# UNSUPERVISED HIERARCHICAL STRUCTURE INDUCTION FOR DEEPER SEMANTIC ANALYSIS OF AUDIO

*Sourish Chaudhuri, Bhiksha Raj*

Language Technologies Institute, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, PA - 15213.
{sourishc, bhiksha}@cs.cmu.edu

## ABSTRACT

Current audio analysis techniques rely on fairly shallow analysis of audio content, using symbols or patterns extracted directly from the observed acoustics. We hypothesize that the observed acoustics actually map to semantics in a hierarchical manner, and that the higher levels of this hierarchy correspond to increasingly higher-level semantics. In this paper, we present a model for deeper analysis of the observed acoustics, that induces a probabilistic tree structure depending on estimated constituent identities and contexts. Audio characterization using the deeper structure outperforms the standard shallow-feature based characterizations.

***Index Terms***— automatic content analysis, structure discovery, unsupervised learning, semantic audio

## 1. INTRODUCTION

The analysis of audio content is required for various common tasks– *e.g.* audio classification, retrieval, segmentation and recounting. Many different approaches have been formulated in the literature for these tasks. Typically, however, most of these approaches work directly off of the observed acoustics. In this paper, we posit that the audio content contains a wealth of information in its structure and sequence that would allow a task-agnostic automatic system to analyze it, and that information from these analyses would directly enable the wide range of tasks mentioned earlier.

While traditional audio content analysis has relied primarily on shallow analysis of the observed acoustics, based on detection or single-level latent variable models, in this paper, we present a hierarchical paradigm for content analysis that can be used for deeper analysis of the audio content. Figure 1 shows an instance of the kind of analysis that we believe can be extracted by our framework, using an example from baseball audio. The lowest level of this tree structure corresponds to low-level, generalized acoustic units, which may not carry discernible semantic information individually, but the sequences or distribution patterns of these units should capture higher-level semantic information– we refer to higher-level patterns as *events*. These *event* units themselves might contain certain patterns, corresponding to still more complex events. In most natural audio, these events themselves do not occur in isolation. They are related to each other in different ways, and event context provides cues for possible future events (event dependencies are indicated by arrows in the figure). Further, the event sequences themselves should carry information about the overall semantic content or class of the audio.

Such hierarchical analysis structures could not only be exploited for various audio analysis tasks, but also to develop a better understanding of relationships between events. The primary issue in estimating such structures for audio is a scarcity of richly annotated data with information at the various hierarchical levels that could be used to provide supervision.

To address this issue, this paper proposes an unsupervised model for structure induction that can leverage easily available, but unlabeled, data. Given simply an audio corpus, our model estimates corresponding, hierarchical tree structures.

Whether the structure induced by such unsupervised models would be consistently semantically coherent or human-interpretable is unknown, at this point in time. However, there are compelling motivations behind such approaches. The process of building richly annotated, hierarchically labeled data sets would be an expensive and time-consuming one, and the output of unsupervised approaches can be used both to obtain labels for and verify some coherent semantic units, as well as use them to seed semi-supervised approaches, thus building up labeled resources. Besides, while the presented approach makes no claims toward modeling human approaches to scene understanding, the extracted co-occurrence information and contextual cues provide a potential basis for comparisons to human reasoning processes in future work.

While an ideal evaluation framework would directly measure the generated structure by comparing it with ground truth structures (generated by annotators), such data is expensive to obtain. We propose, instead, to use the structured information generated by our models as features for characterization of audio for a retrieval task.

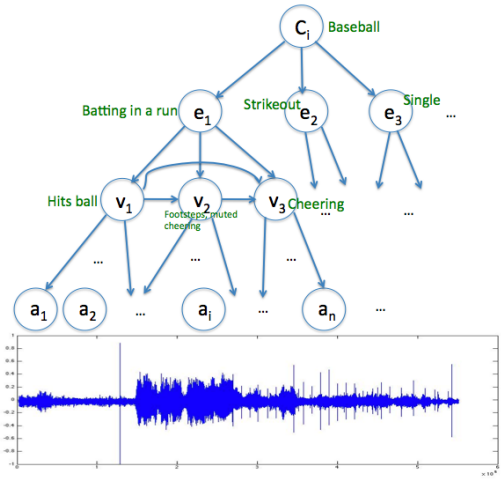


**Fig. 1**. An instance of hierarchical analysis for audio.

The rest of the paper is organized as follows: in Section 2, we review related prior work. We present our model for unsupervised structure induction in Section 3, and present our experiments with using the induced structure for an audio retrieval task in Section 4 before concluding in Section 5.

## 2. RELATED WORK

Audio categorization and retrieval systems largely rely on lexicon-based approaches, where the individual lexical units model sound types or sources. Several approaches have been successfully employed for lexicon learning including detector-based approaches using supervised data as well as unsupervised lexicon learning techniques [1, 2, 3, 4, 5]. These lexical units usually perform well at the task of capturing acoustically consistent phenomena, but cannot explicitly use the structure of the audio data to perform any further semantic analysis. For instance, acoustics similar to a dull metallic collision may be produced by very different semantic sources– a hammer striking an object, a baseball bat hitting a ball, or a car collision. The proposed hierarchical approach aims to automatically leverage different contextual cues to identify differing semantics at the appropriate level in the hierarchy.

Various models can be developed for the individual higher order layers for the framework in Figure 1, such as in [6]. Unlike [6], which estimates higher levels in the hierarchy one layer at a time, the method outlined in this paper estimates the entire tree structure jointly following approaches similar to those used in text parsing, as we shall describe. Distributional clustering approaches have been used for part-of-speech induction [7, 8] and have resulted in high-quality clusters, though the clusters resemblance to classical parts-of-speech varies substantially. Grammar induction approaches have dealt with attempting to uncover tree structures unsupervised, using the assumption that the tree was generated by a probabilistic, context-free grammar [9, 10], using assumptions of fixed structure, linguistic constraints or prior knowledge, but these approaches only had limited success. Subsequent models for generating tree structures, based off of the assumption that valid constituents in a tree should be non-crossing, met with much greater success [11, 12].

## 3. PROPOSED MODEL

Due to the similarity of the task we are attempting to solve with those from the parsing of natural language text, the primary model that we experiment with in this paper to generate the hierarchical structures of the type shown in Figure 1 follows the constituent-context model [11], which proposes a parametric family of models over trees. We begin by first estimating the lowest level acoustic units (indexed by $a$ in Figure 1), to convert the continuous audio sequence to a discrete representation. The process of inducing the higher-level structure works on top of this representation, and we show in our experiments in Section 4 that features derived from this structure improve over characterizations using simply the acoustic units.

Given the low-level acoustic units for an audio recording, the task of inducing a tree structure can be divided into two tasks– first, we need to decide constituent identity, *i.e.* which and how many of the consecutive low-level units should belong to the same higher-level constituent; and second, we need to decide the label for the constituent. While these tasks are correlated, the task of labeling the constituents is the easier of the two, since the distribution of the lower-level acoustic units within the constituents can be used to cluster the various constituents. The task of deciding constituent identity is significantly harder in our case, because the process of estimation the lower-level acoustic units is typically noisy; unlike text, where the observed surface forms typically correspond to ground truth, the observed audio can contain noise both in terms of innate variations in semantic content, as well as additive background noise that can cause errors in estimation of the lower-level units. As in [11], the proposed model relies on two assumptions: (i) constituents of a parse do not cross each other, and (ii) constituents occur in constituent contexts.

Let $\mathcal{A}$ be a sequence of the estimated lower-level acoustic units, such that for any given recording, $\mathcal{A} = a_1 a_2 ... a_{n_i}$. Every subsequence of $\mathcal{A}_i$ occurs in some linear context $c$, where $c(\mathcal{A}_j^k) = a_{j-1}\mathcal{A}_j^k a_{k+1}$, where the context elements correspond to the adjacent acoustic units for the subsequence. Then we can view any tree $t$ over a sequence $\mathcal{A}$ s as a collection of sequences and contexts. Good trees will include nodes whose yields [1] frequently occur as constituents and these constituents are frequently surrounded by expected contexts. To formally model this, we use a log-linear model with the form for the conditional distribution being as shown below:

$$P(t|\mathcal{A},\Theta) = \frac{\exp(\sum_{\{\mathcal{A}_j^k,c\}\in t} \lambda_{\mathcal{A}_j^k} f_{\mathcal{A}_j^k} + \lambda_c f_c)}{\sum_{t:yield(t):\mathcal{A}} \exp(\sum_{\{\mathcal{A}_j^k,c\}\in t} \lambda_{\mathcal{A}_j^k} f_{\mathcal{A}_j^k} + \lambda_c f_c)} \quad (1)$$

Thus, for each tree, we have one feature $f_{\mathcal{A}_j^k}$ for each constituent subsequence $\mathcal{A}_j^k$ in the tree, and its value is the number of nodes in $t$ with yield $\mathcal{A}_j^k$, and one feature $f_c$ for each context $c$ representing the number of times $c$ is the context of the yield of some node in the tree. Joint features over the context and the yield are not used, and no distinction is made between the constituent types at this point.

We model the conditional likelihood of a tree $t$ as $P(t|\mathcal{A},\Theta)$, where $\Theta = \{\lambda_{\mathcal{A}_j^k}, \lambda_c\}, \forall\{j,k\}$ that form constituent subsequences. We use an iterative EM-like procedure to find the best parameter estimate given the observed acoustic unit sequences for the given data. The parameter set $\Theta$ is initialized to zero and each audio recording is initialized to with a random tree structure for the observed acoustic unit sequence for each recording. In alternating steps, then, we find the the best parameter update $\Theta^*$ and the best guess for the tree structure for each of the acoustic unit sequences, given the updated parameters, using a dynamic program. For any $\Theta$, this produces the set of tree structures $T^*$ that maximizes $P(T|\{\mathcal{A}\},\Theta)$. Thus, $P(T^*|\{\mathcal{A}\},\Theta) \geq P(T|\{\mathcal{A}\},\Theta)$ (Here, $T$ refers to the set of most likely set of trees for the set of audio recordings $\{\mathcal{A}\}$). The iterative process then fixes these estimated tree structures to update the parameters. Given the choice of exponential family in Equation 1, we do not have a closed-form update rule for the parameters, and will need to adopt a numerical solution for updation, such as conjugate gradient.

Due to the varied linear contexts that can occur in the lower-level acoustic unit sequences, smoothing plays an important role in determining the quality of the induced tree structures. The current system can model arbitrarily long yields, which occur infrequently. The corresponding parameters for these yields may not significantly change from their initial choices, in spite of multiple learning iterations. Ideally, we would like the weights for unlikely occurrences to have very little influence, by making them as close to zero as possible, thus skewing the distribution of values in $\lambda_{\mathcal{A}_j^k}$ towards low values.

[1] The yield of any non-terminal node in a tree structure refers to the sequence of terminals produced by the subtree rooted at the non-terminal node.

In the conjugate gradient setting, parameter estimates are slow to converge and difficult to smooth with desired priors. Thus, we adopted a different approach that proved to work quite well using the simple smoothed relative frequency estimates, where $\lambda_k = \frac{count(f_k)}{count(k)+M}$. This estimation process ensures that the parameter values lie between 0 and 1, providing a bias toward non-constituency for long subsequences using high values for $M$.

Once the underlying *most-likely* tree structures have been computed for all the audio recordings given their representation as sequences of low-level acoustic units, we then move to the second stage of the process– that of labeling the induced constituents using a clustering technique. The only external input to this system at this stage is the hyperparameter $K$ for the number of clusters.

We performed clustering on the estimated constituent acoustic units using a modified $k$-means procedure, where we first selected a set of $K$ cluster centers. The distance of each induced constituent ($\mathcal{A}_j^k$) to each of these cluster centers ($\mathcal{C}_i$) were computed using a combination of 2 factors– temporal sequence of the constituent acoustic units, as well as distribution of the units in the constituent. The intuition behind using the temporal sequence is apparent, since we would expect similar higher order units to contain similarities in their constituent sequences. However, since the lower level acoustic units are not ground truth, but in fact estimates from noisy decodes, different manifestations of the same higher-order unit might be quite different due to insertions, deletions and substitutions in the true sequence. To account for this, we consider the distribution of the acoustic units as well. , where we estimate the $L2$ distance between the distributions of the constituent and cluster centers.

We use the Levenshtein edit-distance ($\mathcal{L}(\mathcal{A}_j^k, \mathcal{C}_i)$) for any constituent to each cluster center using the actual acoustic unit sequences that occurs in the constituent and the one for the cluster center to model temporal similarity. For each constituent subsequence, the distances to various centers are normalized by the maximum distance to lie between 0 and 1. To compute the distance between the distributions, we compute the cosine divergence between the 2 distributions ($Cos(\mathcal{A}_j^k, \mathcal{C}_i)$. The final computed distance is a product of the 2 individual distances as follows:

$$\mathcal{D}(\mathcal{A}_j^k, \mathcal{C}_i) = \mathcal{L}(\mathcal{A}_j^k, \mathcal{C}_i) \times (1 - Cos(\mathcal{A}_j^k, \mathcal{C}_i)) \quad (2)$$

The constituent $\mathcal{A}_j^k$ is assigned to the cluster center $\mathcal{C}^*$ chosen as:

$$\mathcal{C}^* = \arg\min_{\mathcal{C}_i} \mathcal{D}(\mathcal{A}_j^k, \mathcal{C}_i) \quad (3)$$

While the semantic import of the tree structures and the induced constituent labels cannot be understood directly from this process in the absence of extensive studies using humans in the loop to understand if the constituents consistently capture human-interpretable semantics, we hypothesize that these will provide positive improvements on semantically defined tasks, if they do indeed capture some underlying higher-level semantics. We present results with using characterizations derived from the tree structures on an audio retrieval task in Section 4.

## 4. EXPERIMENTAL RESULTS

Since we do not have labeled data that can be used to directly analyze the accuracy of the estimation process, we evaluate our framework on an audio retrieval task. In this section, we first describe the data and the task used in our experiments in Section 4.1. We discuss the systems that we compare and explain how they are used to obtain a characterization of the audio for this task in Section 4.2, and Section 4.3 describes the classifier we use. Section 4.4 discusses the results of our experiments.

### 4.1. Audio Retrieval Dataset and Task

For our experiments, we use the BBC Sound Effects Library CDs 1 – 20 consisting of 1120 different audio clips [13]. This library consists of various conceptual categories of sound, and audio tracks for the various categories contain complex audio due to the presence of many different sounds; *e.g.* a supermarket audio contains voices, sound from the checkout bell, trolleys and baskets being stacked. Thus, these categories are defined at a higher semantic level than datasets that contain instances of simpler sounds, such as gunshots, laughter, etc. The BBC Sound Library recordings are of a high and consistent quality, and allow us to compare compare different systems in a setting where additional confounding factors are not present, as is often the case in Youtube-style, user-generated content where different recording conditions and equipment introduce channel variance.

The entire dataset was sampled at 16KHz, and 13-dimensional Mel-Frequency Cepstral Coefficients (MFCC) were extracted from the data, and this MFCC representation for the audio was used in all the experiments reported in this paper. While there are a number of categories in this dataset, we only use those that have at least 15 positive instances belonging to the category. Thus, we have the following 10 categories– *Exterior atmospheres*, *Household*, *Interior Backgrounds*, *Transport*, *Animals*, *Audiences*, *Electronic Equipment*, *Water*, *Birds*, *Warfare*. All the other files have a negative label for each of the 10 categories.

The audio retrieval task is defined as follows: given one of the 10 categories as input, the task is to retrieve all audio files belonging to that category from the test collection. We compute *Missed Detection* (MD) and *False Alarm* (FA) rates as follows: suppose there are $N_t$ test files, with $C_i$ belonging to class $i$, and the detector predicts $N_i$ as belonging to class $i$, and $D_i$ of these were correct. Then:

$$MD = \frac{C_i - D_i}{C_i}; FA = \frac{N_i - D_i}{N_t - C_i} \quad (4)$$

We report results using the average Area Under MD-v/s-FA Curves (AUC) for the 10 categories using 5-fold cross validation on the entire data. Since the curve measures error of the system being evaluated, the lower the area under the curve, the better the performance.

### 4.2. Systems Used for Retrieval

Our hypothesis in testing the induced tree structures on an audio retrieval task was that the induced structures over the lower-level units should improve over the performance of retrieval systems based on the lower-level units alone.

We set up 2 baselines using 2 different lower-level unit estimation schemes. The first baseline uses a Vector-Quantization approach to quantize each audio frame into one of several clusters (we refer to this system as **VQ**). The second uses an HMM-based lexicon learning scheme outlined in [1] to represent each audio file as a sequence of acoustic units descriptors, where each such descriptor can span multiple frames (we refer to this system as **AUDs**). Each audio file

is represented using a feature set of dimensionality equal to the number of units in the system, with the feature value for each unit being its relative frequency of occurrence in the file.

We then induce the tree structures over the sequences of acoustic units produced by both the VQ and the AUD systems, and then use the subsequent clustering to generate identifiers for the various constituents. The audio files can be characterized using the relative frequencies of these constituents (we refer to these systems as **VQ-Trees** and **AUD-Trees** respectively). Finally, we can create an additional pair of systems that combines the lower-level units with the structure induction process, by concatenating the pair of feature vectors (we refer to these as **VQ-Comb** and **AUD-Comb** respectively).

### 4.3. Random Forest Classifier

The audio retrieval requires us to predict whether each audio file belongs to a particular class or not. Hence, we train binary classifiers for each of the 10 audio categories to predict whether a test file belongs to the class or not (one-versus-all). The experiments reported employ a Random Forest [14] classifier for each category. While any classifier could have been used for this task, we chose random forest classifiers as they are resistant to overfitting. Random forests are an extension of decision tree classification techniques, where the training process grows many trees instead of a single one, using held out data is used to get an estimate of the error as trees are added to the forest. The trees in the forest are grown as far as possible, and pruning is not used. Given a new test file, each of the trees in the forest returns a class label, which is used in a weighted vote to determine the final predicted label. In our experiments, we use 500 trees. For details of the training process, the reader is referred to [14].

### 4.4. Experimental Results

A comparison of performances using the different systems is shown in Table 1. We obtained the best results when using 64 cluster centers to cluster the constituent blocks of audio units. While we provide tree-only results for our experiments, we note that retrieval based on features from the induced tree alone is not expected to be better for 2 reasons. First, tree structures induced in an unsupervised manner will contain considerable noise. Secondly, and possibly more importantly, even if ground truth labels were available, the higher a node is in the tree hierarchy, the longer is its yield, and information available for such nodes in the training set decreases. The higher level clusters are likely to be broader since we represent the wide range of possibilities by a mapping onto a limited, finite set of clusters, thus collapsing a varied set of concepts together and reducing the discriminative properties of the true higher-level concept. Nonetheless, we do expect that they would provide *additional* information, and the results are consistent with this expectation. While the systems using the induced tree structure information only do not outperform the systems using lower level units only, the combination of the two systems outperforms both the individual systems significantly.

We expect that the primary reason that the tree structure based systems have limited success is due to the fact that they work with the information provided by the lower level units and any error in the estimation of those units is propagated, resulting in the semantics captured being weaker than in a model that can jointly utilize both the observed audio and the estimated units jointly to induce structure. Developing such models remains a focus of our future work.

The fact that the combined systems outperform the individual unit-based or tree-based systems (with consistent trends for both baseline systems) is promising, and shows that the induced structure does capture additional semantics.

| System | Avg. AUC |
|---|---|
| VQ | 0.214 |
| VQ-Trees | 0.246 |
| VQ-Comb | 0.194 |
| AUD | 0.174 |
| AUD-Trees | 0.181 |
| AUD-Comb | 0.169 |

**Table 1**. Comparison of the various systems on average AUC (lower is better)

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a novel unsupervised approach to inducing tree structures for modeling higher-level semantic information that can be applied for different tasks. We presented a unified framework that hypothesizes that the observed acoustics map hierarchically to higher-level semantics, and that estimation of these semantics directly from the audio in a task-agnostic manner could be used to derive characterizations that could be appropriately utilized for the specific task at hand.

We leveraged previous work in unsupervised text parsing as well as acoustic unit estimation to generate hierarchical structures for audio in an unsupervised setting. Presently, the semantic import of the derived structures is unclear, since we do not have labeled data for analysis of the estimated constituent boundaries or parse structures. We expect to address this in future work by obtaining human judgments for the induced structures. Annotations, thus obtained, can then be used for comparison with other techniques such as the one presented in [6] to compare the effects of methods that model higher layers one layer at a time against methods that model the entire hierarchy jointly as presented in this paper. We can further leverage the automatically induced structures to obtain richer semantic annotations for audio datasets in a less expensive manner, so that supervised or weakly-supervised methods become feasible in the future.

## 6. REFERENCES


[1] S. Chaudhuri, M. Harvilla, and B. Raj, "Unsupervised learning of acoustic unit descriptors for audio content representation and classification," in *Interspeech*, 2011, pp. 717–720.

[2] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Interspeech*, 2011.

[3] S.F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. Loui, and J. Luo, "Large-scale multimodal semantic concept detection for consumer video," in *MIR workshop, ACM-Multimedia*, 2007.

[4] M. Slaney, "Mixture of probability experts for audio retrieval and indexing," in *ICME*, 2002.

[5] X. Zhuang, S. Tsakalidis, S. Wu, P. Natarajan, R. Prasad, and P. Natarajan, "Compact audio representation for event detection in consumer media," in *Interspeech*, 2011.

[6] S. Chaudhuri and B. Raj, "Unsupervised structure discovery for semantic analysis of audio," in *Neural Information Processing Systems*, 2012.

[7] S. Finch and N. Chater, "Distributional bootstrapping: From word class to proto-sentence," in *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, 1994.

[8] H. Schutze, "Distributional part-of-speech tagging," in *Proceedings of the European Association for Computational Linguistics*, 1995.

[9] G. Carroll and E. Charniak, "Two experiments on learning probabilistic dependency grammars from corpora," *Working Notes of the Workshop Statistically-Based NLP Techniques*, pp. 1–13, 1992.

[10] K. Lari and S. J. Young, "The estimation of stochastic context-free grammars using the inside-outside algorithm," *Computer Speech and Language*, pp. 35–56, 1990.

[11] D. Klein and C. Manning, "Natural language grammar induction using a constituent-context model," in *Advances in Neural Information Processing Systems*, 2001.

[12] D. Klein and C. Manning, "A generative constituentcontext model for improved grammar induction," in *Proceedings of the Association for Computational Linguistics*, 2002.

[13] "Bbc sound effects library original series, http://www.sound-ideas.com/bbc.html," .

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.