Then we embed b'_m by:

$$\boldsymbol{l}_m = \text{FC}(b_m). \tag{4}$$

We further consider the pair features of visual regions to determine the likelihood of object pairs forming a larger region. Concretely, we maintain a pair relevance matrix $\boldsymbol{M} \in \mathbb{R}^{C\times C}$ for all categories of object pairs in the dataset based on the co-occurrence probability, where C is the total number of object categories. Supposing there are C categories of object label, and for every two object that belong to $c_\alpha, c_\beta \in C$, their score is:

$$\boldsymbol{M}_{\alpha,\beta} = \frac{\text{Co-Count}(\alpha,\beta)}{\sum_{\gamma\in C}(\text{Co-Count}(\alpha,\gamma)+\text{Co-Count}(\gamma,\beta))}, \tag{5}$$

where Co-Count(γ,β) means the count that objects with categories α and β detected in the same image within the entire dataset.

Speech Features. For speech features, we leverage a word discovery model VG-HuBert [41] to segment raw speech into clips $\{c_1,...,c_T\}$, and obtain the clip representations $\{r_1,...,r_T\}$. In the full setting, we take the textual transcripts to force align the speech, while in the textless setting, we use VG-HuBert to perform unsupervised word discovery. Note that each clip corresponds to a single word in the sentence when text is provided as input, i.e., $T = n$. To take into account the structural information of speech signals, we leverage pitch detection and voice activity detection (VAD) to extract intonation and rhythm features. First, we adopt the robust algorithm for pitch tracking (RAPT) [55], extracting the $f0$ value for each speech clip c_i, and encode it to a high dimension representation. Moreover, considering the pitch frequency constraint of normal human speech, we limit the extracted $f0$ value to the range of 50 Hz to 500 Hz [57]. Then, we embed the rounded $f0$ (450 values) to the high-dimension representation, and use an LSTM model to capture the temporal features of pitch changing:

$$\begin{aligned} f0_i &= \text{Round}(\text{Avg}(\text{RAPT}(c_i))),\\ \boldsymbol{p}_i &= \text{Embed}(f0_i), \quad \hat{\boldsymbol{p}}_i = \text{LSTM}(\boldsymbol{p}_i), \end{aligned} \tag{6}$$

where Avg denotes average operation, Round means rounding to an integer. $f0_i$ means the average $f0$ of frames in clip c_i. $\hat{\boldsymbol{p}}_i$ represents the pitch feature and will be used to enhance span representation. At last, we represent the rhythm feature of speech. The voice activity time can be extracted by the robust voice activity detection (rVAD) method [56]. For each clip, VAD outputs the number of speech frames and the non-speech frames. We sum the frame length of speech frames and non-speech frames as voice activity time at_i and non-voice time wt_i for each clip c_i.

4.2. Inside-outside recursive autoencoder

Similar to Diora, VaTiora operates as an autoencoder, encoding the sentence into span representations and then reconstructing the bottom words of the constituent tree, i.e., the terminal nodes. The encoding process involves mapping the input words, denoted as X, to a latent constituent tree. To efficiently explore all valid trees, a dynamic programming-based inside-outside algorithm [36] is employed. Notably, VaTiora incorporates external multi-modal cues during the inside-outside pass, thereby extending the functionality of the Diora framework. Fig. 3 provides an overview of the interaction of features and the inside-outside pass.

Feature Interaction. We first initialize the bottom-most terminal nodes[1] $\boldsymbol{h}^{in}_{i,i}$ with the aligned word embedding $\boldsymbol{W} = \{\boldsymbol{w}_i\}$ and speech clip representation $\boldsymbol{R} = \{\boldsymbol{r}_i\}$ via cross-attention:

$$\begin{aligned} \{\boldsymbol{h}^{in}_{i,i}\} &= \text{CrossAttn}(\boldsymbol{W},\boldsymbol{R}),\\ \text{CrossAttn}(\boldsymbol{W},\boldsymbol{R}) &= \text{Softmax}((\boldsymbol{Q}\boldsymbol{W})(\boldsymbol{K}\boldsymbol{R}))\boldsymbol{V}\boldsymbol{R}, \end{aligned} \tag{7}$$

where the $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ are training parameters. The bottom-most node span score $s^{in}_{i,i}$ is initialized with 0. Similar to Diora, the VaTiora maintains a $N\times N$ chart $\boldsymbol{T}$ to store intermediate span vectors and scores, i.e., $\boldsymbol{h}^{in}_{i,j}$, $\boldsymbol{h}^{out}_{i,j}$ and $s^{in}_{i,j}$, $s^{out}_{i,j}$, for inside and outside representation respectively. We further fuse the visual and speech features into span vector $\boldsymbol{h}^{in}_{i,j}$ and score $s^{in}_{i,j}$. First, for each span (i,j) and its decomposition (i,k,j), we enhance $\boldsymbol{h}^{in}_{i,j}$ by:

$$\begin{aligned} \hat{\boldsymbol{h}}^{in}_{i,k} &= \boldsymbol{h}^{in}_{i,k} + \gamma\boldsymbol{u}_{i,k} + \lambda\hat{\boldsymbol{p}}_{i,j},\\ \boldsymbol{h}^{in}_{i,j,k} &= \text{MLP}(\hat{\boldsymbol{h}}^{in}_{i,k};\hat{\boldsymbol{h}}^{in}_{k+1,j}),\\ \boldsymbol{h}^{in}_{i,j} &= \sum_k \boldsymbol{h}^{in}_{i,j,k}\cdot\underset{k}{\text{Softmax}}(s^{in}_{i,j,k}), \end{aligned} \tag{8}$$

where $1\le i<k<j\le N$, and $(;)$ denotes the concatenation. γ and λ are the fusion weights. The MLP (Multi-Layer Perceptron) is the network contains multiple linear layers, i.e., $\boldsymbol{h}_{l+1} = \sigma(\boldsymbol{W}_l\boldsymbol{h}_l + \boldsymbol{b}_l)$, where σ is an activation function, $\boldsymbol{W}_l$ and $\boldsymbol{b}_l$ are training parameters. We adopt a two-layer MLP here, which is used as the composition function to merge two spans (see Diora [8]). The $s^{in}_{i,j,k}$ is the decomposition score calculated during the inside pass (Eq. (12)). Then $\boldsymbol{h}^{in}_{i,j}$ could be weighted sum of $\boldsymbol{h}^{in}_{i,j,k}$. $\boldsymbol{u}_{i,j}$ and $\boldsymbol{p}_{i,j}$ are the visual and pitch feature, obtained by:

[1] In the textless settings, we initialize $\boldsymbol{h}^{in}_{i,i}$ with only segmented speech clips $\boldsymbol{r}_i$, i.e., $\boldsymbol{h}^{in}_{i,i} = \text{Norm}(\boldsymbol{r}_i)$.

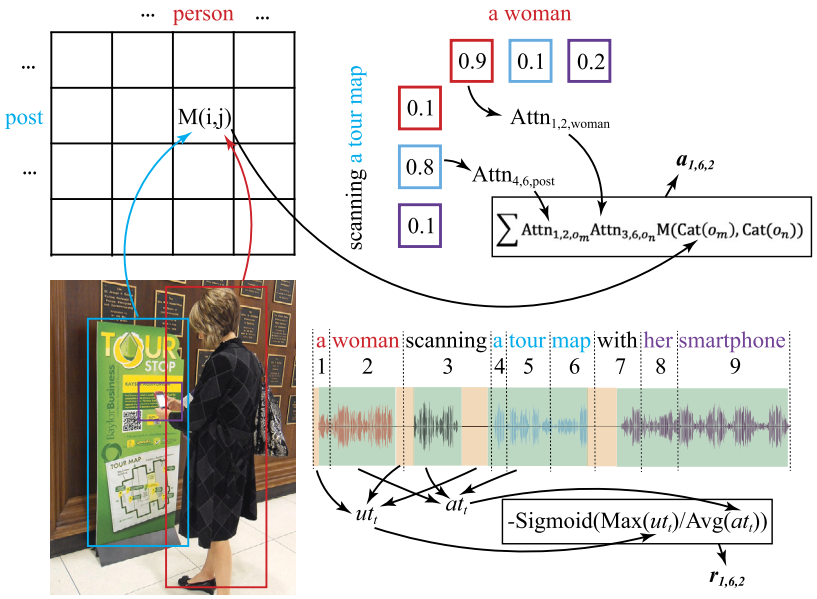


Fig. 4. Illustration of pair feature and voice activity feature.

$$\begin{aligned} \boldsymbol{u}_{i,j} &= \sum_{m}^{M} (\boldsymbol{v}_m + \boldsymbol{l}_m) \cdot \text{Softmax}(\boldsymbol{v}_m^\top \boldsymbol{h}_{i,j}^{in}), \\ \hat{\boldsymbol{p}}_{i,j} &= \underset{k \in [i,j]}{\text{AvgPool}}(\hat{\boldsymbol{p}}_k), \end{aligned} \tag{9}$$

where $\boldsymbol{v}_m$ is the object-level region features and $\boldsymbol{l}_m$ is the embedding of the bouding box coordinates b_m. $\hat{\boldsymbol{p}}_k$ is the pitch feature obtained in Eq. (6). AvgPool$(\cdot)$ is the vector averaging operation.

For calculating the span score $s_{i,j}^{in}$, we consider fusing visual region composition and voice activity features. Specifically, for each span (i, j), we take its decomposition (i, k, j) into account, where $1 \leq i < k < j \leq N$. Technically, we calculate a decomposition score $s_{i,j,k}^{in}$ with visual region composition score $r_{i,j,k}$ and the voice activity time-based decomposition score $a_{i,j,k}$. For $r_{i,j,k}$, we compute:

$$\begin{aligned} r_{i,j,k} &= \sum_{o_m, o_n} \text{Attn}_{i,k,o_m} \cdot \text{Attn}_{k+1,j,o_n} \cdot \boldsymbol{M}_{\text{Cat}(o_m),\text{Cat}(o_n)}, \\ \text{Attn}_{i,k,o_m} &= \text{Softmax}(\boldsymbol{v}_m^\top \boldsymbol{h}_{i,k}^{in}), \\ \text{Attn}_{k+1,j,o_n} &= \text{Softmax}(\boldsymbol{v}_n^\top \boldsymbol{h}_{k+1,j}^{in}) \end{aligned} \tag{10}$$

where $\boldsymbol{M}$ is the pair relevance matrix in Eq. (5) and Cat$(*)$ retrieve the category index of an object. For $a_{i,j,k}$, we compute:

$$a_{i,j,k} = -\text{Sigmoid}(\text{Max}(ut_t)/\text{Avg}(at_t)),\ t \in [i, j) \tag{11}$$

where ut_t and at_t are the non-voice time and voice activity time of speech clip t. Max is the maximum functions. $a_{i,j,k}$ represents the density degree of the span (i, j) in speech rhythm view. Then the $s_{i,j,k}^{in}$ is computed as:

$$\begin{aligned} s_{i,j,k}^{in} =& (\hat{\boldsymbol{h}}_{i,k}^{in})^\top \hat{\boldsymbol{W}}_\theta (\hat{\boldsymbol{h}}_{k+1,j}^{in}) + r_{i,j,k} \\ &+ a_{i,j,k} + s_{i,k}^{in} + s_{k+1,j}^{in}, \end{aligned} \tag{12}$$

where the $\hat{\boldsymbol{W}}_\theta$ is the learnable parameter. The final $s_{i,j}^{in}$ is obtained:

$$s_{i,j}^{in} = \sum_{k} s_{i,j,k}^{in} \underset{k}{\text{Softmax}}(s_{i,j,k}^{in}). \tag{13}$$

Fig. 4 visually illustrates the contributions of these features.

Inside-Outside Pass. The enhanced span vector $\boldsymbol{h}_{i,j}^{in}$ and score $s_{i,j}^{in}$ are then utilized to perform an inside-outside pass similar to Diora. The process is based on the inside-outside algorithm [1], which is used for the derivation process of probabilistic context-free grammar. This method produces representations for all internal nodes via recursively filling a chart, where each cell represents a node in the latent tree.

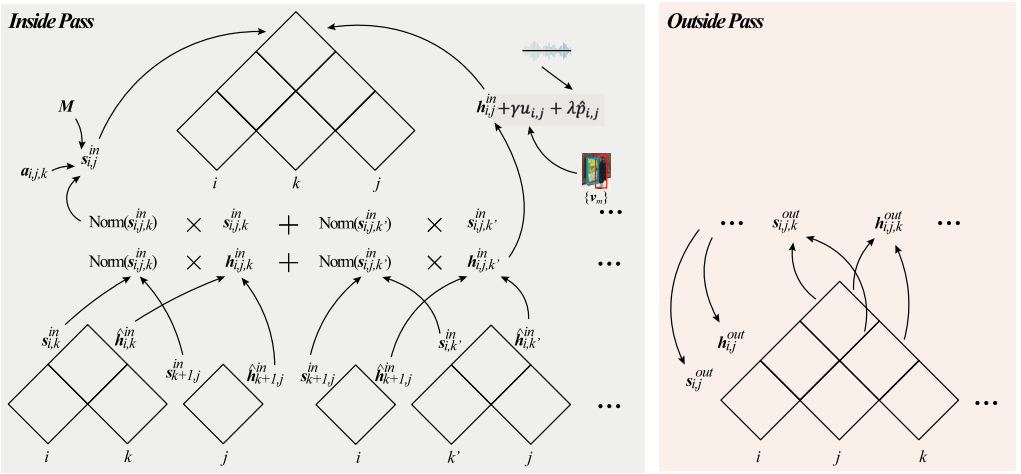


Fig. 5. The process of inside and outside pass, respectively.

The inside pass of this method recursively compresses the input sequence, at each step inputting the vector representations of the two children into a composition function (i.e. Equation (8)) that outputs an inside vector representation of the parent. This process continues up to the root of the tree, eventually yielding a single vector representing the entire sentence (Fig. 5 left part). In the inside pass, each span (i, j) computes inside vector $\boldsymbol{h}_{i,j}^{in}$ and score $s_{i,j}^{in}$ by weighted summing all possible substructures.

Following that, the outside pass computes the outside presentations via a top-down process (Fig. 5 right part). The outside representations are encoded by looking at only the context of a given sub-tree. The root node of the outside chart is learned as a bias. Descendant cells are predicted using a disambiguation over the possible outside contexts. Each component of the context consists of a sibling cell from the inside chart and a parent cell from the outside chart. In the outside pass, we compute $\boldsymbol{h}_{i,j}^{out}$ and $s_{i,j}^{out}$ from top to bottom, where the bottom-most vector $\boldsymbol{h}_{i,i}^{out}$ is used to reconstruct the words. For a span (i, j) and a k out of (i, j):

$$\boldsymbol{h}_{i,j,k}^{out} = \begin{cases} \text{MLP}(\boldsymbol{h}_{i,k}^{out}; \boldsymbol{h}_{j+1,k}^{in}) & k > j \\ \text{MLP}(\boldsymbol{h}_{k,j}^{out}; \boldsymbol{h}_{k,i-1}^{in}) & k < i \end{cases} \tag{14}$$

$$s_{i,j,k}^{out} = \begin{cases} (\boldsymbol{h}_{i,k}^{out})^{\top} \boldsymbol{W} (\boldsymbol{h}_{j+1,k}^{in}) \\ \quad + s_{i,k}^{out} + s_{j+1,k}^{in} & k > j \\ (\boldsymbol{h}_{k,j}^{out})^{\top} \boldsymbol{W} (\boldsymbol{h}_{k,i-1}^{in}) \\ \quad + s_{k,j}^{out} + s_{k,i-1}^{in} & k < i \end{cases} \tag{15}$$

Similarly to the inside pass, the outside span representation $\boldsymbol{h}_{i,j}^{out}$ and score $s_{i,j}^{out}$ is:

$$\begin{aligned} \boldsymbol{h}_{i,j}^{out} &= \sum_k \boldsymbol{h}_{i,j,k}^{out} \cdot \underset{k}{\text{Softmax}}(s_{i,j,k}^{out}), \\ s_{i,j}^{out} &= \sum_k s_{i,j,k}^{out} \cdot \underset{k}{\text{Softmax}}(s_{i,j,k}^{out}), \end{aligned} \tag{16}$$

Finally, the span score is calculated as $q(i, j) = s_{i,j}^{in} \cdot s_{i,j}^{out} / s_{1,n}^{in}$ to measure how likely the span (i, j) exists.

4.3. Overall training

Structure Reconstruction and Contrastive Learning. Following [58,8], we adopt structure reconstruction and contrastive learning. For reconstruction, a self-supervised blank-filling objective is defined as:

$$\mathcal{L}_{rec} = -\frac{1}{n} \sum_i \log P(x_i | \boldsymbol{h}_{i,i}^{out}). \tag{17}$$

For contrastive learning, we randomly select unpaired image I' and span $(i, j)'$ as negative samples within a training batch, and calculate the contrastive objective:

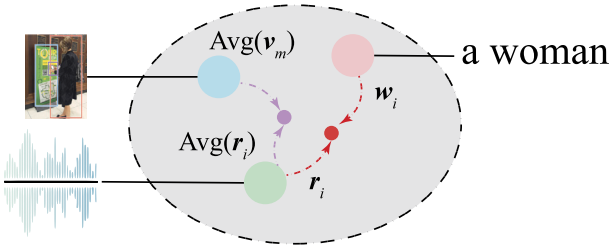


Fig. 6. Illustration of the representation learning to map word-level speech clip r_i and word embedding w_i, and whole image Avg($\{v_m\}$) and speech Avg($\{r_i\}$).

$$
\begin{aligned}
l_{span}(I,i,j) &= \max\{0, d(I,(i,j)') - d(I,(i,j)) + \epsilon\} \\
&\quad + \max\{0, d(I',(i,j)) - d(I,(i,j)) + \epsilon\}, \\
d(I,(i,j)) &= sim((i,j),I) \times q(i,j), i \neq j, \\
sim(i,j,I) &= \max_{m\in[0,M]} \{v_m^\top (h_{i,j}^{in} + h_{i,j}^{out})\},
\end{aligned}
\tag{18}
$$

where ϵ is the positive margin. $q(i,j)$ is the span score as mentioned above ($q(i,j) = s_{i,j}^{in} \cdot s_{i,j}^{out} / s_{1,n}^{in}$). For bottom-most (i,i), e.g., the word w_i (or the speech clip c_i in textless setting), we compute:

$$
l_{word}(I,i) = -\log \frac{\exp(sim(i,I))}{\sum_{\hat{I}\in batch} \exp(sim(i,\hat{I}))},
\tag{19}
$$

where $sim(i,I) = \max_{m\in[0,M]}\{v_m^\top h_{i,i}\}$ The final contrastive loss will be:

$$
\mathcal{L}_{cl} = \sum_{i,j,i\neq j} l_{span}(I,i,j) + \sum_i l_{word}(I,i).
\tag{20}
$$

Representation Learning. Furthermore, we propose to use a representation learning objective to align the vectors of multi-modal inputs in the feature space (see Fig. 6):

$$
\begin{aligned}
&//\textit{Full setting} \\
&\mathcal{L}_{rep} = \text{Cos}(r_i, w_i) + \text{Cos}(\text{Avg}(\{r_i\}), \text{Avg}(\{v_m\})), \\
&//\textit{Textless setting} \\
&\mathcal{L}_{rep} = \text{Cos}(\text{Avg}(\{r_i\}), \text{Avg}(\{v_m\})),
\end{aligned}
\tag{21}
$$

where Cos means cosine similarity. Avg($\{r_i\}$), Avg($\{v_m\}$) means average of the clip and object vectors, representing the whole speech and image feature. In textless setting, there are only Cos(Avg($\{r_i\}$), Avg($\{v_m\}$)). The whole training loss will be:

$$
\mathcal{L} = \mathcal{L}_{rec} + \alpha_1 \mathcal{L}_{cl} + \alpha_2 \mathcal{L}_{rep}.
\tag{22}
$$

In the inference stage, we follow the conventional GI method to predict the tree with the maximum inside scores via the CKY algorithm.

5. Novel dataset for VAT-GI

VAT-GI can be evaluated on the SpokenCOCO dataset [20], which contains approximately 600,000 speech recordings of MSCOCO image captions.[2] Nevertheless, SpokenCOCO can fall prey to limitations in terms of its *visual scenario styles, short captions*, and *monotonous speech tone*, resulting in inferior generalization performance of VAT-GI parser across diverse language environments. Thus, we introduce a more challenging test set, namely SpokenStory. The dataset is built by extending the image caption data with speech records. We comprises 1,000 images extracted from OpenImages [33] across various scenarios, and accompanied by corresponding caption annotations from [47]. OpenImages is widely used for computer vision research and applications, which contains over 9 million labeled images sourced from the web and covers more than 6,000 topics and categories. Each image is annotated with labels that describe the objects or concepts present in the image.

We adopt the annotations of Localized Narratives [47] for aligned captions, and we manually record speech for each caption by human speakers, who are recruited in our research team, producing the aligned transcripts and speech audios. The average sentence length of SpokenStory is 20 words, with an average speech duration of 5.01 seconds, roughly twice that of SpokenCOCO. Fig. 7 shows the comparison of captions between our SpokenStory and SpokenCOCO, where the sentence in SpkenStory has longer constituents. The speech recordings encompass five distinct tone styles, i.e., "*Monotonous*", "*Excited*", "*Tense*", "*Peaceful*" and "*Eccentric*", with each

[2] https://cocodataset.org/.

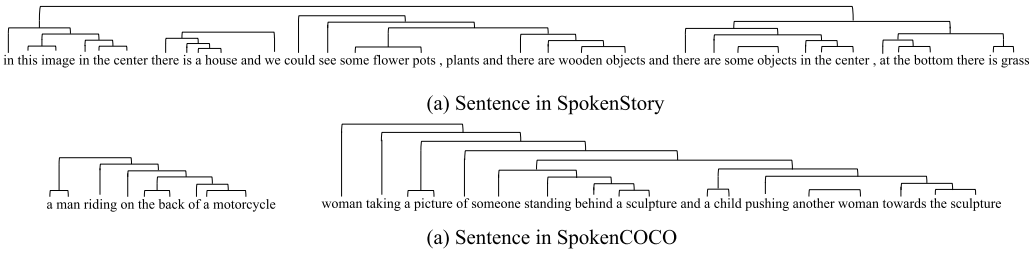


Fig. 7. Comparison of captions between our SpokenStory and SpokenCOCO.

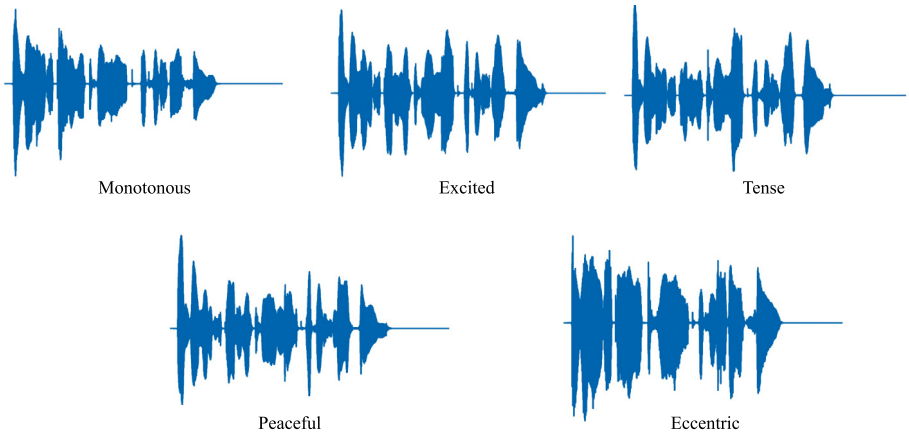


Fig. 8. Visualized waveform for the sentence "*A woman scanning a tour map with her smartphone*" in the five emotions.

Table 2
Statistics of SpokenCOCO and SpokenStory. We follow [25] to split SpokenCOCO. "Avg. Sppech Len." means average speech length. "Agb. Sent. Len." means average sentence length.

	SpokenCOCO	**SpokenStory**
Split	Train/Val/Test	Test
Images	80 K/5 K/5 K	1 K
Visual scenario	91	> 300
Avg. Speech Len. (second)	2.54	5.01
Tone Styles	Monotonous	5 Styles
Avg. Sent. Len. (word)	10.46	20.69

characterized by specific intonation and rhythm. We do not suggest speakers how to express each emotion while letting them express all by their own comprehension. Fig. 8 illustrates the waveform of five types of speech in the same sentence. We develop a tool for recording and transcript annotation. The speakers first record the captions and the system will return a pseudo transcript via an off-the-shelf speech segmentation model. After recording, the speakers are asked to adjust the transcript to check the word boundaries. Then the recording and the transcripts will be saved into the database. Overall, the new test set poses a more challenging setting, due to its longer sentence, complex constituents and diverse intonations and tones. Table 2 compares two datasets.

6. Experiments

6.1. Settings

Following [64], we use the Faster-RCNN model to detect object regions and extract visual features of the image. For fair comparison, we follow [58] that use ELMo [14] for text embedding. Note the PLM or LLM embedding may contain potential knowledge of language structures during the pre-training stage, which may introduce immeasurable interference to the experiments. Following previous work, the ground-truth constituent structures are parsed with the Benepar [29], and then we take the hand-checked text set from Shi et al. [51]. We use VG-HuBert [41] model for the speech encoding and word segmentation, which is pre-trained on unlabeled speech-image pairs. For the inside-outside recursive autoencoder, we follow [8] to use an MLP as the composition function for both inside and outside passes. We compare our VaToria with current state-of-the-art methods on two kinds of settings: 1) text-only grammar induction. 2) Visual-text grammar induction on the same dataset, so that we could explicitly explore the influence of different modalities. The model hyper-parameters are listed in Table 3. Other settings have defaulted to Diora.

Table 3
Model hyperparameters.

Hyper-param.	Value	Hyper-param.	Value
dimension of ROI feature	2048	α_2	0.5
dimension of Speech hiddens	1024	max text length	80
dimention of word embeddings	400	optimizer	Adam
number of RoI	36	dropout	0.1
tIoU threshould of SCF1	0.5	learning rate	1e-4
γ	0.5	bach size	64
λ	0.5	epoch	30
α_1	0.5		

Table 4
Main results of Corpus-F1 and Sent-F1 on the SpokenCOCO test set. In the VAT-GI-Textless setting, we report Corpus-SCF1 and Sent-SCF1. The up arrows represent the significance level of our model to the baselines.

	Corpus-F1	Sent-F1
• **Text-only GI**		
Left Branch	15.1(↑↑↑)	15.7(↑↑↑)
Right Branch	51.0↑↑↑	51.8(↑↑↑)
Random	24.2(↑↑↑)	24.6(↑↑↑)
C-PCFG	53.6(↑↑↑)	53.7(↑↑↑)
Diora	58.3(↑↑↑)	59.1(↑↑↑)
• **Visual-Text GI**		
VG-NSL	50.4(↑↑↑)	-
VG-NSL+HI	53.3(↑↑↑)	-
VC-PCFG	59.3(↑↑↑)	59.4(↑↑↑)
Cliora	61.0(↑↑)	61.7(↑↑↑)
• **VAT-GI**		
VaTiora	**62.7**	**63.0**
	Corpus-SCF1	**Sent-SCF1**
• **VAT-GI-Textless**		
VaTiora	32.5	31.2

6.2. Baseline specification

We compare our model with baselines in two types of settings:

Text only methods which only takes text as inputs.

- **Left Branching**, A rule-based construction method that compose each span with its next word to a higher span from left to right.
- **Right Branching**, A rule-based construction method that compose each span with its prior word to a higher span from right to left.
- **Random**, Randomly composing spans to a tree.
- **C-PCFG**, A method based on a compound probabilistic context-free grammar, where the grammar's rule probabilities are modulated by a per-sentence continuous latent variable.
- **Diora**, A deep inside-outside recursive autoencoder, which incorporates the inside-outside algorithm to compute the constituents and construct the constituency structure. This is also the backbone of our framework.

Visual-text methods which takes aligned image and description text as inputs.

- **VG-NSL**, A visually grounded neural syntax learner that learns a parser from aligned image-sentence pairs. The model is optimized via REINFORCE, where the reward is computed by scoring the alignment of images and constituents.
- **VG-NSL+HI**, VG-NSL model with the design of head-directionality inductive biases, encouraging the model to associate abstract words with the succeeding constituents instead of the preceding ones.
- **VC-PCFG**, An extended model of C-PCFG with visually grounded learning.
- **Cliora**, An extended model of Diora with a fine-grained visual-text alignment via mapping RoI regions and spans. The model computes a matching score between each constituent and image region, trained via contrastive learning.

6.3. Significance test

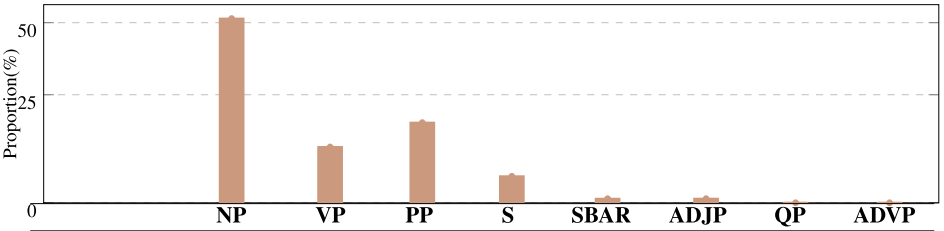
We employ five different random seeds to initialize the model parameters and the data shuffling to obtain a set of results, which are used for significance test. We apply two-tailed T-test [46] for the significance and use the p value to measure the significance

Table 5
Corpus-F1 on the SpokenCOCO test set and SpokenStory set. "AvgSentLen" means the average sentence length. All the models are trained on the SpokenCOCO train set.

	SpokenCOCO	SpokenStory
• **AvgSentLen**	10.46	20.69
Diora	58.3(↑↑↑)	35.74(↑↑↑)
Cliora	61.0(↑↑)	37.82(↑↑↑)
VaTiora	**62.70**	**53.19**

Table 6
Module ablation results. "Pitch" and "VA' denote pitch feature and voice activity feature, i.e., $\boldsymbol{p}_i$ and $a_{i,j,k}$. "w/o Vision" means ablating all the visual features. "Region" denotes region features, ablating which means using the feature of the whole image instead. "Pair" means region pair score, i.e., $\boldsymbol{M}$.

Speech	Pitch	VA	Visual	Region	Pair	Corpus-F1
✓	✓	✓	✓	✓	✓	62.70
✓	✓	✗	✓	✓	✓	62.23
✓	✗	✓	✓	✓	✓	62.10
✓	✗	✗	✓	✓	✓	61.38
✗	-	-	✓	✓	✓	60.07
✓	✓	✓	✓	✓	✗	62.31
✓	✓	✓	✓	✗	-	61.24
✓	✓	✓	✗	-	-	60.72
✓	✓	✗	✗	-	-	59.76
✓	✗	✓	✗	-	-	60.07
✓	✗	✗	✗	-	-	59.21
Text-only (Diora)						58.30



	NP	VP	PP	S	SBAR	ADJP	QP	ADVP
Left Branch	33.20	0.00	0.10	0.00	0.00	4.90	0.00	0.00
Right Branch	37.50	**94.50**	71.10	0.00	**97.80**	20.90	0.00	79.10
Random	32.80	18.40	24.40	0.10	17.70	26.80	0.00	20.90
C-PCFG†	43.00	85.00	78.40	-	90.60	36.60	-	**87.40**
Diora⋆	68.20	55.89	52.96	71.02	54.56	53.76	71.29	61.68
VG-NSL†	64.30	28.10	32.20	-	16.90	13.20	-	5.60
VG-NSL+HI†	61.00	33.50	62.70	-	42.00	13.90	-	65.90
VC-PCFG†	54.90	83.20	**80.90**	-	89.00	38.80	-	86.30
Cliora⋆	70.38	56.82	51.70	73.34	57.01	56.91	68.74	66.48
VaTiora	**71.92**	60.17	53.05	**74.58**	59.40	**58.22**	**71.35**	65.10
w/o Vision	67.04	58.82	50.97	75.11	59.40	58.21	71.35	65.30
w/o Speech	70.40	55.95	52.97	73.40	59.40	58.01	71.66	65.10

Fig. 9. Recall of eight constituent labels on SpokenCOCO test set. The best mean number in each column is in bold. † indicates results reported by Zhao and Titov [63]. ⋆ indicates the results are obtained by running their code on our dataset. The bar chart describes the distribution of each label.

levels of all our results. The significance levels are reported in each table via a postfix tag, where '↑', '↑↑', '↑↑↑' denote $p < 0.1$, $p < 0.05$, and $p < 0.01$. The '-' denote that the results are not significant or lower than the baselines.

6.4. Main results

As shown in Table 4, we compare our VAT-GI with two settings, i.e., text-only and visual-text. When comparing the PCFG-based methods (i.e., C-PCFG and VC-PCFG) and Diora-based methods (i.e., Diora and Cliora) separately, we can see the models in visual-text setting perform better, demonstrating that the visual information benefits the GI. Overall, the Diora-based methods outperform the other methods. When extending to VAT-GI, our model outperforms the state-of-the-art Cliora by leveraging both visual and speech information, demonstrating the multiple modality features helps better grammar induction. In the textless setting, the performance declines sharply, which is reasonable due to the lack of text-level structure. But we can still learn the latent structure by speech and

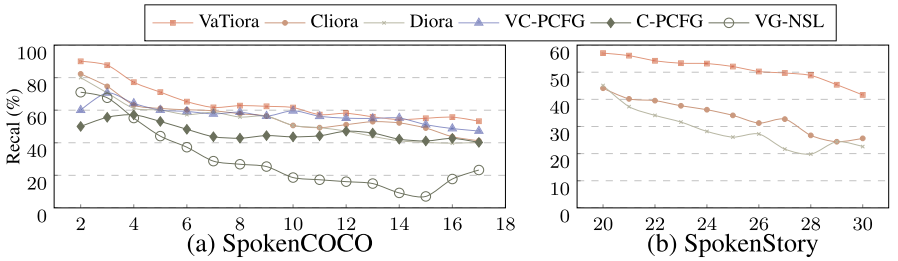


Fig. 10. Recall curves varying by constituent length on SpokenCOCO and SpokenStory.

outperforms the random setting. We also evaluate our model on the new proposed dataset SpokenStory. For a fair comparison, we report the three Diora-based methods trained on the SpokenCOCO train set. The results are presented in Table 5. Among these methods, our VaTiora demonstrates significant improvements, particularly in the more challenging SpokenStory dataset, gaining about 15.37 Corpus-F1 over the current best-performing model, Cliora. The results largely indicate the potential of VaTiora in effectively handling complex constituent structures.

6.5. Module ablation

We display the quantified contributions of each component of VaTiora in Table 6. The results reveal that speech features and visual features contribute 2.63 and 1.98 corpus-F1, respectively, with speech exhibiting a more significant impact. More concretely, we conduct ablations on fine-grained speech and vision features individually. For speech, the pitch and voice activity features collectively contribute 1.32 F1, with pitch playing a more substantial role. Concerning vision, using the feature of the entire image instead of mapping object regions to spans results in a performance decay to 61.24, as the entire image compromises the perception of visual hierarchical structures. Additionally, the region pairs score (i.e. $\boldsymbol{M}$) provides valuable information on high-level region composition. Moreover, the cross-attention and LSTM modules in VaTiora enhance the system by facilitating multi-modal interaction and capturing pitch patterns.

6.6. Analysis and discussions

We now take one step further, exploring the task and the proposed method from various angles, so as to gain a deeper understanding.

Performance on Span Labels. We present a detailed analysis of the performance of eight frequent constituency labels. Fig. 9 displays a bar chart indicating the proportion of each span label in SpokenCOCO, while the table provides the recall scores for each method. Overall, our model consistently outperforms other diora-based methods across all span types and achieves optimal performance on NP, S, ADJP, and QP constituents. The performance of NPs is generally better than other constituents across most methods, primarily due to their dominant proportion (approximately 50%). Vision-based methods (VC-PCFG and Cliora) notably outperform their language-only counterparts (C-PCFG and Dliora) on NPs. This can be attributed to the visual information aiding in optimizing the models to focus more on NPs, which are typically noun phrases describing visual scenes. It is noteworthy that the right branching approach exhibits remarkably high recall on VPs (94.5%) and SBARs (97.8%). This can be explained by the fact that in a sentence, the predicate and its object phrase, often located on the right side, combine to form a VP constituent. Similarly, for SBARs, the subordinate clause is usually positioned after the main clause, which is effectively captured by the right branching strategy. We also compare the results when removing visual features or speech features, shown in the last two rows in the table. The performance on NP spans drops a lot without visual features while keep comparable when only removing speech features. On the contrary, the speech features contribute a lot to the performance on VP spans. The results explain the different impacts of the two modalities on different grammar components.

Influence of Constituent Length. In Fig. 10, we analyze model performance for constituent lengths on SpokenCOCO and SpokenStory, where on SpokenStory we mainly report the results for super long constituents (more than 20 tokens). Overall, the performance of all models tends to weaken as the sentence length increases. Diora-based methods exhibit consistent performance across varying sentence lengths, while VG-NSL experiences a significant drop. PCFG-based methods initially lag behind when the sentence length is short and maintain stable performance for sentence lengths above 5. Notably, our VaTiora consistently outperforms other methods across all scenarios. For long constituents in the SpokenStory, we evaluate the performance of Diora, Cliora, and VaTiora. Since the average length of SpokenStory sentences exceeds 20, we can specifically examine these long constituents that are not present in the SpokenCOCO. It can be noted that VaTiora still showcases the best performance among them.

Influence of Speech Quality. We explore the influence of speech quality by comparing the model on real human voice and mechanically synthetic speech. Specifically, we evaluate the model on SpokenCOCO (with human voice) and SpeechCOCO [3] (with synthetic

[3] https://zenodo.org/record/4282267.

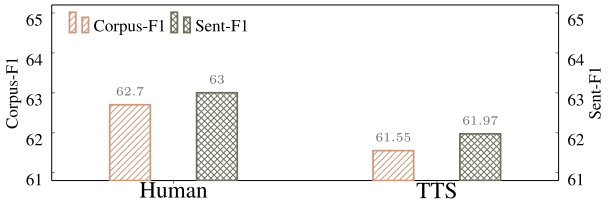


Fig. 11. Comparison of SpeechCOCO ("TTS") and SpokenCOCO ("Human").

Table 7
Corpus-F1 of VaTiora on the SpokenCOCO with different SNR. Smaller SNR means heavier noise.

SNR (dB)	Corpus-F1 (%)
-5	55.31
0	58.24
5	61.92
10	62.04

Table 8
Phrase grounding results on the built 1,000 COCO test samples.

	Acc(%)
Cliora	61.51
VaTiora	**62.76**

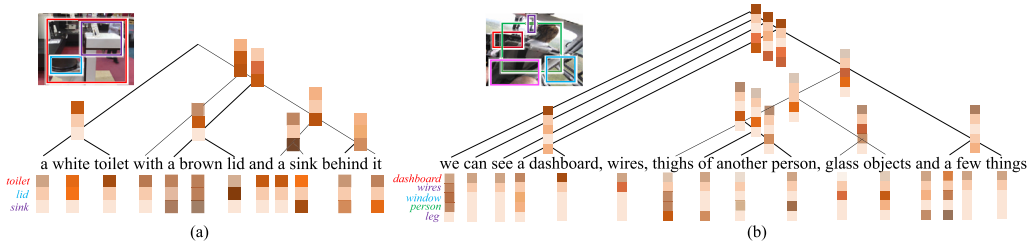


Fig. 12. Cases of grounding between spans and regions.

speech constructed by test-to-speech (TTS) technology). Note that we specially select SpeechCOCO as the comparison because it adopts an outdated TTS technology, where the synthetic speech has significant difference to the human voice in speech quality. As shown in Fig. 11, the VaTiora demonstrates superior performance in the real human voice. This is because synthetic speech in SpeechCOCO exhibits a rigid tone while human voice is more natural and contains more reasonable intonation and rhythm owing to human language intuition.

Further, we explore the model's performance in a noisy environment. Table 7 shows the model performance under different signal-to-noise ratios (SNR). It is evident that light noise shows a relatively minor impact on performance as it causes minimal corruption to the pitch frequency and the voice activity time [37]. However, as the noise becomes substantial enough to contaminate the entire waveform, the model's performance experiences a noticeable decline.

Analysis of Grounding. We report the phrase localization accuracy to measure the grounding performance of our model. We use a similarity score to ground each span and image region:

$$g(i, j, o_m) = \text{Softmax}(\boldsymbol{v}_m^\top \boldsymbol{h}_{i,j}^{in}), \tag{23}$$

where $\boldsymbol{v}_m$ is the object feature in Table 1, and $\boldsymbol{h}_{i,j}^{in}$ is the span representation in Equation (8). Then we use $Argmax_m(g(i, j, o_m))$ to find the grounded object. Due to the fact that the COCO dataset has not phrase grounding annotations, we manually build test set containing 1,000 samples. The results are shown in Table 8, where our model demonstrates good performance and outperforms previous Cliora method.

We also show the visualized analysis for span-region grounding in Fig. 12. Then we calculate the attention weights ($\text{Softmax}(\boldsymbol{v}_m^\top \boldsymbol{h}_{i,j}^{in})$) for each span and visualize them in a heat map chart. Example (a) is chosen from SpokenCOCO, which contains three main objects in a hierarchical structure, i.e., the "lid" and the "sink" combine into the "toilet". We can see the NP phrases representation could focus on the corresponding object regions, demonstrating a good grounding performance. The high-level span also has the attention weights to the union regions of its sub-spans. Example (b) is chosen from SpokenStory, which has longer sentences and more detected

Table 9
SCF1 results on different word segmentations on SpokenCOCO test set.

Segmentation Methods	tIoU	SCF1
VG-HuBert-Golden	100	**53.07**
VG-HuBert	44.02	32.50
VG-W2V2	45.56	32.42
HuBERT	37.72	31.59
W2V2	38.56	31.77
ResDAVEnet-VQ	23.78	30.02

Table 10
Corpus-F1 results of different word embeddings on SpokenCOCO test set.

	Corpus-F1
Random Initialization	62.58
Elmo	**62.70**
Glove	62.49

Table 11
Corpus-F1 results of different RoI numbers in each image on SpokenCOCO test set.

RoI NUM	Corpus-F1
3	58.01
10	60.10
36	**62.70**
100	62.58

objects. Our model could also allocate a roughly right attention distribution for each span. For the NPs that have a clear name, such as “a dashboard”, or “wires”, the model could learn a sharp attention distribution, while for spans about “objects” and “things” that represent general objects, it performs poorly to attend to the right targets. However, this weakness shall be the prevalent problem in unsupervised learning. Overall, our model could fully make use of the hierarchical structures of images via span-object mapping.

Influence of Word Segmentation in Textless Setting To explore the influence of word segmentation, we compare the results with predicted word segments in Table 9. For evaluation, we adopt the Corpus-F1 to measure the predicted structure and use tIoU to measure the accuracy of word segmentation. We can see that the model with golden word segmentation could achieve comparable performance (53.07) with text-based GI. When the accuracy of word segmentation falls, the final F1 score decreases.

Influence of Word Embedding We compare the results when using different word embeddings in Table 10. The Elmo outperforms the others, while there is not a very large margin among these methods. The random embedding could also achieve comparable performance.

6.7. Influence of object detection

We compare the results of different numbers of RoI in Table 11. We list the results of 3, 10, 36, 100 RoIs per image. When the RoI number is 36 (which is the default number of Fast-RCNN), the model achieves the best performance. When the number is too small, i.e., 3 or 10, there is insufficient information for further grammar induction. When the number is too large, such as 100, there may be more noisy objects with low confidence that decay the performance.

6.8. Detailed case study

We visualize the predicted tree structures of VaTiora in Fig. 13, where example (a) is chosen from SpokenCOCO and example (b) is chosen from SpokenStory. In example (a), VaTiora wrongly predicts the span “a pink” and “a blue” which should be “pink shirt” and “blue metal”. For most spans, VaTiora gives the correct prediction. In example (b), the sentence contains compound sentences and specific symbols. VaTiora wrongly predicts the span “t-shirt, smiling and standing on the ground”, which should be “wearing t-shirt” and “smiling and standing on the ground”. The VaTiora failed to handle the span containing a comma, which seldom appears in the SpokenCOCO training set. In example (b), we can see the VaTiora could handle the long constituents, such as “can see... on the ground”. Fig. 14 shows the predicted structures in textless VAT-GI. We can see the model may predict wrong clip segments, further leading to missing or incorrect spans. The textless VAT-GI is more challenging due to the lack of text features. In example (a), the predicted tree degenerates into a right-branching tree. In example (b), generally, the model predicts a mediocre result, containing some right structures.

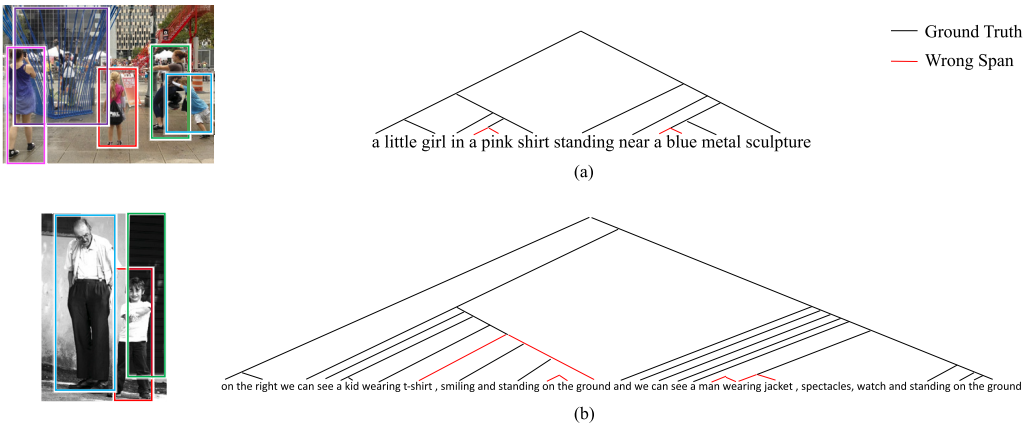


Fig. 13. Case presentation. Example (a) is chosen from SpokenCOCO and example (b) is chosen from SpokenStory. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

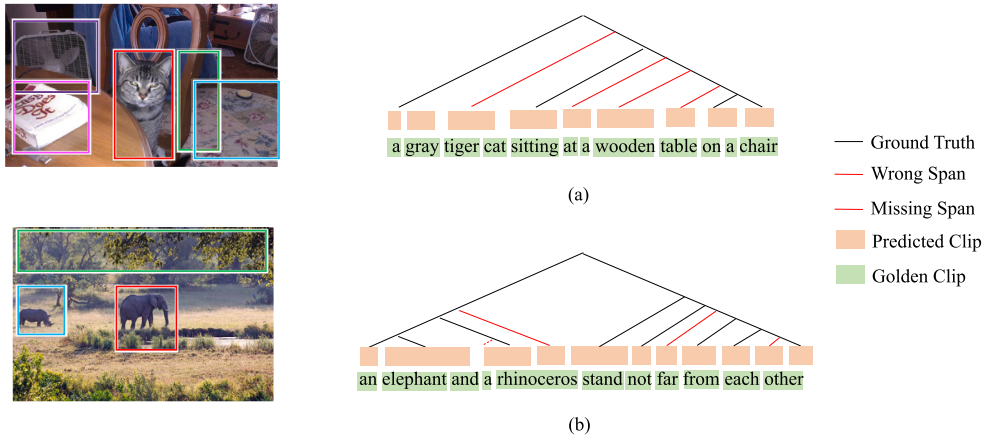


Fig. 14. Case presentation of the textless setting chosen from SpokenCOCO.

7. Limitation and future work

In this work, we propose the VAT-GI task that leverages the multi-modal information from image, speech and text to enhance grammar induction. VAT-GI explores the phylogenetic language acquiring process from the machine learning perspective. Our proposed VaTiora can properly integrate multi-modal structure features into the grammar parsing, providing a beginning attempt for VAT-GI. We further discuss the limitations. First, in VaTiora, the multi-modal feature extraction and grammar parsing is conducted in a pipeline way, which may import noise information due to the limitation of upstream algorithms. Second, the textless setting of VAT-GI is not fully explored, which may reveal the core mechanism of language acquisition and help enhance the current language models.

As future work of VAT-GI, the following aspects are worth exploring:

1) End-to-end multi-modal feature integration. Studying the method to extract multi-modal features from inputs in an end-to-end manner. Currently, the feature extraction in VaTiora tackles three types of inputs, i.e., text, image and speech, respectively and fuse them through a rough vector combination. In this way, the feature quality depends on the design of upstream algorithms, which have inherent limitations. A better way to take an end-to-end architecture is to extract the final feature from multiple inputs, where the fusion is processed in an internally implicit way, such that the noise can be avoided.

2) Embedding the GI into LLMs for enhanced compositionality. Exploring enhancing the large language models (LLMs), and multi-modal LLMs, with the strategy of grammar induction, to help them obtain better performance in multi-modal alignment and strengthen the compositionality ability of language and other modalities.

3) Stronger GI methods. Proposing stronger methods for grammar induction. Diora-based methods are designed for text-only setting, which may not adaptive in multi-modal settings. Also, Diora models the structure through a bottom-up and top-down encoding decoding, which may ignore the latent high-level semantics [8], thus it is necessary to explore more approaches for grammar induction.

4) Exploration on Textless GI. In this work, we also introduce the textless setting of VAT-GI which may reveal the core mechanism of language acquisition and help enhance the current language models. We design our VaTiora system to also support the textless setting. However, this setting has not been fully explored, and we leave this as a promising potential future work.

8. Conclusion

In this paper, we introduce a novel task, visual-audio-text grammar induction (VAT-GI), unsupervisedly inducing the constituent trees from aligned images, text, and speech inputs. First, we propose a visual-audio-text inside-outside recursive autoencoder (VaTiora) to approach the VAT-GI task, in which the rich modal-specific and complementary features are fully leveraged and effectively integrated. To support the textless setting for VAT-GI, we newly devise an aligned span-clip F1 metric (SCF1), which helps conveniently measure the textless constituent structures of the aligned sequences. Further, we construct a new test set SpokenStory, where the data characterizes richer visual scenes, timbre-rich audio and deeper constituent structures, posing new challenges for VAT-GI. Experiments on two benchmark datasets indicate that leveraging multiple modality features helps better grammar induction, and also our proposed VaTiora system is more effective in incorporating the various multimodal signals for stronger performance.

CRediT authorship contribution statement

Yu Zhao: Writing – review & editing, Writing – original draft, Software, Investigation. **Hao Fei:** Writing – review & editing, Writing – original draft, Validation, Conceptualization. **Shengqiong Wu:** Writing – review & editing, Resources, Methodology. **Meishan Zhang:** Writing – review & editing, Validation, Methodology. **Min Zhang:** Writing – review & editing, Resources, Project administration. **Tat-seng Chua:** Writing – review & editing, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is supported by National Natural Science Foundation of China (NSFC) under Grant 62336008.

Data availability

Data will be made available on request.

References

[1] James K. Baker, Trainable grammars for speech recognition, J. Acoust. Soc. Am. 65 (S1) (1979) S132.
[2] Ted Briscoe, Grammatical acquisition: inductive bias and coevolution of language and the language acquisition device, Language 76 (2) (2000) 245–296.
[3] Tara McAllister Byun, Anne-Michelle Tessier, Motor influences on grammar in an emergentist model of phonology, Lang. Linguist. Compass 10 (9) (2016) 431–452.
[4] Jihun Choi, Kang Min Yoo, Sang-goo Lee, Learning to compose task-specific tree structures, in: Sheila A. McIlraith, Kilian Q. Weinberger (Eds.), Proceedings of AAAI, 2018, pp. 5094–5101.
[5] Noam Chomsky, Logical structure in language, J. Am. Soc. Inf. Sci. 8 (4) (1957) 284.
[6] Noam Chomsky, On certain formal properties of grammars, Inf. Control 2 (2) (1959) 137–167.
[7] Shay Cohen, Noah B. Smith, The shared logistic normal distribution for grammar induction, in: Proceedings of the NIPS Workshop on Speech and Language: Unsupervised Latent-Variable Models, 2008.
[8] Andrew Drozdov, Patrick Verga, Mohit Yadav, Mohit Iyyer, Andrew McCallum, Unsupervised latent tree induction with deep inside-outside recursive autoencoders, in: Proceedings of the NAACL, 2019.
[9] Hao Fei, Yafeng Ren, Donghong Ji, Retrofitting structure-aware transformer language model for end tasks, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2151–2161.
[10] Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, Tat-Seng Chua, Lasuie: unifying information extraction with latent adaptive structure-aware generative language model, Adv. Neural Inf. Process. Syst. 35 (2022) 15460–15475.
[11] Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, Donghong Ji, Better combine them together! Integrating syntactic constituency and dependency representations for semantic role labeling, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 549–559.
[12] Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, Shuicheng Yan, Enhancing video-language representations with structural spatio-temporal alignment, IEEE Trans. Pattern Anal. Mach. Intell. 2024 (2024).
[13] Suzanne Flynn, Wayne O'Neil, Linguistic Theory in Second Language Acquisition, vol. 8, Springer Science & Business Media, 2012.
[14] Ross Girshick, Fast r-cnn, in: Proceedings of CVPR 2015, 2015, pp. 1440–1448.
[15] Dave Golland, John DeNero, Jakob Uszkoreit, A feature-rich constituent context model for grammar induction, in: Proceedings of ACL, 2012, pp. 17–22.
[16] David Harwath, Wei-Ning Hsu, James R. Glass, Learning hierarchical discrete linguistic units from visually-grounded speech, in: Proceedings of ICLR, 2020.
[17] David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, James R. Glass, Jointly discovering visual objects and spoken words from raw sensory input, in: Proceedings of ECCV, 2018, pp. 659–677.
[18] Jeffrey Heinz, Computational phonology - Part II: grammars, learning, and the future, Lang. Linguist. Compass 5 (4) (2011) 153–168.
[19] Yining Hong, Qing Li, Song-Chun Zhu, Siyuan Huang, VLGrammar: grounded grammar induction of vision and language, in: Proceedings of ICCV, 2021, pp. 1645–1654.
[20] Wei-Ning Hsu, David Harwath, Tyler Miller, Christopher Song, James R. Glass, Text-free image-to-speech synthesis using learned segmental units, in: Proceedings of ACL/IJCNLP, 2021, pp. 5284–5300.
[21] Xiang Hu, Haitao Mi, Liang Li, Gerard de Melo, Fast-R2D2: a pretrained recursive neural network based on pruned CKY for grammar induction and text representation, in: Proceedings of EMNLP, 2022, pp. 2809–2821.
[22] Elias Iosif, Ioannis Klasinas, Georgia Athanasopoulou, Elisavet Palogiannidi, Spiros Georgiladakis, Katerina Louka, Alexandros Potamianos, Speech understanding for spoken dialogue systems: from corpus harvesting to grammar rule induction, Comput. Speech Lang. 47 (2018) 272–297.

[23] Aren Jansen, Emmanuel Dupoux, Sharon Goldwater, Mark Johnson, Sanjeev Khudanpur, Kenneth Church, Naomi Feldman, Hynek Hermansky, Florian Metze, Richard C. Rose, Mike Seltzer, Pascal Clark, Ian McGraw, Balakrishnan Varadarajan, Erin Bennett, Benjamin Börschinger, Justin T. Chiu, Ewan Dunbar, Abdellah Fourtassi, David Harwath, Chia-ying Lee, Keith D. Levin, Atta Norouzian, Vijayaditya Peddinti, Rachael Richardson, Thomas Schatz, Samuel Thomas, A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition, in: Proceedings of ICASSP, 2013, pp. 8111–8115.
[24] Lifeng Jin, Finale Doshi-Velez, Timothy A. Miller, William Schuler, Lane Schwartz, Depth-bounding is effective: improvements and evaluation of unsupervised PCFG induction, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, 2018, pp. 2721–2731.
[25] Andrej Karpathy, Li Fei-Fei, Deep visual-semantic alignments for generating image descriptions, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 664–676.
[26] Taeuk Kim, Jihun Choi, Daniel Edmiston, Sang-goo Lee, Are pre-trained language models aware of phrases? Simple but strong baselines for grammar induction, in: Proceedings of ICLR, 2020.
[27] Yoon Kim, Chris Dyer, Alexander M. Rush, Compound probabilistic context-free grammars for grammar induction, in: Proceedings of ACL, 2019, pp. 2369–2385.
[28] Yoon Kim, Alexander M. Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, Gábor Melis, Unsupervised recurrent neural network grammars, in: Proceedings of NAACL-HLT, 2019, pp. 1105–1117.
[29] Nikita Kitaev, Dan Klein, Constituency parsing with a self-attentive encoder, in: Proceedings of ACL, 2018, pp. 2676–2686.
[30] Ioannis Klasinas, Alexandros Potamianos, Elias Iosif, Spiros Georgiladakis, Gianluca Mameli, Web data harvesting for speech understanding grammar induction, in: Proceedings of INTERSPEECH, 2013, pp. 2733–2737.
[31] Dan Klein, Christopher D. Manning, A generative constituent-context model for improved grammar induction, in: Proceedings of the ACL, 2002, pp. 128–135.
[32] Kenichi Kurihara, Taisuke Sato, Variational Bayesian grammar induction for natural language, in: Proceedings of ICGI, vol. 4201, 2006, pp. 84–96.
[33] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, Vittorio Ferrari, The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale, Int. J. Comput. Vis. 2020 (2020).
[34] Cheng-I Jeff Lai, Freda Shi, Puyuan Peng, Yoon Kim, Kevin Gimpel, Shiyu Chang, Yung-Sung Chuang, Saurabhchand Bhati, David D. Cox, David Harwath, Yang Zhang, Karen Livescu, James R. Glass, Audio-visual neural syntax acquisition, CoRR, arXiv:2310.07654 [abs], arXiv:2310.07654, 2023.
[35] Cheng-I Jeff Lai, Freda Shi, Puyuan Peng, Yoon Kim, Kevin Gimpel, Shiyu Chang, Yung-Sung Chuang, Saurabhchand Bhati, David D. Cox, David Harwath, Yang Zhang, Karen Livescu, James R. Glass, Audio-visual neural syntax acquisition, in: IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023, IEEE, 2023, pp. 1–8.
[36] Karim Lari, Steve J. Young, The estimation of stochastic context-free grammars using the inside-outside algorithm, Comput. Speech Lang. 4 (1) (1990) 35–56.
[37] Yuzhou Liu, DeLiang Wang, Time and frequency domain long short-term memory for noise robust pitch tracking, in: Proceedings of ICASSP, 2017, pp. 5600–5604.
[38] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Process. Syst. 26 (2013).
[39] Karin Müller, Probabilistic context-free grammars for phonology, in: Proceedings of SIGMORPHON, 2002, pp. 70–80.
[40] Daniel N. Osherson, Lila R. Gleitman, An Invitation to Cognitive Science: Language, vol. 1, MIT Press, 1995.
[41] Puyuan Peng, David Harwath, Word discovery in visually grounded, self-supervised speech models, in: Proceedings of INTERSPEECH, 2022, pp. 2823–2827.
[42] Yella Dezas Perdani, Enhancing the students' grammar comprehension by utilizing the video-based instruction, in: Proceedings of ICETC, 2022, pp. 316–321.
[43] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer, Deep contextualized word representations, in: Proceedings of NAACL-HLT, 2018, pp. 2227–2237.
[44] James M. Pickett, The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology, Allyn & Bacon, 1999.
[45] A.J. Piergiovanni, Anelia Angelova, Michael S. Ryoo, Differentiable grammars for videos, in: Proceedings of AAAI, 2020, pp. 11874–11881.
[46] David B. Pillemer, One-versus two-tailed hypothesis tests in contemporary educational research, Educ. Res. 20 (9) (1991) 13–17.
[47] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, Vittorio Ferrari, Connecting vision and language with localized narratives, in: Proceedings of ECCV, 2020.
[48] Yikang Shen, Zhouhan Lin, Chin-Wei Huang, Aaron C. Courville, Neural language modeling by jointly learning syntax and lexicon, in: Proceedings of ICLR, 2018.
[49] Yikang Shen, Shawn Tan, Alessandro Sordoni, Aaron C. Courville, Ordered neurons: integrating tree structures into recurrent neural networks, in: Proceedings of ICLR, 2019.
[50] Haoyue Shi, Karen Livescu, Kevin Gimpel, On the role of supervision in unsupervised constituency parsing, in: Proceedings of EMNLP, 2020, pp. 7611–7621.
[51] Haoyue Shi, Jiayuan Mao, Kevin Gimpel, Karen Livescu, Visually grounded neural syntax acquisition, in: Proceedings of ACL, 2019, pp. 1842–1861.
[52] Haoyue Shi, Jiayuan Mao, Tete Xiao, Yuning Jiang, Jian Sun, Learning visually-grounded semantics from contrastive adversarial samples, in: Emily M. Bender, Leon Derczynski, Pierre Isabelle (Eds.), Proceedings of COLING, 2018, pp. 3715–3727.
[53] Valentin I. Spitkovsky, Hiyan Alshawi, Daniel Jurafsky, Christopher D. Manning, Viterbi training improves unsupervised dependency parsing, in: Proceedings of CoNLL, 2010, pp. 9–17.
[54] Robert P. Stockwell, The place of intonation in a generative grammar of English, Language 36 (3) (1960), 1960, pp. 360–367.
[55] David Talkin, W. Bastiaan Kleijn, A robust algorithm for pitch tracking (RAPT), in: Speech Coding and Synthesis, vol. 495, 1995, p. 518.
[56] Zheng-Hua Tan, Achintya Kumar Sarkar, Najim Dehak, rVAD: an unsupervised segment-based robust voice activity detection method, Comput. Speech Lang. 59 (2020) 1–21.
[57] Valerie L. Trollinger, Relationships between pitch-matching accuracy, speech fundamental frequency, speech range, age, and gender in American English-speaking preschool children, J. Res. Music Educ. 51 (1) (2003) 78–95.
[58] Bo Wan, Wenjuan Han, Zilong Zheng, Tinne Tuytelaars, Unsupervised vision-language grammar induction with shared structure modeling, in: Proceedings of ICLR, 2022.
[59] Pengyu Wang, Phil Blunsom, Collapsed variational Bayesian inference for PCFGs, in: Proceedings of CoNLL, 2013, pp. 173–182.
[60] Dani Yogatama, Phil Blunsom, Chris Dyer, Edward Grefenstette, Wang Ling, Learning to compose words into sentences with reinforcement learning, in: Proceedings of ICLR, 2017.
[61] Songyang Zhang, Linfeng Song, Lifeng Jin, Haitao Mi, Kun Xu, Dong Yu, Jiebo Luo, Learning a grammar inducer from massive uncurated instructional videos, in: Proceedings of EMNLP 2022, 2022, pp. 233–247.
[62] Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, Jiebo Luo, Video-aided unsupervised grammar induction, in: Proceedings of NAACL-HLT, 2021, pp. 1513–1524.
[63] Yanpeng Zhao, Ivan Titov, Visually grounded compound PCFGs, in: Proceedings of EMNLP, 2020, pp. 4369–4379.
[64] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, Jianfeng Gao, RegionCLIP: region-based language-image pretraining, in: Proceedings of CVPR, 2022, pp. 16772–16782.