Combining Task-Dependent Information with Auditory Attention Cues for Prominence Detection in Speech

Ozlem Kalinli and Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
Department of Electrical Engineering-Systems
University of Southern California, Los Angeles, CA, USA
kalinli@usc.edu, shri@sipi.usc.edu

Abstract

Auditory attention is a highly complex mechanism that involves the process of low-level acoustic features of sound together with higher level cognitive rules. In this paper, a novel method that combines biologically inspired auditory attention cues with higher level lexical and syntactic information is proposed to model task-dependent influences on a given task. The feature maps are extracted from sound at multi-scales by mimicking the processing stages in the human auditory system, and converted to low-level auditory gist features. Then, the auditory attention model biases the gist features based on the task to maximize target detection. The top-down task-dependent influence of lexical and syntactic information is incorporated into the model using a probabilistic approach. The combined model is tested to detect prominent syllables in speech using the BU Radio News Corpus. The model achieves 88% prominence detection accuracy at syllable level, which is comparable to reported human performance on this task.

Index Terms: auditory attention, auditory gist, stress, accent, prominence, syntax, lexical rules.

1. Introduction

Attention allows primates to efficiently allocate neural resources to locations of interest in order to precisely interpret a scene or to search for a target. There are two types of attentional mechanisms: rapid saliency-driven (*task-independent*) *bottom-up* attention and slower *task-dependent top-down* attention [1]. First, rapid bottom-up processing of the whole scene occurs that attracts attention towards conspicuous or salient locations in an unconscious manner. Then, the top-down processing shifts the attention voluntarily towards locations of cognitive interest. Only the selectively attended incoming stimuli is allowed to progress through cortical hierarchy for high-level processing to recognize and further analyze the details of the stimuli [1, 2].

Bottom-up attention is a rapid, saliency-driven mechanism, and it detects the objects that perceptually "pop-out" of a scene by significantly differing from their neighbors. For example, a red flower among green leaves of a plant will be salient, so will gun shots or sudden explosions in a street. The top-down task-dependent process uses prior knowledge and learned past expertise to focus attention on the target locations in a scene. In [3], it was shown that gaze patterns depend on the task performed while viewing the same scene. The gaze of an observer fell on faces when estimating the people's age, but fell on clothing when estimating the people's material conditions. Similarly, it is the selective attention that allows a listener to extract one stream of speech in the presence of others by focusing on variety of acoustic cues such as pitch, timbre, spatial location [4].

There has been extensive research to explore the influence of attention on the neural responses in sensory systems. It has been revealed that the top-down task-dependent attention modulate the neuron responses in the visual and auditory cortex [5, 6]. This modulation mostly occurs by enhancing the response of neurons tuned to the features of target stimulus, whereas attenuating the response of neurons to stimuli that did not match the target feature. In [7], a guided visual search model was proposed to emulate this modulation effect of task on neuron responses. The model combines the weighted feature maps in a top-down manner, i.e., when the task is to detect a red bar, the feature maps which are sensitive to red gain a larger weight.

In speech, while processing the sound stimuli, the brain is influenced by higher level information such as lexical information, syntax, semantics, and the discourse context [8, 9]. For example, the recorded neurophysiological brain response was larger for one's native language than unfamiliar sounds in the experiments in [8]. Also, the experiments at the level of meaningful language units has revealed that words elicit a larger brain response than meaningless pseudo-words. The psychophysical experiments indicate that some words that carry semantically important information, i.e., one's name, can capture attention, and so can some syllable/word strings that make sense [4, 9].

In our previous work [10], we proposed a task-independent bottom-up auditory attention model. The model computes an *auditory saliency map* that encodes perceptual influence of each part of sound spectrum. It was demonstrated with the experiments that the model could detect prominent syllables in speech in an unsupervised manner. The motivation of the present paper is to analyze the effect of task-dependent influences on auditory attention, and on the task performance. In other words, when the subjects are asked to find the prominent (stressed) syllable, they use their prior task-relevant knowledge, i.e., prominent syllables have longer duration. Specifically, the focus is to analyze the effect of the task-dependent influences captured via syntactic and lexical cues working in conjunction with an auditory attention model in the context of prominence detection.

The feature maps are extracted from sound by using the front-end of the bottom-up auditory attention model proposed by us in [10], which mimics the various processing stages in the human auditory system (HAS). First, an auditory spectrum of the sound is computed based on early stages of HAS. Then, feature maps are extracted from the spectrum in multi-scales based on the processing stages in the central auditory system, and converted to low-level *auditory gist* features. The attention model biases the auditory gist features to imitate the modulation ef-

fect of task on neuron responses using the weights learned for a given task. Next, the top-down influence of lexical and syntactic information is incorporated into the model using a probabilistic approach. The lexical information is integrated into the system by using a probabilistic language model. The syntactic knowledge is represented using the part-of-speech (POS) tags, and a neural network is used to model influence of syntax on prominence. The combined model is used to detect prominent syllables in experiments conducted on the Boston University Radio News Corpus (BU-RNC) [11], and achieves 88% accuracy, providing approximately a 12% absolute improvement over using just the bottom-up attention model.

The paper is organized as follows: the database used for the experiments is discussed in Section 2. The auditory attention model with acoustic cues combined with top-down high level information is presented in Section 3, and followed by results in Section 4. The conclusion is presented in Section 5.

2. Database

The BU-RNC is used for the experiments in this work to test the proposed task-dependent attention model. The BU-RNC is a broadcast news-style read speech corpus that consists of speech from 3 female and 3 male speakers, totaling about 3 hours of data. A significant portion of the data has been manually labelled with prosodic tags. We mapped all pitch accent types (H*, L*, L*+H, etc..) to a single stress label, reducing the task to a two-class problem. Hence, the syllables annotated with any type of pitch accent were labelled "prominent", and otherwise "non-prominent". The database also contains the orthography corresponding to each spoken utterance together with time alignment information, and part-of-speech (POS) tags for the orthographic transcriptions. The database consists of approximately 49,000 syllables, and the prominent syllable fraction is 34.3% (chance level). We chose this database for two main reasons: i) syllables are stress labelled based on human perception ii) since it carries labelled data, it helps us to learn the weights of the task-dependent influences in a supervised fashion.

3. Top-Down Task Dependent Model

The task-dependent model has two types of evidence: 1) acoustical evidence captured with the auditory gist features 2) higher level top-down information captured with lexical and syntactic models. These two separate streams of evidence are combined using a probabilistic approach for the prominence detection task. Next, we discuss modelling of each cue in details.

3.1. Acoustic Cues

The bio-inspired auditory gist feature extraction mimics the processing stages in the early and central auditory systems as illustrated in Fig. 1. First, the auditory spectrum of the sound is estimated using an early auditory (EA) system model. The EA model used here consists of cochlear filtering, inner hair cell, and lateral inhibitory stages mimicking the process from basilar membrane to the cochlear nucleus in the auditory system [12]. The cochlear filtering is implemented using a bank of 128 overlapping constant-Q asymmetric band-pass filters. For analysis, audio frames of 20 milliseconds (ms) with 10 ms shift are used.

The output of the EA model is a two-dimensional (2D) auditory spectrum with time and frequency axes, and here it is referred as *"scene"*. The scene is analyzed by extracting a set of multi-scale features which consist of *intensity (I), frequency contrast (F), temporal contrast (T)* and *orientation (O)* feature channels. They are extracted using 2D spectro-temporal receptive filters mimicking the analysis stages in the primary auditory cortex. Each of the receptive filters (RF) simulated for feature extraction is illustrated with gray scaled images in Fig. 1 next to its corresponding feature channel. The excitation phase, and inhibition phase are shown with white and black color, respectively. For example, the frequency contrast filter corresponds to RF with an excitatory phase and simultaneous symmetric inhibitory side bands. The RF for generating frequency contrast, temporal contrast and orientation features are implemented using 2D Gabor filters with angles (θ) 0^o, 90^o, $\{45^o, 135^o\}$, respectively. The RF for intensity feature is implemented using a 2D Gaussian kernel. The multi-scale features are obtained using a dyadic pyramid: the input spectrum is filtered, and decimated by a factor of two, and this is repeated. Finally, eight scales are created (if the scene duration D is large enough; otherwise there are fewer number of scales). The details of feature extraction and filters can be found in [10].

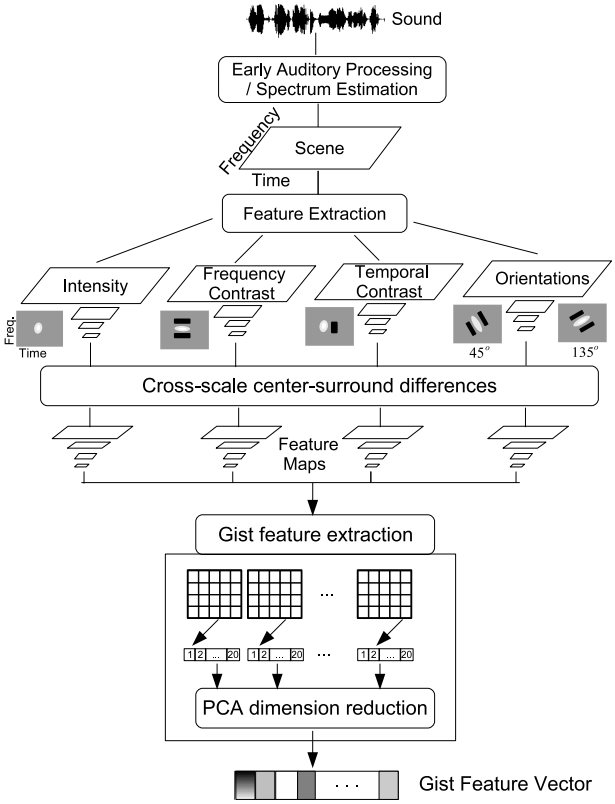

Figure 1: Auditory gist feature extraction [16]

After extracting features at multiple scales, the model computes "center-surround" differences akin to the properties of local cortical inhibition. It is simulated by across scale subtraction ($\ominus$) between a "center" fine scale c and a "surround" coarser scale s yielding a feature map $\mathcal{M}(c,s)$:

$$\mathcal{M}(c,s) = |\mathcal{M}(c) \ominus \mathcal{M}(s)|, \;\; \mathcal{M}\epsilon\{I,F,T,O_\theta\} \qquad (1)$$

The across scale subtraction between two scales is computed by interpolation to the finer scale and pointwise subtraction. Here, $c = \{2,3,4\}$, $s = c + \delta$ with $\delta\epsilon\{3,4\}$ are used. Next, the gist of the scene is extracted from the feature maps.

The gist of a scene is captured by humans rapidly, and it describes the overall properties of the scene [13]. Gist processing

guides attention to focus into a subset of stimuli to analyze the details of target locations. A gist vector is extracted from the feature maps of I, F, T, O_θ such that it covers the whole scene at low resolution. A feature map is divided into m by n grid of subregions, and the mean of each subregion is computed to capture rough information about the region, which results in a gist vector with length $m \times n$. For a feature map $\mathcal{M}_i$ with height h and width v, the computation of gist features can be written as:

$$G_i^{k,l} = \frac{mn}{vh} \sum_{u=\frac{kv}{n}}^{\frac{(k+1)v}{n}-1} \sum_{z=\frac{lh}{m}}^{\frac{(l+1)h}{m}-1} \mathcal{M}_i(u,z), \text{ for} \quad (2)$$
$$k = \{0, \cdots, n-1\}, l = \{0, \cdots, m-1\}.$$

Averaging operation is the simplest neuron computation, and the use of other second-order statistics such as variance did not provide any appreciable benefit for our application. An example of gist feature extraction with $m = 4, n = 5$ is shown in Fig. 1. After extracting a gist vector from each feature map, we obtain the cumulative gist vector by combining them. Then, principal component analysis (PCA) is used to reduce the dimension to make the subsequent machine learning more practical.

Here, the term *"scene"* is used to refer to the sound around a syllable, and the task is to determine whether a prominent syllable exists in the scene. For the experiments, a scene is generated for each syllable in the database by extracting the sound around it with an analysis window of duration D that centers on the syllable. As stated earlier, the top-down task-dependent model biases neuron responses to perform a given task, and here the weights to combine the features are learned in a supervised manner using a 3-layer neural network. We use auditory gist features as the input of the neural network, and the output returns the class posterior probability $p(c_i|f_i)$ for the i^{th} syllable, where f_i auditory gist feature, and $c_i \epsilon \{0, 1\}$ where 1 denotes that the syllable is prominent, while 0 denotes that it is non-prominent. Then, the most likely prominence sequence $\mathbf{C}^* = \{c_1, c_2, \ldots, c_M\}$ given the gist features $\mathbf{F} = \{f_1, f_2, \ldots, f_M\}$ can be found using a maximum a-posteriori (MAP) framework as follows:

$$\mathbf{C}^* = \underset{\mathbf{C}}{\text{argmax}}\ p(\mathbf{C}|\mathbf{F}) \quad (3)$$
$$= \underset{\mathbf{C}}{\text{argmax}} \prod_{i=1}^{M} p(c_i|f_i) \quad , \quad (4)$$

where syllable prominence classes are assumed independent.

3.2. Lexical Cues

The lexical information is included in the top-down model using a probabilistic language model. Syllable tokens and prominence classes of syllables are used together to build the language model. Given only the lexical information of syllable sequence $\mathbf{S} = \{s_1, s_2, \ldots, s_M\}$, the most likely prominence sequence $\mathbf{C}^*$ can be found as:

$$\mathbf{C}^* = \underset{\mathbf{C}}{\text{argmax}}\ p(\mathbf{C}|\mathbf{S}) \quad (5)$$

Here, $p(\mathbf{C}|\mathbf{S})$ is modelled within a bounded n-gram context as:

$$p(\mathbf{C}|\mathbf{S}) = \underset{\mathbf{C}}{\text{argmax}} \prod_{i=1}^{M} p(c_i|c_{i-n+1}^{i-1}, s_{i-n+1}^{i}) \quad (6)$$

It is difficult to robustly estimate the language model even within n-gram context due to the size of the database. Hence, a factored language model is built to overcome data sparsity using the SRILM toolkit [14]. We use a back-off path such that we first drop the most distant syllable variable s_{i-n+1} from the set on the right of the conditioning bar, then the most distant class variable c_{i-n+1}, so on so forth. We use a 4-gram model for the prominence class history, and a trigram model for the syllable tokens history. This n-gram order selection is also validated with the experiments.

3.3. Syntactic Cues

In our model, the syntactic information is represented using the POS tags provided with the database. POS tags are associated with words, so the most likely prominence sequence $\mathbf{C}^*$ for a word string $\mathbf{W}$ given only syntactic information can be computed as:

$$\mathbf{C}^* = \underset{\mathbf{C}}{\text{argmax}}\ p(\mathbf{C}|\mathbf{POS}(\mathbf{W})) \quad (7)$$

Assuming word tokens are independent, Eq. 7 can be written as

$$\mathbf{C}^* = \underset{\mathbf{C}}{\text{argmax}} \prod_{i} p(c_i|\mathbf{POS}_\mathbf{i}^\mathbf{L}(w_i)) \quad (8)$$

In Eq. 8, c_i represent the prominence of the i^{th} word w_i, $\mathbf{POS}_\mathbf{i}^\mathbf{L}$ is the POS tags sequence that is neighboring w_i :

$$\mathbf{POS}_\mathbf{i}^\mathbf{L} = (POS_i^{i-(L-1)/2}, \ldots, POS_i^i, \ldots , POS_i^{i+(L-1)/2}) \quad (9)$$

$\mathbf{POS}_\mathbf{i}^\mathbf{L}$ is chosen such that it contains syntactic information from a fixed window of L words centered at the i^{th} word, and $L = 5$ performs well for prominence detection [15]. In our implementation, a set of 34 POS tags are used as those used in the BU-RNC database. Each POS feature is mapped into a 34 dimensional binary vector and a 3-layer neural network with 34×5 inputs ($L = 5$) and two outputs (since this is a two-class problem) is used to compute the class posterior probability $p(c_i|\mathbf{POS}_\mathbf{i}^\mathbf{L}(w_i))$ in Eq. 8.

The syntactic model is built at the word level since POS tags are associated with words. Let us assume that the i^{th} word w_i consists of n_i syllables. Then, w_i is non-prominent *if and only if* all the syllables within w_i are non-prominent. Hence,

$$p(c_i = 0|\mathbf{POS}(w_i)) = \prod_{k=1}^{n_i} p(c_k = 0|\mathbf{POS}(s_k)) \quad (10)$$

where $p(c_i = 0|\mathbf{POS}(w_i))$ is the probability of w_i being non-prominent given the POS tags, and $p(c_k = 0|\mathbf{POS}(s_k))$ denotes the probability of the syllable s_k within the word w_i being non-prominent given the POS tags. We approximate the posterior probability of a syllable in a word being non-prominent as:

$$p(c_k = 0|\mathbf{POS}(s_k)) \approx \sqrt[n_i]{p(c_i = 0|\mathbf{POS}(w_i))}. \quad (11)$$

Then, the probability of the syllable s_k being prominent can be computed as:

$$p(c_k = 1|\mathbf{POS}(s_k)) = 1 - p(c_k = 0|\mathbf{POS}(s_k)) \quad (12)$$

In practice, to bring word level syntactic information to syllable level, Eq. 11 and 12 are used for the experiments.

3.4. Combining Acoustical and Higher Level Cues

Given auditory gist features $\mathbf{F}$, lexical evidence $\mathbf{S}$, and syntactic evidence $\mathbf{POS}$, the most likely prominence sequence can be found using a MAP framework as follows:

$$\mathbf{C}^* = \underset{\mathbf{C}}{\text{argmax}}\ p(\mathbf{C}|\mathbf{F}, \mathbf{S}, \mathbf{POS})$$
$$= \underset{\mathbf{C}}{\text{argmax}}\ p(\mathbf{F}, \mathbf{S}, \mathbf{POS}|\mathbf{C})p(\mathbf{C}) \quad (13)$$

The joint distribution cannot be robustly estimated since the vocabulary size is very large compared to the training data, so a naïve Bayesian approximation is used to simplify Eq 13 as:

$$\begin{aligned} \mathbf{C}^* &= \underset{\mathbf{C}}{\operatorname{argmax}}\ p(\mathbf{F}|\mathbf{C})p(\mathbf{S}|\mathbf{C})p(\mathbf{POS}|\mathbf{C})p(\mathbf{C}) \\ &= \underset{\mathbf{C}}{\operatorname{argmax}}\ \frac{p(\mathbf{C}|\mathbf{F})}{p(\mathbf{C})}p(\mathbf{C}|\mathbf{S})\frac{p(\mathbf{C}|\mathbf{POS})}{p(\mathbf{C})} \end{aligned} \quad (14)$$

The combined top-down model which includes auditory gist features, lexical and syntactic information, reduces to the product of individual probabilistic model outputs.

4. Experiments and Results

The BU-RNC data corpus is split into training and test sets to carry the experiments on prominence detection in speech. All of the experimental results presented here are estimated using the average of five-fold cross validation, i.e., in each set 80% of the data is used for training and the remaining 20% of the data is used for testing. The average out-of-vocabulary (OOV) syllables in the test sets was 12.7% relative to the test vocabulary. When, we compare the performance of different information streams, the Wilcoxon signed rank test is used to report the confidence level in terms of significance values ($p-$values).

A detailed discussion about experiments with auditory gist features, scene duration and grid resolution can be found in [16]. The mean syllable duration in BU-RNC is $\approx 0.2\ s$. In [16], it was found that the prominence of syllables is affected by the neighboring syllables and based on those experiments, here, the scene duration is set to $D = 0.6s$ considering both prominence detection performance and computational cost (since a larger scene duration yields a larger dimensional gist vector). The grid size of $(m, n) = (1, v)$ was sufficient for this task, where v is the width of the feature map. In other words, we had full temporal resolution and minimal frequency resolution. This results in a 48 dimensional auditory gist feature vector after PCA. Thus, the 3-layer neural network used together with acoustic features has 48 inputs and two-outputs, and it is designed such that the outputs return the posterior probability.

The results are presented in Table 1. Using only the auditory gist features, a 84.46% accuracy with an F-score of 0.77 is achieved for the prominence detection task at the syllable level. The performance obtained with using only lexical information is 83.35% (F-score=0.76). We achieve 82.50% accuracy (F-score=0.84) for the prominence detection task at the word level using only syntactic information. The best performance of 87.95% accuracy (F-score=0.82) is achieved when acoustic, lexical, and syntactic cues are combined all together, and this significantly outperforms all the individual feature performances at $p \leq 0.001$. The combined top-down model provides approximately 12% absolute improvement over the bottom-up model (BU in Table 1), and the results compare well against the previously reported performance levels with the BU-RNC database, e.g. a supervised model obtained 74.1% accuracy using only acoustical features, and 86.1% accuracy using acoustical, lexical and syntactic features in [17].

Table 1: Prominent Syllable Detection Performance

TD Cues	Acc.	Pr.	Re.	F-sc.
Auditory Feat. only	84.46%	0.80	0.73	0.77
Lexical only	83.85%	0.77	0.76	0.76
Syntactic only	82.50%	0.82	0.87	0.84
Combined	**87.95%**	**0.83**	**0.82**	**0.82**
BU	75.9%	0.64	0.79	0.71

5. Conclusions

In this paper, a model that combines auditory attention model with task-dependent higher level lexical and syntactic evidence is presented. The auditory gist features are extracted from sound by mimicking the analysis stages in the HAS, and biased with weights learned in a supervised manner to incorporate the task-dependent influences on the neural responses. The lexical information is modelled by building and using a language model with syllable tokens, and syntactic information is captured using POS tags within a neural network. Finally, the information from different cues is combined within a probabilistic framework. The model was demonstrated to successfully detect prominent syllables in read speech with 88% accuracy. The results compare well to the human performance on stress labelling reported with BU-RNC: the average inter-transcriber agreement for manual annotators was 85-90% for presence vs. absence of stress labelling [11].

The model proposed here is not only limited to prominence detection. For example, the top-down model with auditory gist features can be used in other spoken language processing tasks and general computational auditory scene analysis (CASA) applications to classify ambient scenes.

6. Acknowledgements

Research was support in part with funds from DARPA, ONR and Army.

7. References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] C. Alain and S. R. Arnott, "Selectively attending to auditory objects," *Front. Biosci.*, vol. 5, pp. d202–212, 2000.

[3] A. Yarbus, *Eye movements during perception of complex objects*. New York, NY: Plenum Press, 1967.

[4] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *The Journal of the Acoustical Society of America*, vol. 25, p. 975, 1953.

[5] D. Hubel, C. Henson, A. Rupert, and R. Galambos, "Attention Units in the Auditory Cortex," *Science*, vol. 129, no. 3358, pp. 1279–1280, 1959.

[6] R. Desimone and J. Duncan, "Neural Mechanisms of Selective Visual Attention," *Annual Review of Neuroscience*, vol. 18, no. 1, pp. 193–222, 1995.

[7] J. Wolfe, "Guided Search 2.0: A revised model of guided search," *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.

[8] F. Pulvermüller and Y. Shtyrov, "Language outside the focus of attention: The mismatch negativity as a tool for studying higher cognitive processes," *Progress in Neurobiology*, vol. 79, no. 1, pp. 49–71, 2006.

[9] A. Johnson and R. W. Proctor, *Attention: Theory and Practice*. Sage Publications, 2004.

[10] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Proc. of Interspeech*, August 2007.

[11] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, *The Boston University Radio Corpus*, 1995.

[12] S. Shamma, "On the role of space and time in auditory processing," *Trends Cogn. Sci*, vol. 5, pp. 340–8, 2001.

[13] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, pp. 300–312, 2007.

[14] K. Kirchhoff, J. Bilmes, and K. Duh, "Factored language models tutorial," Dept. of EE, U. Washington, Tech. Rep. UWEETR–2007–0003, June 2007.

[15] K. Chen, M. Hasegawa-Johnson, A. Cohen, and J. Cole, "A maximum likelihood prosody recognizer," in *Proc. of Speech Prosody*, Nara, Japan, 2004.

[16] O. Kalinli and S. Narayanan, "A top-down auditory attention model for learning task dependent influences on prominence detection in speech," in *Proc. of ICASSP*, April 2008.

[17] S. Ananthakrishnan and S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, 2008.