Prominence Detection Using Auditory Attention Cues and Task-Dependent High Level Information

Ozlem Kalinli, *Student Member, IEEE*, and Shrikanth Narayanan, *Fellow, IEEE*

***Abstract*—Auditory attention is a complex mechanism that involves the processing of low-level acoustic cues together with higher level cognitive cues. In this paper, a novel method is proposed that combines biologically inspired auditory attention cues with higher level lexical and syntactic information to model task-dependent influences on a given spoken language processing task. A set of low-level multiscale features (intensity, frequency contrast, temporal contrast, orientation, and pitch) is extracted in parallel from the auditory spectrum of the sound based on the processing stages in the central auditory system to create feature maps that are converted to auditory gist features that capture the essence of a sound scene. The auditory attention model biases the gist features in a task-dependent way to maximize target detection in a given scene. Furthermore, the top-down task-dependent influence of lexical and syntactic information is incorporated into the model using a probabilistic approach. The lexical information is incorporated by using a probabilistic language model, and the syntactic knowledge is modeled using part-of-speech (POS) tags. The combined model is tested on automatically detecting prominent syllables in speech using the BU Radio News Corpus. The model achieves 88.33% prominence detection accuracy at the syllable level and 85.71% accuracy at the word level. These results compare well with reported human performance on this task.**

***Index Terms*—Accent, auditory attention, auditory gist, lexical rules, prominence, stress, syntax, task-dependent.**

I. Introduction

HUMANS can precisely process and interpret complex scenes in real time, despite the tremendous number of stimuli impinging the senses and the limited resources of the nervous system. One of the key enablers of this capability is believed to be a neural mechanism, called "*attention*." Even though the term "attention" is commonly used in daily life, researchers in sensory processing and systems have not reached an agreement on the structure of the attention process. Attention can be viewed as a gating mechanism that allows only a small part of the incoming sensory signals to reach memory and awareness [1]. It can also be viewed as a spotlight that is directed towards a target of interest in a scene to enhance the processing in the area while ignoring the stimuli that fall outside of the spotlighted area [2], [3]. It has been suggested that attention can be oriented by a bottom-up saliency-driven mechanism and a top-down task-dependent mechanism [4]–[6]. It is assumed that the bottom-up process is based on scene-dependent features and may attract attention towards conspicuous or salient locations of a scene in an unconscious manner, whereas the top-down process shifts attention voluntarily toward locations of cognitive interest in a task-dependent manner [4]–[6]. Further, psychophysical studies suggest that only the selectively attended incoming stimuli are allowed to progress through the cortical hierarchy for high-level processing to recognize and further analyze the details of the stimuli [7], [8], [3].

Bottom-up attention is considered to be a rapid, saliency-driven mechanism, which detects the objects that perceptually pop out of a scene by significantly differing from their neighbors. For example, in vision, for an observer, a red flower among green leaves of a plant will be salient. Similarly, in audition, the sound of a gunshot in a street will perceptually stand out of the traffic/babble noise of the street. The experiments in [9] have indicated that bottom-up attention acts early, and top-down attention takes control within an order of 100 ms. The top-down task-dependent (goal-driven) process is considered to use prior knowledge and learned past expertise to focus attention on the target locations in a scene. For example, in vision, it was shown that gaze patterns depend on the task performed while viewing a scene [10]. The gaze of the observer fell on faces when estimating the people's ages, but fell on clothing when estimating the people's material conditions. For example, in audition, it is the selective attention that allows a listener to extract a particular person's speech in the presence of others (the cocktail party phenomenon) by focusing on a variety of acoustic cues such as pitch, timbre, and spatial location [11], [12]. The bottom-up attention mechanism may play a vital role for primates by making them quickly aware of possible dangers around them, whereas the top-down attention mechanism may play a key role for extracting the signal of interest from cluttered and noisy backgrounds. However, in the current paper, the proposed top-down task-dependent model is used to detect prominent regions of the speech utterances from a single-channel signal.

There has been extensive research to explore the influence of attention on the neural responses in the sensory systems. It has been revealed that the top-down task-dependent influences modulate the neuron responses in the visual and auditory cortex [1], [13]–[15]. This modulation mostly occurs by enhancing the response of neurons tuned to the features of the target stimulus, whereas attenuating the response of neurons to stimuli that did

Manuscript received July 30, 2008; revised December 21, 2008. Current version published June 10, 2009. This work was supported in part by grants from the Office of Naval Research (OFR), the Defense Advanced Research Projects Agency (DARPA), and the U.S. Army. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hiroshi Sawada.

The authors are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: kalinli@usc.edu; shri@sipi.usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2014795

not match the target feature [1], [14], [16], [17]. In addition to this, psychophysical experiments on selective attention have demonstrated that top-down processing can selectively focus attention on a limited range of acoustic feature dimensions [12]. An extensive review of task specific influences on the neural representation of sound is presented in [16]. Psychophysical experiments on selective attention have been recently reviewed in [12].

A. Related Work

In the literature, computational attention models have been mostly explored for vision. In [8] and [18], the concept of *saliency map* was proposed to model bottom-up visual attention in primates. A set of low-level features (such as color, intensity, orientation) is extracted in parallel from an image in multiscales to produce topographic "feature maps" and combined into a single saliency map which indicates the perceptual influence of each part of the image. The saliency map is then scanned to find the locations that attract attention, and it was verified by virtue of eye movement that the model could replicate several properties of human overt attention, i.e., detecting traffic signs, detecting colors, etc., [8].

In [19]–[21], the top-down influence of a task on visual attention was modeled. A guided visual search model was proposed in [19] that combines the weighted feature maps in a top-down manner, i.e., when the task is to detect a red bar, the feature maps that are sensitive to red color get a larger gain factor. This top-down biasing is based on the evidence that neural responses are modulated by task dependent attention [1]. The authors in [20] presented a model that tunes the bottom-up features based on the properties of both the target and the distracter items for visual search problems, i.e., finding the cell phone on a cluttered scene of a desk. A model that combines bottom-up and top-down attention was proposed in [21] to predict where subjects' gaze patterns fall while performing a task of interest. This method was shown to perform well when tested with recorded eye movements of people while playing video games.

Analogies between auditory and visual perception have been widely discussed in the neuroscience literature. The common principles of visual and auditory processing are discussed in [22], and it is suggested that, although early pathways of visual and auditory systems have anatomical differences, there exists a unified framework for central visual and auditory sensory processing. Intermodal and cross-modal interaction between auditory and visual attention is discussed in [16] and [23]. Based on the assumption of parallel processing between vision and audition, an *auditory saliency map*, inspired by the visual saliency map of [8], was proposed for audition in [24] and [25]. In [24], intensity, temporal, and frequency contrast features were extracted from the Fourier spectrum of the sound in multiscales and contributed to the final saliency map in a bottom-up manner. This model was able to replicate some overt properties of auditory scene perception, i.e., the relative salience of short, long, and temporally modulated tones in noisy backgrounds. In our earlier work [25], it was shown that the proposed bottom-up saliency-driven attention model could detect prominent syllables in speech in an unsupervised manner. The motivation behind choosing the prominent syllable detection task was that, during speech perception, a particular phoneme or syllable can be perceived to be more salient than the others due to the coarticulation between phonemes and other factors such as accent and the physical and emotional state of the talker [26], [27]. This information encoded in the acoustical signal is perceived by the listeners, and we showed that these salient syllables can be detected using the bottom-up task-independent auditory attention model proposed by us in [25].

B. Our Model and Contribution

The motivation for the present paper is to analyze the effect of task-dependent influences on auditory attention-inspired speech processing. For example, it is known that when human subjects are asked to find the prominent (stressed) word/syllable, they use their prior task-relevant knowledge, such as prominent words have longer duration [28]. The first goal of the present paper is to provide a detailed analysis of the various acoustic features used in the context of auditory attention-based speech processing. While processing incoming speech stimuli, in addition to acoustic cues, it is well known the brain is also influenced by higher level information such as lexical information, syntax, semantics, and the discourse context [23], [29]. Hence, the second goal of this paper is to analyze the effect of the task-dependent influences, captured via syntactic and lexical cues, working in conjunction with an auditory attention model for automatic prominence detection from speech.

The prominent syllable/word detection can play an important role in speech understanding. For instance, it is important in terms of finding salient regions in speech that may carry critical semantic information. This has applications in speech synthesis for generating more naturally sounding speech when used together with other cues, such as boundary times and intonation patterns [30]. Similarly in speech-to-speech translation systems where it is important to capture and convey concepts from the source to the target language, the ability to handle such salient information contained in speech is critical. Prominent syllable detection also plays a role in word disambiguation and hence in word recognition and synthesis. For example, it has been shown in [31] and [32] that integrating prominence patterns into the automatic speech recognition improved the speech recognizer performance. In summary, extraction of knowledge sources human use beyond segmental level and integration of them into current machine speech processing systems can yield much improved performance. Also, the auditory attention model proposed here is not limited to the prominence detection task; it is a general bio-inspired model and can be applied to other spoken language processing tasks and computational auditory scene analysis applications as discussed in Section VI.

In this paper, we describe a novel task-dependent auditory attention model that works together with higher level lexical and syntactic cues. The auditory attention model proposed here is based on the "gist" phenomenon commonly studied for vision. Given a task and an acoustic scene, the attention model first computes the biologically inspired low-level auditory gist features to capture the overall properties of the scene. Then, the auditory gist features are biased to imitate the modulation effect of task on neuron responses to reliably detect a target or to perform a specific task. In parallel to the auditory attention cues,

the task-dependent influence of lexical and syntactic information is also incorporated into the model using a probabilistic approach (maximum *a posteriori* (MAP) framework). 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) [33], and achieves 88.33% accuracy at the syllable level, providing approximately a 12.4% absolute improvement over using just the bottom-up attention model.

Some related work and preliminary results were presented in our previous work [34] and [35]. Some salient aspects of the current work are enumerated below.

1) A comprehensive analysis of the acoustic features used in the auditory attention model is presented. We present detailed results using mutual information and prominent syllable detection performance.
2) In addition to the intensity, temporal contrast, frequency contrast, and orientation features used in [34] and [35], in this work, a novel bio-inspired pitch feature is also included within the auditory attention model.
3) While in [34] only the acoustic auditory attention cues for prominence detection were used, here an integrated model that utilizes acoustical, lexical, and syntactic information is developed and evaluated.
4) Additionally, in contrast to [35], here we systematically analyze the importance of each piece of acoustic, lexical and syntactic information in the prominence detection performance. Also, in [35] we initially used a lower resolution auditory gist feature extraction method in consideration of the computation cost, whereas here a higher resolution feature set is used in the combined model for achieving further improved performance. In summary, our results indicate that the combined model achieves higher accuracy compared to the results presented in [35] (87.95% versus 88.33%, and it is verified with the Wilcoxon signed rank test (reference Section V) that this is significant at $p \leq 0.005$).
5) We also present prominence detection results at the word level in addition to the prominence detection at the syllable level [34], [35].

The rest of the paper is organized as follows: In Section II, the database used for the experiments is introduced. The auditory attention and the task-dependent higher level cues are presented in Section III. Section IV explains the probabilistic approach proposed to combine the acoustic, lexical, and syntactic cues. The experiments conducted to analyze the auditory attention features and the prominence detection performance obtained with all three cues are presented together with the results in Section V. The discussion and conclusion are presented in Section VI.

II. Database

The Boston University Radio News Corpus (BU-RNC) database [33] is used in the experiments reported in this paper. Being one of the largest speech databases with manual prosodic annotations has made the radio news corpus highly popular for prosodic event detection experiments in the literature. The corpus contains recordings of broadcast news-style read speech that consists of speech from seven speakers (three females and four males). Data for six speakers has been manually labeled with tones and break indices (ToBI) [36] style prosodic tags, totaling about 3 h of acoustic data. The database also contains the orthographic transcription corresponding to each spoken utterance together with time alignment information at the phone and word level. To obtain syllable level time-alignment information, the orthographic transcriptions are syllabified using the rules of English phonology [37], and then the syllable level time-alignments are generated using the phone level alignment information given with the corpus. POS tags for the orthographic transcriptions are also provided with the corpus.

We mapped all pitch accent types ($\mathrm{H}^*, \mathrm{L}^*, \mathrm{L}^* + \mathrm{H}$, 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 labeled "prominent" and otherwise "non-prominent." Also, we derived word level prominence tags from the syllable level prominence tags. The words that contain at least one prominent syllable are labeled as prominent. The database consists of 48 852 syllables; the prominent syllable fraction is 34.3%, and the prominent word fraction is 54.2%. In summary, we chose this database for two main reasons: 1) syllables are stress labeled based on human perception, and 2) since it carries labeled data, it helps us to learn the task-dependent influences carried by lexical and syntactic information, and task-dependent biasing of auditory attention cues as discussed in Section III-A3.

III. Top-Down Task-Dependent Model

The proposed task-dependent model uses two types of evidence: 1) acoustic information captured with the auditory gist features, and 2) higher level top-down information captured with lexical and syntactic models. Next, we discuss modeling of each piece of acoustic, lexical, and syntactic information in detail.

A. Auditory Attention Cues

The block diagram of the auditory gist feature extraction is illustrated in Fig. 1. The feature maps are extracted from sound by using the front-end of the bottom-up auditory attention model proposed by us in [25], 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 the HAS. This 2-D time–frequency auditory spectrum is akin to an image of a scene in vision. Then, a set of multiscale features are extracted in parallel from the auditory spectrum of sound based on the processing stages in the central auditory system (CAS) to produce feature maps. The *auditory gist* of the scene is extracted from the feature maps at low resolution to guide attention during target search, and the attention model biases the gist features to imitate the modulation effect of task on neuron responses using the weights learned for a given task. It should be noted that the proposed task-dependent auditory attention model is a generic model with a variety of applications, i.e., speaker recognition, scene change detection, context recognition, etc., as discussed in Section VI; however, here the task is designed to be prominence

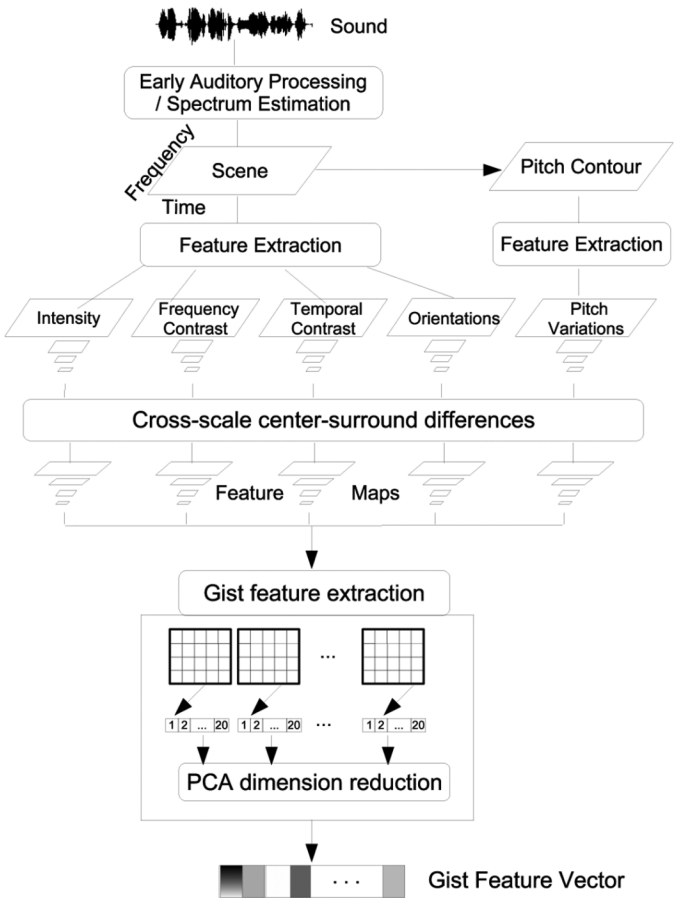

Fig. 1. Diagram of auditory gist feature extraction. The auditory spectrum of the sound (referred as *scene*) is estimated based on the early stages of the HAS. The scene is analyzed by extracting multiscale features from the scene in parallel by mimicking the various stages in the central auditory system. The features (*intensity, frequency contrast, temporal contrast, orientations, and pitch*) are extracted using different sets of receptive filters (ref. Fig. 2). Next, the center-surround differences of features are computed which result in feature maps. Finally, the *auditory gist* of the scene is extracted from the feature maps by capturing the overall properties of the scene at low resolution.

detection in speech. The steps of the auditory attention model: multiscale feature maps and auditory gist extraction followed by task-dependent biasing of the gist features are discussed next.

1) Multiscale Feature Map Generation: 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 based on the information processing stages in the early auditory (EA) system [38]. The EA model used here consists of cochlear filtering, inner hair cell (IHC), and lateral inhibitory stages mimicking the process from basilar membrane to the cochlear nucleus in the human auditory system. The raw time-domain audio signal is filtered with a bank of 128 overlapping constant-Q asymmetric bandpass filters with center frequencies that are uniformly distributed along a tonotopic (logarithmic) frequency axis analogous to cochlear filtering. This is followed by a differentiator, a nonlinearity, and a low-pass filtering mimicking the IHC stage, and finally a lateral inhibitory network [38]. Here, sound is analyzed using a 20-ms window shifted every 10 ms; i.e., each 10-ms audio frame is represented by a 128-dimensional vector.

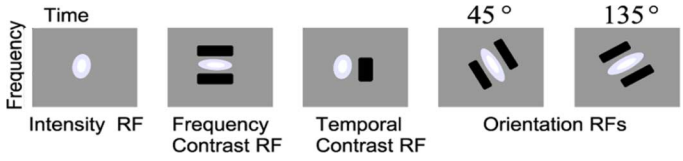

Fig. 2. 2-D spectro–temporal receptive filters. These filters mimic the analysis stages in the primary auditory cortex. The excitation and inhibition phase are shown with white and black color, respectively. For example, the frequency contrast filters correspond to the receptive fields in the auditory cortex with an excitatory phase and simultaneous symmetric inhibitory side bands.

The output of the EA model is an *auditory spectrum* with time and frequency axes, and here it is referred to as a "*scene*." In the next stage, the scene is analyzed by extracting a set of multiscale features that are similar to the information processing stages in the CAS. Auditory attention can be captured by (bottom-up) or selectively directed (top-down) to a wide variety of acoustical features such as intensity, frequency, temporal, pitch, timbre, FM direction or slope (called "orientation" in the current paper) and spatial location [16], [23]. Here, five features are included in the model to encompass all the aforementioned features except spatial location, and spatial location information is beyond the scope of this paper. The features included in the model are *intensity* (I), *frequency contrast* (F), *temporal contrast* (T), *orientation* (O), and *pitch* (P), and they are extracted in multiscales using 2-D spectro–temporal receptive filters mimicking the analysis stages in the primary auditory cortex [39], [22]. All the receptive filters (RF) simulated here for feature extraction are illustrated in Fig. 2. The excitation phase (positive values) and inhibition phase (negative values) are shown with white and black color, respectively.

The intensity filter mimics the receptive fields in the auditory cortex with only an excitatory phase selective for a particular region [39] and can be implemented with a Gaussian kernel. The multiscale intensity features $I(\sigma)$ are created using a dyadic pyramid: the input spectrum is filtered with a 6×6 Gaussian kernel [1,5,10,10,5,1]/32 and reduced by a factor of two, and this is repeated [40]. If the scene duration D is large (i.e., $D > 1.28$ s), the number of scales is determined by the number of bandpass filters used in the EA model, hence eight scales $\sigma = \{1, \ldots, 8\}$ are created yielding size reduction factors ranging from 1:1 (scale 1) to 1:128 (scale 8). Otherwise, there are fewer scales.

Similar to $I(\sigma)$, the multiscale $F(\sigma)$, $T(\sigma)$, $O_\theta(\sigma)$ features are extracted using the filters described below on eight scales (when the scene duration D is large enough), each being a resampled version (factor 2) of the previous one. The frequency contrast filters correspond to the receptive fields with an excitatory phase and simultaneous symmetric inhibitory side bands, and the temporal contrast filters correspond to the receptive fields with an inhibitory phase and a subsequent excitatory phase as described in [24], [39], and they are shown in Fig. 2. The filters used for extracting frequency and temporal contrast features can be interpreted as horizontal and vertical orientation filters used in the visual saliency map [8], [24]. These filters are implemented using a 2-D Gabor filter (product of a cosine function with 2-D Gaussian envelope [40]) with orientation $\theta = 0°$ for frequency contrast $F(\sigma)$ and $\theta = 90°$ for temporal

contrast $T(\sigma)$. In the lowest scale, the frequency contrast filter has 0.125 octave excitation with same width inhibition side bands (24 channels/octave in EA model), and the temporal contrast filter is truncated such that it has a 30-ms excitation phase flanked by a 20-ms inhibition phase. The orientation filters mimic the dynamics of the auditory neuron responses to moving ripples [22], [39]. This is analogous to motion energy detectors in the visual cortex. To extract orientation features $O_\theta(\sigma)$, 2-D Gabor filters with $\theta = \{45°, 135°\}$ are used. They cover approximately 0.375 octave frequency band in the lowest scale. The exact shapes of the filters used here are not important as long as they can manifest the lateral inhibition structure, i.e., an excitatory phase with simultaneous symmetric inhibitory sidebands [41].

Pitch information is also included in our model because it is an important property of sound; using a method of extracellular recording, it was shown that the neurons of the auditory cortex also respond to pitch [42]. Further, in [11] it was shown that participants noticed the change in the pitch of the sound played in the unattended ear in a dichotic listening experiment, which indicates that pitch contributes to auditory attention. In general, there are two hypotheses for the encoding of pitch in the auditory system: temporal and spectral [22]. We extract pitch based on the temporal hypothesis which assumes that the brain estimates the periodicity of the waveform in each auditory nerve fiber by autocorrelation [43]. Then, a piecewise second-order polynomial model is fit to the estimated pitch values in the voiced regions for smoothing. We mapped the computed pitch values to the tonotopic cortical axes assuming that the auditory neurons in the cochlear location corresponding to the pitch are fired. Then, the multiscale pitch features $P(\sigma)$ are created using the following 2-D spectro–temporal receptive filters similar to those discussed above.

- Frequency Contrast RF: to capture pitch variations over the tonotopic axis, i.e., spectral pitch changes. These pitch features are denoted as P_F
- Orientation RFs with $\theta = 45°$ and $\theta = 135°$: to capture raising and falling pitch behavior. These pitch features are denoted as P_{O_θ}.

In summary, eight feature sets are computed in the model; one feature set for each I, F, T, two feature sets for O_θ (with $\theta = 45°$ and $\theta = 135°$), and three feature sets for P (P_F, P_{O_θ} with $\theta = 45°$ and $\theta = 135°$).

As shown in Fig. 1, after extracting features at multiple scales, "center-surround" differences are calculated resulting in "feature maps." The center-surround operation mimics the properties of local cortical inhibition, and it is simulated by across scale subtraction ($\ominus$) between a "center" fine scale c and a "surround" coarser scale s followed by rectification [8], [44]

$$\mathcal{M}(c,s) = |\mathcal{M}(c) \ominus \mathcal{M}(s)|, \quad \mathcal{M}\epsilon\{I, F, T, O_\theta, P\}. \tag{1}$$

The across scale subtraction between two scales is computed by interpolation to the finer scale and point-wise subtraction. Here, $c = \{2, 3, 4\}$, $s = c + \delta$ with $\delta\epsilon\{3, 4\}$ are used. Thus, six feature maps are computed per feature set resulting in total 48 feature maps when the features are extracted in eight scales. Next, the feature maps are used to extract the gist of a scene.

2) Gist Features: The auditory attention model proposed here is based on the "gist" phenomenon commonly studied for vision. [45] defines two types of gist: perceptual gist and conceptual gist. Perceptual gist refers to the representation of a scene built during perception, and conceptual gist includes the semantic information inferred from a scene and stored in memory. Here, we focus on perceptual gist. A reverse hierarchy theory related to perceptual gist was proposed in [3] for vision. Based on this theory, gist processing is a pre-attentive process and guides attention to focus on a particular subset of stimuli locations to analyze the details of the target locations. The gist of a scene is captured by humans rapidly within a few hundred milliseconds of stimulus onset and describes the type and overall properties of the scene; i.e., after very brief exposure to a scene, a subject can report general attributes of the scene, i.e., whether it was indoors, outdoors, kitchen, street traffic, etc. [3], [46]. In [47], a review of gist perception is presented, and it is argued that gist perception also exists in audition.

Our gist algorithm is inspired by the work in [3], and [48]. We formalize gist as a relatively low-dimensional acoustic scene representation which describes the overall properties of a scene at low-resolution; hence, we represent gist as a feature vector [48]. Then, the task-dependent top-down attention is assumed to focus processing to the specific dimensions of the gist feature vector to maximize the task performance, which is implemented using a learner as discussed in Section III-A3. We present the details of the auditory gist extraction algorithm in the rest of this subsection.

A gist vector is extracted from the feature maps of I, F, T, O_θ, P 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 $m * n$. For a feature map $\mathcal{M}_i$ with height h and width v, the computation of gist features can be written as

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

and i is the feature map index, i.e., $i = \{1, \ldots, 48\}$ for features extracted at eight scales.

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 redundancy and the dimension to make the subsequent machine learning more practical while still retaining 99% of the variance. The final auditory gist feature (after PCA), is denoted with f, and the dimension of f is denoted with d in the rest of the paper.

3) Task-Dependent Biasing of Acoustic Cues: As stated in Section I, the top-down task-dependent influences modulate neuron responses in the auditory cortex while searching for a target [1], [13]–[15]. This modulation mostly occurs by

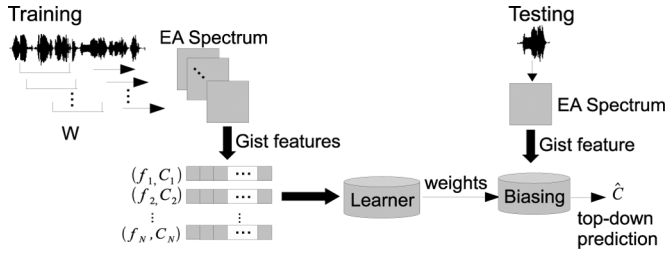

Fig. 3. Auditory attention model. Training phase: the weights are learned in supervised manner. Testing phase: auditory gist features are biased with the learned weights to estimate the top-down model prediction.

enhancing the response of neurons tuned to the features of target stimuli, whereas attenuating the response of neurons to stimuli that did not match the target feature [1], [14], [16], [17]. Thus, we formalize the task-dependent top-down process as follows: given a task (which is prominence detection in the current paper), the top-down task-dependent auditory attention model biases the auditory gist features with weights learned in a supervised manner for the task such that it enhances specific dimensions of the gist features that are related to the task, while attenuating the effect of dimensions which are not related to the task. Here, the weights are learned in a supervised manner as illustrated in Fig. 3; first the data is split into training and test sets. In the training phase, gist features f_i are extracted from the scenes in the training set and compiled together with their corresponding prominence class categories C_i. The features are stacked and passed through a "learner" (a machine learning algorithm) to discover the weights. In the testing phase, scenes that are not seen in the training phase are used to test the performance of the top-down model. For a given test scene, the gist of the scene is extracted and biased with the learned weights to estimate the top-down model prediction $\hat{C}$. Here, a three-layer neural network is used to implement the learner in Fig. 3 as discussed in detail in Section V-A. The reason for using neural network is that they are biologically well motivated; it mimics the modulation effect of task dependent attention on the neural responses.

In this context, 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 sound surrounding a syllable with an analysis window of duration D that centers on the syllable. The analysis of scene duration D is described later in Section V-A1.

B. Task-Dependent Higher Level Cues

Speech is one of the most important sound sources for human listeners. While processing speech stimuli, the brain is influenced by higher level information such as lexical information, syntax, semantics, and the discourse context [23], [29]. For example, one famous result from the dichotic listening experiments reported in [49] is that people may respond to the messages on the unattended channel when they heard their own names. In the experiments of [49], 8% of the participants responded to the message "you may stop now" when it was presented on the unattended channel, whereas 33% of the participants responded when the message was preceded by the participant's name. This is similar to what happens in the cocktail party phenomenon. For example, one may hear her/his name being mentioned by someone else across the room, even though she/he was not consciously listening for it. In the experiments of [29], the recorded neurophysiologic brain response was larger for one's native language than unfamiliar sounds. Also, the experiments at the level of meaningful language units have revealed that real words elicit a larger brain response than meaningless pseudowords [29]. The psychophysical experiments indicate that some words that carry semantically important information, e.g., one's name, can capture attention, as can some syllable/word strings that form a meaningful word/sentence [11], [23].

In addition to these, earlier studies have revealed that there is dependency between prominence and lexical information and also between prominence and syntax [28], [27]. The authors of [28] show that content words are more likely to be prominent than function words. Also, a statistical analysis presented in [50] indicates that some syllables have a higher chance of being prominent than others; i.e., the syllable "can" has an 80% chance of being prominent, whereas the syllable "by" has a 13% chance of being prominent.

We incorporate the task-dependent higher level cues into our model using lexical and syntactic information for prominence detection in speech. Specifically, this information is used to create probabilistic models for the current application of prominent syllable detection. The lexical information is incorporated in the system by building a language model with syllables as explained in Section IV-B. The syntactic knowledge is represented using part-of-speech tags, and a neural network is used to model the influence of syntax on prominence as detailed in Section IV-C.

IV. Probabilistic Approach for Task-Dependent Model

The task-dependent model is influenced by acoustic and other higher level cues. In this section, we present a system to combine the auditory attention cues discussed in Section III-A together with lexical and syntactic information in a probabilistic framework for prominence detection in speech. The probabilistic model is based on a MAP; given acoustic, lexical and syntactic information, the model estimates the sequence of prominence that maximizes the posterior probability. First, we discuss modeling of each piece of information separately, and then we discuss how they are combined in a MAP framework.

A. Task-Dependent Model With Auditory Gist Features

A multilayer perceptron (MLP) is used to implement the learner in Fig. 3 to bias the auditory gist features to mimic the top-down influences of task on neuron responses. We use auditory gist features as the input to the neural network, and the output returns the class posterior probability $p(c_i|f_i)$ for the ith syllable, where f_i is the auditory gist feature, and $c_i \epsilon \{0, 1\}$ where 1 denotes that the syllable is prominent, while 0 denotes that it is nonprominent. 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 MAP framework as follows:

$$\mathbf{C}^* = \underset{\mathbf{C}}{\operatorname{argmax}}\, p(\mathbf{C}|\mathbf{F}). \tag{3}$$

Assuming that the syllable prominence classes are independent, (3) can be approximated as

$$\mathbf{C}^* = \underset{\mathbf{C}}{\operatorname{argmax}} \prod_{i=1}^{M} p(c_i|f_i) \tag{4}$$

when the gist features are the only information considered in the top-down model.

B. Task-Dependent Model With Lexical Cues

The lexical evidence is included in the top-down model using a probabilistic 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

$$\mathbf{C}^* = \underset{\mathbf{C}}{\operatorname{argmax}}\, p(\mathbf{C}|\mathbf{S}) \tag{5}$$

$$= \underset{\mathbf{C}}{\operatorname{argmax}}\, p(\mathbf{C}, \mathbf{S}). \tag{6}$$

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

$$p(\mathbf{C}, \mathbf{S}) = \underset{\mathbf{C}}{\operatorname{argmax}} \prod_{i=1}^{M} p\left(c_i, s_i|c_{i-n+1}^{i-1}, s_{i-n+1}^{i-1}\right). \tag{7}$$

For example, in a bigram context model, $p(\mathbf{C}, \mathbf{S})$ can be approximated as

$$p(\mathbf{C}, \mathbf{S}) = p(c_1, s_1) \cdot \prod_{i=2}^{M} p(c_i, s_i|c_{i-1}, s_{i-1}). \tag{8}$$

We assume that the utterance transcripts are available and use the actual transcripts provided with the database in the experiments. Then, the syllable s_i becomes known; hence, we replace $p(c_i, s_i|c_{i-n+1}^{i-1}, s_{i-n+1}^{i-1})$ term in (7) with $p(c_i|c_{i-n+1}^{i-1}, s_{i-n+1}^{i})$. It is difficult to robustly estimate the language model even within n-gram context due to the size of the available training database. Specifically, the BU-RNC database used in this study is small compared to the large number of syllables in the dictionary. Hence, a factored language model is built to overcome data sparsity. The factored language model is a flexible framework for incorporating various information sources, such as morphological classes, stems, roots, and any other linguistic features, into language modeling [51]. In a factored language model scheme, when the higher order distribution cannot be reliably estimated, it backs off to lower order distributions. Using a factored language model also helps with out-of-vocabulary (OOV) syllables seen in the test sets because when an OOV syllable is observed on the right side of the conditioning bar in $p(c_i, s_i|c_{i-n+1}^{i-1}, s_{i-n+1}^{i-1})$, the backed off estimate that does not contain that variable is used instead. The back-off graph used for creating the language model is shown in Fig. 4. We use a back-off path such that the most distant variable is dropped first from the set on the right side of the conditioning bar in $p(c_i, s_i|c_{i-n+1}^{i-1}, s_{i-n+1}^{i-1})$. Here, both s_{i-n+1} and c_{i-n+1} are the most distant variables. For instance, we first drop the most distant syllable variable s_{i-n+1} and then the most distant class variable c_{i-n+1} and so on. As shown in Fig. 4, the graph includes both possible paths of starting to drop from either s_{i-n+1} or c_{i-n+1}. The language model is built using the SRILM toolkit [52]. 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.

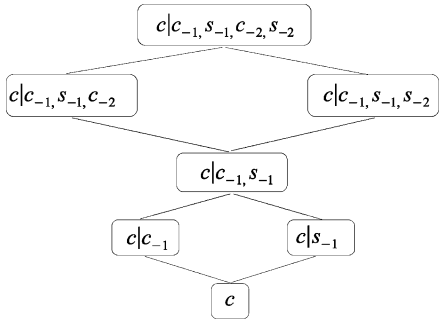

Fig. 4. Backoff graph of lexical-prominence language model moving from the trigram model $p(c|c_{-1}, s_{-1}, c_{-2}, s_{-2})$ down to $p(c)$, where c denotes prominence class, and s represents syllable token. The most distant variable s_{-2}/c_{-2} is dropped first.

C. Task-Dependent Model With Syntactic Cues

In our model, the syntactic information is captured using POS tags provided with the database. POS tagging provides details about a particular part of speech within a sentence or a phrase by looking at its definition as well as its context (relationship with adjacent words), i.e., whether a word is a noun, verb, adjective, etc. The syntactic evidence is included in the model using a similar approach to the lexical evidence. 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

$$\mathbf{C}^* = \underset{\mathbf{C}}{\operatorname{argmax}}\, p(\mathbf{C}|\mathbf{POS}(\mathbf{W})) \tag{9}$$

Assuming that word tokens are independent, (9) can be rewritten as

$$\mathbf{C}^* = \underset{\mathbf{C}}{\operatorname{argmax}} \prod_{i} p\left(c_i|\mathbf{POS_i^L}(w_i)\right). \tag{10}$$

In (10), c_i represents the prominence of the ith word w_i in the word sequence, and $\mathbf{POS_i^L}$ is the POS tags sequence that is neighboring the ith word

$$\mathbf{POS_i^L} = \left(\mathrm{POS}_i^{i-(L-1)/2}, \ldots, \mathrm{POS}_i^{i}, \ldots, \mathrm{POS}_i^{i+(L-1)/2}\right). \tag{11}$$

$\mathbf{POS_i^L}$ is chosen such that it contains syntactic information from a fixed window of L words centered at the ith word, and $L = 5$ performs well for prominence detection [53]. The class posterior probability $p(c_i|\mathbf{POS_i^L}(w_i))$ in (10) is computed using a neural network as detailed in Section V-C. In the rest of

the paper, $\mathbf{POS}$ is used to denote $\mathbf{POS_i^L}$ to make the notation simpler.

The syntactic model is built at the word level since POS tags are associated with words. Thus, the syntactic model indicates whether one or more syllables within the word are prominent. However, it does not provide information regarding which one is the prominent one. On the other hand, if a word is nonprominent, so are the syllables composing the word.

Let us assume that the ith word w_i consists of n_i syllables. Then the syllable string for w_i can be written as $\mathbf{S_i} = [s_1, s_2, \ldots, s_{n_i}]$. The word w_i is nonprominent if and only if all the syllables within w_i are nonprominent. Hence

$$p(c_i = 0|\mathbf{POS}(w_i)) = \prod_{k=1}^{n_i} p(c_k = 0|\mathbf{POS}(s_k)) \tag{12}$$

where $p(c_i = 0|\mathbf{POS}(w_i))$ is the probability of w_i being nonprominent 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 nonprominent given the POS tags. From (12), we approximate the posterior probability of a syllable in a word being nonprominent as follows:

$$p(c_k = 0|\mathbf{POS}(s_k)) \approx \sqrt{[n_i]}p(c_i = 0|\mathbf{POS}(w_i)). \tag{13}$$

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

$$p(c_k = 1|\mathbf{POS}(s_k)) = 1 - p(c_k = 0|\mathbf{POS}(s_k)). \tag{14}$$

In practice, to bring word level syntactic information to syllable level, (13) and (14) are used for the experiments.

D. Combined Model With Acoustic and Higher Level Cues

The top-down task-dependent model uses acoustic and higher level cues while performing a task. These cues are combined using a probabilistic approach for the purpose of prominence detection. Given auditory gist features $\mathbf{F}$, lexical evidence $\mathbf{S}$, and syntactic evidence $\mathbf{POS}$, the most likely prominence sequence can be found as

$$\begin{aligned}\mathbf{C}^* &= \underset{\mathbf{C}}{\operatorname{argmax}}\ p(\mathbf{C}|\mathbf{F}, \mathbf{S}, \mathbf{POS}) \\ &= \underset{\mathbf{C}}{\operatorname{argmax}}\ p(\mathbf{F}, \mathbf{S}, \mathbf{POS}|\mathbf{C})p(\mathbf{C}).\end{aligned} \tag{15}$$

Assuming that the auditory gist features are conditionally independent of the lexical and syntactic features given the prominence class information, (15) can be written as

$$\mathbf{C}^* = \underset{\mathbf{C}}{\operatorname{argmax}}\ p(\mathbf{F}|\mathbf{C})p(\mathbf{S}, \mathbf{POS}|\mathbf{C})p(\mathbf{C}). \tag{16}$$

The joint distribution $p(\mathbf{S}, \mathbf{POS}|\mathbf{C})$ in (16) 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 it as follows:

$$p(\mathbf{S}, \mathbf{POS}|\mathbf{C}) \approx p(\mathbf{S}|\mathbf{C})p(\mathbf{POS}|\mathbf{C}). \tag{17}$$

Then (16) can be rewritten as

$$\begin{aligned}\mathbf{C}^* &= \underset{\mathbf{C}}{\operatorname{argmax}}\ p(\mathbf{F}|\mathbf{C})p(\mathbf{S}|\mathbf{C})p(\mathbf{POS}|\mathbf{C})p(\mathbf{C}) \\ &= \underset{\mathbf{C}}{\operatorname{argmax}}\ \frac{p(\mathbf{C}|\mathbf{F})}{p(\mathbf{C})}p(\mathbf{S}, \mathbf{C})\frac{p(\mathbf{C}|\mathbf{POS})}{p(\mathbf{C})}.\end{aligned} \tag{18}$$

The combined top-down model, which includes auditory gist features and lexical and syntactic information, finally reduces to the product of individual probabilistic model outputs as shown in (18).

V. Experiments and Results

This section presents the details of the experiments conducted, including the results for the automatic prominence detection tests. For the experiments, we report prominent syllable detection accuracy (Acc) together with precision (Pr), recall (Re) and F-score (F-sc) values. These measures are

$$\begin{aligned}\mathrm{Acc} &= \frac{tp + tn}{tp + fp + tn + fn} \times 100, \\ \mathrm{Pr} &= \frac{tp}{tp + fp}, \\ \mathrm{Re} &= \frac{tp}{Tp + fn}, \\ \mathrm{F-sc} &= \frac{2.\mathrm{Pr}.\mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}}\end{aligned} \tag{19}$$

where tp and tn denote true positive and negative, and fp and fn denote false positive and false negative. All of the experimental results presented here are estimated using the average of fivefold cross-validation. We used all the data manually labeled with ToBI style prosody tags and randomly split it into five groups $(G1, G2, G3, G4, G5)$ to create five cross-validation sets. In each set, four groups were used for training, and one group was used for testing, i.e., 80% of the data was used for training and the remaining 20% of the data was retained for testing. For example, $G2, G3, G4, G5$ were used for training, when $G1$ was used for testing. On average, the number of syllables in the training and test sets was 39 082 and 9770, respectively. The number of unique syllables in the training sets was 2894, while it was 1728 for the test sets (averaged over the five cross-validation sets). The average number of OOV syllables in the test sets was 220 (12.7% relative to the test vocabulary). The baseline prominence accuracy, which is the chance level, is 65.7% at the syllable level.

The Wilcoxon signed rank test is used to report the confidence level in terms of significance values (p-values) whenever we make comparisons throughout the paper. Here, the comparison is performed in terms of achieved accuracy values for the samples. The Wilcoxon signed rank test is a nonparametric test, and it does not make any assumption about the distributions of samples. The test is available as part of MATLAB software, and more information about the test can be found in [54].

This section is organized as follows. First, the experiments and results obtained with the auditory attention model are presented in Section V-A. More specifically, Section V-A

includes scene duration and grid size selection analyses presented in Subsections V-A1 and V-A2, respectively. This is followed by an analysis of auditory attention features using mutual information and prominence detection experiments conducted with each individual feature and their combinations in Subsection V-A3. Next, the prominence detection results obtained with lexical and syntactic information are presented in Section V-B and V-C, respectively. Finally, the results with the combined acoustic, lexical, and syntactic model are presented in Section V-D.

A. Experiments and Results With Auditory Attention Model

In this section, we present the experiments conducted with the auditory attention model which was discussed in Section III-A. The learner in Fig. 3 is implemented using a three-layer neural network with d inputs, $(d+n)/2$ hidden nodes and n output nodes, where d is the length of gist feature vector after PCA dimension reduction, and $n=2$ since this is a two-class problem as discussed in Section II. The output of the neural network can be treated as class posterior probability, and the class with higher probability is assumed to be the top-down model prediction. The reason for using a neural network is that they are biologically well motivated; it mimics the modulation effect of task dependent attention on the neural responses.

1) Analysis of "Scene" Duration: The role of scene duration is investigated in the experiments, and the results are discussed in this section. A scene is generated for each syllable in the database. As explained earlier, scenes are produced by extracting the sound around each syllable with an analysis window that centers on the syllable. We used the statistics of the BU-RNC to determine a range of values for the scene window duration D. It was found that the mean syllable duration is approximately 0.2 s with 0.1-s standard deviation, and the maximum duration is 1.4 s for the database. The scene duration is varied starting from 0.2 s, considering only the syllable by itself, up to 1.2 s considering the neighboring syllables.

In order to get full temporal resolution while analyzing the scene duration, at the gist feature extraction stage each feature map is divided into 1-by-v grids, where v is width of a feature map. This results in a $(1*v)=v$ dimensional gist vector for a single feature map, and v varies with scene duration and center scale c. For instance, when $D=0.6$ s, the early auditory (EA) model outputs a 128×60 dimensional scene since there are 128 bandpass filters used for cochlear filtering in the EA model, and the analysis window is shifted by 10 ms. Then, we can extract features up to six scales, which enables the center-surround operation at scales $(c-s)=\{(2-5),(2-6),(3-6)\}$, and in turn generates three feature maps. v width of a feature map is a function of scene duration and center scale c; given that the analysis window shift is 10 ms, $v=(D/0.01)/2^{(c-1)}$. When the dimension of a scene is 128×60 and 1-by-v is the grid size, the dimension of the gist vector of a single feature set is $(30+30+15)=75$ (since v is 30 at $c=2$ and 15 at $c=3$), finally resulting in a cumulative gist vector of $(75*8)=600$ dimension (one feature set for each I,F,T and two sets for O_θ since $\theta=\{45^\circ,135^\circ\}$, and three sets for P (P_F,P_{O_θ} with $\theta=\{45^\circ,135^\circ\}$), total eight sets). After principal component analysis, the dimension of the gist vector is reduced to $d=60$.

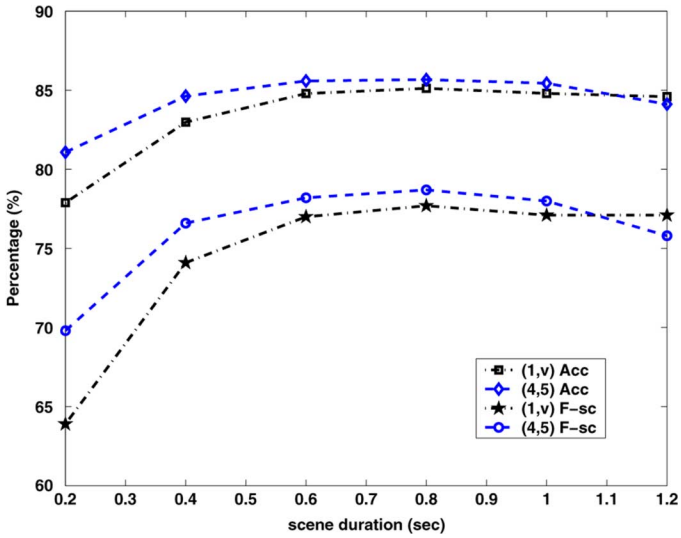

Fig. 5. Performance of prominence detection as a function of scene duration for grid sizes of 1-by-v and 4-by-5 (v: width of a feature map, Acc: Accuracy (%), F-sc: F-score (%)).

TABLE I
PROMINENT SYLLABLE DETECTION PERFORMANCE WITH TASK-DEPENDENT AUDITORY ATTENTION MODEL

D (s)	1-by-v grids			4-by-5 grids		
	d	Acc.	F-sc	d	Acc.	F-sc
0.2	12	77.89%	0.64	36	81.08%	0.7
0.4	37	82.98%	0.74	84	84.63%	0.77
0.6	60	84.79%	0.77	94	85.59%	0.78
0.8	**98**	**85.12%**	**0.78**	**134**	**85.67%**	**0.79**
1.0	123	84.80%	0.77	130	85.44%	0.78
1.2	153	84.59%	0.77	138	84.12%	0.76

The performance of prominence detection obtained with 1-by-v grid size is shown in Fig. 5 as a function of scene duration. It can be observed from Fig. 5 that the accuracy does not change significantly for varying scene durations for $D\geq 0.6$ s. The best performance achieved is when $D=0.8$ s. The prominence performance is poor for short scene durations ($D<0.6$ s), especially for the case when $D=0.2$ s (scene approximately includes only the syllable by itself). This essentially indicates that the prominence of a syllable is affected by its neighboring syllables.

In Table I, the results for selected values of scene duration are presented with the values of accuracy, F-score (F-sc), and the dimension of gist features after PCA (d). The gist feature dimension d gets larger, requiring a larger neural network for training when the scene duration is larger as shown in Table I. When 1-by-v grid size is used while extracting auditory gist features, the best prominent syllable detection performance achieved is 85.1% accuracy with an F-score of 0.78 obtained when $D=0.8$ s.

2) Analysis of Grid Size Selection: In this section, the effect of grid size on the prominence detection performance is examined. The resolution in the frequency domain is increased by a factor of four while reducing the temporal resolution so that the dimension stays compact. At the gist feature extraction stage each feature map is divided into 4-by-5 grids, resulting in a fixed $(4*5)=20$ dimensional gist vector for a single feature map.

TABLE II
PROMINENT SYLLABLE DETECTION PERFORMANCE WITH BOTTOM-UP SALIENCY-BASED ATTENTION MODEL [25]

D (s)	d	Acc	Pr	Re	F-sc
0.6	NA	75.9%	0.64	0.79	0.71

This is different from the one in Section V-A1 where the dimension of a gist vector for a feature map varies with v (hence with scene duration). As in the previous example in Section V-A1, when $D = 0.6$ s, the model generates three feature maps per feature set and hence a $20*3*8 = 480$ dimensional cumulative gist feature vector. After the principal component analysis, the dimension is reduced to 94. As listed in Table I, for all scene durations, this selection results in a larger dimensional gist feature vector compared to 1-by-v grid size, i.e., for scene duration of 0.6 s the dimension of gist feature with 1-by-v grid size is 60 whereas it is 94 with 4-by-5 grid size. This indicates that the gist features obtained with 4-by-5 grids carries more diverse information about the scene compared to the one obtained with 1-by-v grids. The results obtained with 4-by-5 grids for varying scene durations are also reported in Fig. 5 and Table I. The best performance achieved with 4-by-5 grid size is 85.7% accuracy with an F-score $= 0.79$, and obtained again when $D = 0.8$ s.

The performance obtained with 4-by-5 grid size selection is better than the one obtained with 1-by-v grid size ($p \leq 0.001$) except for scene duration of 1.2 s (reference Fig. 5). This might be due to the fact that the temporal resolution is not adequate with 4-by-5 grid size selection for large scene durations. This also indicates that, while choosing the grid size, the scene duration should be factored in while choosing the temporal grid size that determines temporal resolution. Larger scene durations might need larger temporal grids in order to obtain adequate temporal resolution. Also, even though the best performance obtained with both grid sizes is with scene duration of $D = 0.8$ s, this is not significantly better than the results obtained with scene duration of 0.6 s (at $p \leq 0.001$). Hence, we fix the scene duration as 0.6 s in the rest of the experiments since it is computationally less expensive (the feature dimension is smaller, and so is the neural network).

The results obtained with our unsupervised bottom-up (BU) attention model from [25] are also summarized in Table II for comparison purpose. The top-down auditory attention model provides approximately 10% absolute improvement over the unsupervised bottom-up auditory attention model.

3) Analysis of Auditory Attention Features: We present an analysis of the auditory attention features using mutual information and prominence detection experiments conducted with each individual feature and their combinations in this subsection. The scene duration is fixed at 0.6 s for the analysis due to its sufficient performance as discussed in Section V-A1 and V-A2. First, pitch feature sets are analyzed to provide insight into the features extracted with different receptive filters. Then, mutual information estimations are presented to measure the amount of redundancies between features and also the amount of information each feature set and their combinations provide about the syllable prominence.

a) Analysis of Pitch Features: These results indicate that the gist features obtained from the pitch contour using the $F, O_{45^\circ}, O_{135^\circ}$ receptive filters capture the pitch variations and behavior. Also, there is no need for normalization since the gist features capture variations rather than the absolute values. Finally, the prominence detection performances obtained with using only pitch features are detailed in Table III for both 1-by-v and 4-by-5 grid sizes. The best performance is achieved with pitch when all three RFs $(F, O_{45^\circ}, O_{135^\circ})$ are used to extract pitch gist features. Also, 4-by-5 grids performs better than 1-by-v grids, and the best achieved performance is 81.26% accuracy with an F-score of 0.69 via using 4-by-5 grids. In the rest of the paper, the pitch features are extracted from the pitch contour using all three receptive filters $(F, O_{45^\circ}, O_{135^\circ})$ and pitch features are denoted with P to prevent confusion with other features. As described in Section III-A, pitch is extracted from the auditory spectrum and then mapped onto the tonotopic axis, assuming that the auditory neurons in the cochlear location corresponding to the pitch are fired. Then, this 2-D representation is analyzed to capture pitch behavior using frequency contrast and orientation receptive filters. Pitch analysis results for a sample speech scene are illustrated in Fig. 6. The top figure shows the mapped pitch contour itself. The gist feature vectors obtained from this contour using frequency contrast and orientation filters are shown below it. Here, only the raw gist vector (without PCA) obtained from the feature map with center scale $c = 2$, surround scale $s = 5$ and grid size 1-by-v (v is width of a feature map) is shown. The vector is interpolated to scale 1 for time alignment purpose. It

TABLE III
PROMINENT SYLLABLE DETECTION PERFORMANCE WITH ONLY PITCH FEATURES

Pitch Feature	1-by-v grids			4-by-5 grids		
	d	Acc.	F-sc	d	Acc.	F-sc
P_F	21	73.90%	0.57	17	79.44%	0.67
$P_{O_{45^o}}$	15	76.10%	0.60	30	79.48%	0.67
$P_{O_{135^o}}$	14	74.99%	0.58	29	78.65%	0.65
$P_{O_{45^o}}$ & $P_{O_{135^o}}$	26	78.88%	0.66	44	80.80%	0.69
P_F & $P_{O_{45^o}}$ & $P_{O_{135^o}}$	42	**80.13%**	0.68	54	**81.26%**	0.70

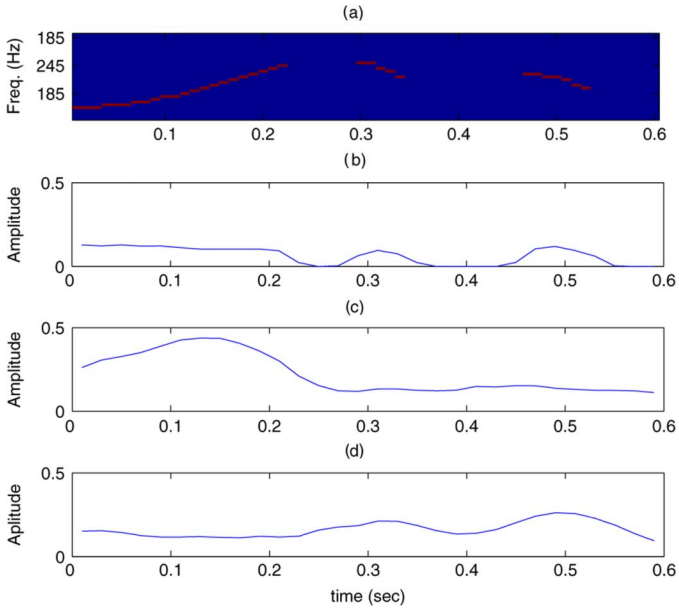

Fig. 6. Pitch analysis of a speech scene with grid size of 1-by-v (a) pitch. Output obtained with (b) frequency contrast RF. (c) Orientation RF with 45° rotation. (d) Orientation RF with 135° rotation.

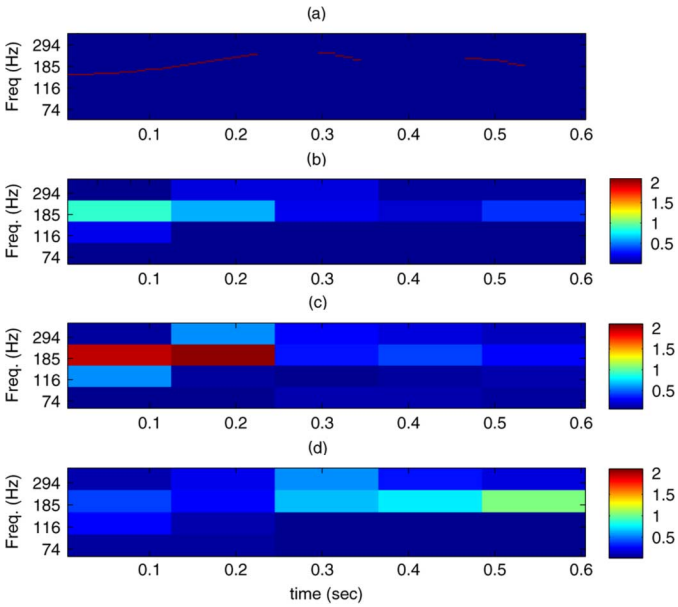

Fig. 7. Pitch analysis of a speech scene with grid size of 4-by-5 (a) pitch. Output obtained with (b) frequency contrast RF. (c) Orientation RF with 45° Rotation. (d) orientation RF with 135° rotation.

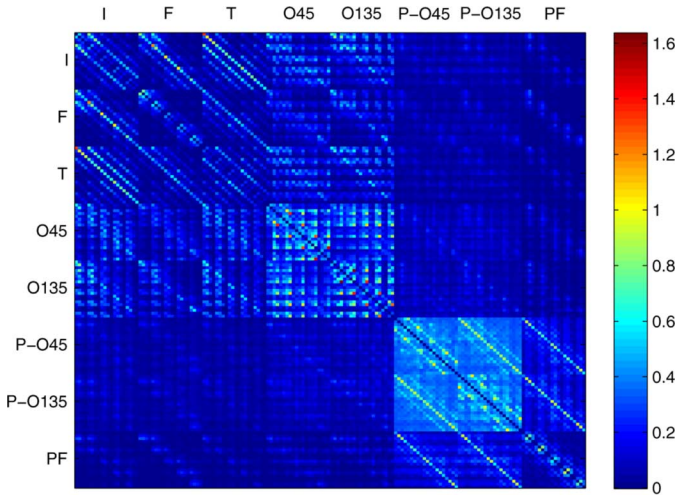

Fig. 8. Pair-wise MI between raw gist features created using 4-by-5 grid size and 0.6-s scene duration. Only the gist vector extracted from the feature map with $c = 2, s = 5$ is illustrated. The diagonal elements are set to zero.

can be observed from the figure that, for the segments where pitch is rising, the gist values obtained with 45° orientation RF show high activity, whereas for the segments where pitch is falling, the gist values obtained with 135° orientation RF show high activity. Also, it can be observed that when the duration of pitch rising/falling is longer, in other words when pitch variation is larger, the gist components are larger, i.e., in Fig. 6 the pitch rising from 0.1 to 0.2 s results in gist values with O_{45° that are larger compared to the gist values obtained with O_{135° when pitch is falling for a shorter duration around 0.3 s or 0.5 s. The gist vector extracted with frequency contrast RF helps to detect voiced regions, especially the segments where there is a pitch plateau. The same speech segment is analyzed with 4-by-5 grid selection and this is illustrated in Fig. 7. It has similar characteristics as the gist features extracted with 1-by-v grids, except that with 4-by-5 grid size the frequency resolution is higher and the range of pitch values is also roughly encoded in the gist features (place coding).

b) Feature Analysis with Mutual Information Measure: Here, we use mutual information (MI) measure to analyze the intensity, frequency contrast, temporal contrast, orientations, and pitch features. In particular, the MI between the individual features (and their combinations) and the prominence classes are computed to measure the statistical dependency between each feature and the syllable prominence. Also, the MI between feature pairs is computed to measure the redundancy between features.

The MI between continuous random variables X and Y can be written in terms of the differential entropies as

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{20}$$

where

$$H(X) = -\int p_x(x) \log p_x(x) dx \tag{21}$$

$$H(Y) = -\int p_y(y) \log p_y(y) dy \tag{22}$$

$$H(X,Y) = -\int\int p_{xy}(xy) \log p_{xy}(x,y) dx dy. \tag{23}$$

The mutual information is always non-negative, and it is zero if and only if X and Y are independent. The joint probability density function $p_{xy}(x,y)$ is required to estimate the mutual information between X and Y. When it is not available, usually, X and Y are quantized with a finite number of bins, and MI is approximated by a finite sum as follows:

$$I_{\text{quan}}(X,Y) = \sum_{ij} p_{xy}(i,j) \log \frac{p_{xy}(i,j)}{p_x(i) p_y(j)}. \tag{24}$$

When the sample size is infinite, and all bin sizes tend to zero, $I_{\text{quan}}(X,Y)$ converges to $I(X,Y)$. However, the amount of the data is usually limited, as in our experiments. Also, these methods are usually limited with the use of one or two dimensional variables, whereas here we use large dimensional vectors in the experiments. To avoid explicit estimation of the joint probability density function, we compute the mutual information using the method proposed in [55], which is based on entropy estimates from k-nearest neighbor distances. This method is data efficient, adaptive, and the MI estimate is numerically shown to be exact for the independent random variables. As suggested in [55], a midrange value for $k(k = 3)$ is used in the experiments.

First, the amount of redundancy in the gist features is analyzed. In Fig. 8, the pair-wise MI between all raw gist components (without PCA) extracted using 4-by-5 grids when the scene duration is 0.6 s is illustrated. Here, only the gist vector extracted from the feature map with center $c = 2$ and surround $s = 5$ scales is shown to make the figure more readable, so the gist vector dimension is $4 \times 5 = 20$ (each square block in the figure is 20×20). The MI $I(G_{X_i}, G_{Y_j})$ is computed for each pair (G_{X_i}, G_{Y_j}), where $G_X, G_Y \epsilon \{I, F, T, O_\theta, P_{O_\theta}, P_F\}$ and $i, j = \{1, \ldots, 20\}$. It can be observed that there are nonzero MI results. In the figure, the diagonal square blocks have nonzero

TABLE IV
MUTUAL INFORMATION BETWEEN THE GIST FEATURE VECTORS OF THE FEATURE SETS

	I	F	T	O_{45}	O_{135}	PO_{45}	PO_{135}	PF
I	x	5.0263	5.5761	3.6738	3.6722	1.3910	1.3796	1.2383
F	5.0263	x	3.9401	3.4344	3.7634	1.6117	1.6728	1.6398
T	5.5761	3.9401	x	3.4691	3.5711	1.3249	1.3285	1.1152
O_{45^o}	3.6738	3.4344	3.4691	x	3.3932	2.0453	1.4230	1.2214
O_{135^o}	3.6722	3.7634	3.5711	3.3932	x	1.5166	2.1615	1.2828
PO_{45^o}	1.3910	1.6117	1.3249	2.0453	1.5166	x	3.7945	2.7435
PO_{135^o}	1.3796	1.6728	1.3285	1.4230	2.1615	3.7945	x	2.9778
PF	1.2383	1.6398	1.1152	1.2214	1.2828	2.7435	2.9778	x

TABLE V
MI BETWEEN I, F, T, O FEATURES AND SYLLABLE PROMINENCE CLASS

Individual Feat.		Combined Features					
I	0.2368	TI	0.2636	TO	0.3149	IFT	0.2926
F	0.2596	IF	0.2710	IO	0.3096	TOI	0.3138
T	0.2502	FT	0.2900	FO	0.3207	IFO	0.3229
O	0.3102					FTO	0.3278
						$IFTO$	0.3300

TABLE VI
MI BETWEEN PITCH FEATURES AND SYLLABLE PROMINENCE CLASS

Individual Feat.		Combined Features			
P_F	0.2015	P_O	0.2650	$\mathbf{IFTO\&P_{OF}}$	**0.3490**
$P_{O_{45}}$	0.2323	P_{OF}	0.2700		
$P_{O_{135}}$	0.2163				

TABLE VII
PROMINENT SYLLABLE DETECTION PERFORMANCE WITH 0.6-s SCENE

Receptive Filter	1-by-v grids			4-by-5 grids		
	d	Acc.	F-sc	d	Acc.	F-sc
I	22	79.62%	0.69	21	80.92%	0.71
F	15	78.31%	0.66	20	81.79%	0.72
T	32	80.83%	0.71	25	81.77%	0.72
O	24	82.52%	0.73	46	84.34%	0.76
P	42	80.13%	0.68	54	81.26%	0.70
$IFTO$	58	84.62%	0.77	80	85.45%	0.78
$IFTOP$	60	84.79%	0.77	94	85.59%	0.78

MI, i.e., $I(G_{X_i}, G_{X_j}) \neq 0$ for some (i, j) pairs. In other words, as one can expect, the gist of a feature map has interdependencies with itself for each feature set since they are extracted using the same receptive filter. Also, there is redundancy across feature sets; i.e., $I(G_{I_i}, G_{F_j})$ is high for some (i, j) pairs. The comparison of redundancy across feature sets can be made more clearly from the MI results presented in Table IV. In Table IV, the mutual information between feature sets is computed using all the gist features extracted from all feature maps as a vector. In other words, the table holds the values $I(\mathbf{G_X}, \mathbf{G_Y})$, where $\mathbf{G_X}$ and $\mathbf{G_Y}$ are multicomponent vectors, and $\mathbf{X}, \mathbf{Y}$ are both $20 \times 3 = 60$ dimensional vectors since each set has three feature maps and each feature map generates a 20-dimensional gist vector. It can be concluded that the gist features extracted from the intensity feature set (I) are highly redundant with the ones extracted from the frequency (F) and the temporal contrast (T) feature sets. The orientation gist features (O_θ) have moderate redundancy with other feature sets' gist features. Finally, the gist features of pitch (P) feature sets have the least redundancy with the gist features of the remaining feature sets I, F, T, O_θ and more redundancy among themselves since they are limited to representing only the pitch characteristic of the scene. Due to the redundancy in the gist features, PCA is applied to the cumulative gist feature vector in the proposed model as shown in Fig. 1.

Next, the MI between the individual features and the prominence classes are computed to measure the amount of knowledge about syllable prominence provided by each feature. Hence, $I(\mathbf{G_X}, C)$ is estimated, where $\mathbf{G_X}$ is a multicomponent vector, $X \epsilon \{I, F, T, O_\theta, P\}$, and C is the syllable prominence class. The MI results are listed in Table V for I, F, T, O_θ. In Table VI, the MI values between pitch features and prominence are listed. The most informative pitch features about prominence are the ones captured with orientation RFs, $P_{O_{45}}$ and $P_{O_{135}}$. The contribution of P_F features is smaller compared to P_{O_θ}. These results are in agreement with the prominence detection results using only the pitch features which are reported in Table III. We can conclude from the results in Table III that the contribution of P_F features is significant when 1-by-v grid size is used. However, in the case of using 4-by-5 grid size, pitch values are roughly place coded (frequency resolution is higher) and P_F features do not contribute much to the prominence detection performance when combined with P_{O_θ} features.

We can conclude from the results in Table V and VI that, when the individual features are compared, the most informative feature about syllable prominence is the orientation (in the tables O represents the combinations of both directional orientations, i.e., it contains both O_{45} and O_{135}). Also, even though the features have across redundancy as listed in Table IV, adding each feature increases the amount of information on the syllable prominence, and the highest MI is obtained when all five features IFTOP are combined. These results are in agreement with the prominence detection results listed in Table VII. The individual feature which achieves the highest accuracy is the orientation with 84.34% accuracy when grid size is 4-by-5. The highest accuracy of 85.6% is achieved when all five features are combined. However, the performance achieved with IFTO and IFTOP features is not significantly different at $p \leq 0.001$.

The results obtained with 1-by-v grid size are also listed in Table VII. We can conclude that, usually, 4-by-5 grid size results in larger dimensional feature vectors after PCA, indicating having more diverse information about the scene. Also, the prominence results obtained with 4-by-5 grid size are significantly higher compared to the ones obtained with 1-by-v grid size at $p \leq 0.001$. In the remaining experiments, the scene duration for auditory gist feature extraction is set to $D = 0.6$ s, and the grid size is set as 4-by-5. Also, the combination of I, F, T, O_θ features is used; i.e., the pitch features are excluded in the rest of the experiments since IFTOP performance is not significantly different than the results of IFTO (refer to Table VII). The results are reported in Table VIII, and as

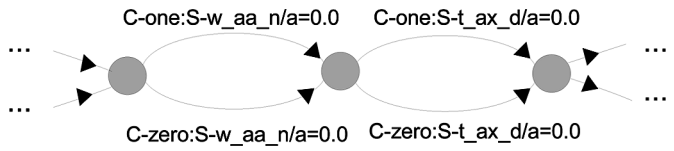

Fig. 9. Sausage lattice with only lexical evidence.

TABLE VIII
PROMINENT SYLLABLE DETECTION PERFORMANCE OF INDIVIDUAL ACOUSTIC, LEXICAL AND SYNTACTIC CUES

TD Evidence	Acc.	Pr.	Re.	F-sc.
Auditory Feat. only	85.45%	0.82	0.75	0.78
Lexical only	83.85%	0.77	0.76	0.76
Syntactic only (word)	82.50%	0.82	0.87	0.84
Syntactic only (syl.)	68.01%	0.54	0.53	0.53

reported earlier, an accuracy of 85.45% with an F-score of 0.78 is achieved for prominence detection task at the syllable level using only the auditory features. Table VIII also includes the results of the other types of top-down evidence which are discussed next. We first present the experimental results with the lexical and syntactic models in Section V-B and V-C, respectively, followed by the combined model results in Section V-D.

B. Top-Down Model Prediction With Lexical Evidence

The top-down model prediction using lexical information is implemented by creating sausage lattices for the test sets using the test transcriptions. The lattice arcs hold the syllable tokens together with the possible prominence class categories. For example, a part of a lattice that includes the word "wanted" is shown in Fig. 9. The arcs carry two syllables "w_aa_n" and "t_ax_d" the word contains together with prominent (C-one) and non-prominent (C-zero) class categories. When the only available evidence is the lexical rule, the arcs of the lattices do not carry any acoustic score, i.e., they are all set to zero ($\mathrm{a} = 0.0$ in Fig. 9). After constructing the lattice, it is scored with the factored n-gram lexical language model which was detailed in Section IV.B. The most likely prominence sequence is obtained by Viterbi decoding through the lattices. The results obtained with only the lexical model are reported in Table VIII; the prominence detection performance achieved with using only lexical information is 83.85% with an F-score of 0.76. We observe that the auditory features alone perform 1.6% better than the lexical features (85.45% versus 83.85%), and this result is significant at $p \leq 0.001$.

C. Top-Down Model Prediction With Only Syntactic Evidence

The class posterior probability $p(c_i|\mathbf{POS_i^L}(w_i))$ in (10) is computed using a neural network [53]. We use a three-layer neural network with d_{pos} inputs and n output nodes, where d_{pos} is the length of feature vector produced with POS tags, and $n = 2$ since this is a two-class problem. In our implementation, a set of 34 POS tags are used as those used in the Penn Treebank [56]. Each POS feature is mapped into a 34-dimensional binary vector. The neural network has $d_{\mathrm{pos}} = 34 \times 5$ inputs, since the syntactic information in our model includes the information from a window of $L = 5$ words.

As mentioned earlier, POS tags are associated with the words, so the neural network is trained using the word level POS tags.

TABLE IX
COMBINED TOP-DOWN MODEL PERFORMANCE FOR PROMINENT SYLLABLE DETECTION

TD Evidence	Acc.	Pr.	Re.	F-sc.
Auditory Feat. + Lexical	88.01%	0.83	0.82	0.82
Auditory Feat. + Syntactic	86.23%	0.81	0.79	0.80
Auditory Feat. + Syntactic + Lexical	**88.33%**	0.83	0.83	0.83
Combined Feat. word level	**85.71%**	0.87	0.86	0.87

Using only syntactic information, we achieve 82.50% accuracy ($\mathrm{F-score} = 0.84$) for the prominence detection task at the word level as detailed in Table VIII. Then, using (13) and (14), we convert word level posterior probability to the syllable level posterior probability. To obtain the baseline performance for the syntactic model, we combine prior chance level observed in the training data with the syntactic model posterior probabilities. The prominence detection accuracy achieved is 68.01% ($\mathrm{F-score} = 0.53$) at the syllable level using the syntactic evidence. This is slightly better than the chance level for the BU-RNC which is 65.7% accuracy at the syllable level. The syntactic features alone (68.01%) perform significantly worse than both auditory features (85.45%) and lexical features (83.83%) at the prominence detection task at the syllable level ($p \leq 0.001$). This is not surprising because of the fact that POS carries information at the word level. When a multisyllabic word is detected as prominent, there is no information about which syllable/s is/are prominent within the word. Hence, (13) and (14) are only approximations. Nevertheless, the combination of the syntactic information leads to a statistically significant performance improvement, as shown in the next section.

D. Combined Model With Auditory, Syntactic, Lexical Cues

We combine auditory gist features together with syntactic and lexical information using a probabilistic approach as presented in (18). First, the syllable level syntactic and auditory gist feature model outputs are combined and embedded in the lattice arcs. Then, the lattices are scored with the lexical language model, and a Viterbi search is conducted to find the best sequence of prominence labels. The combined model achieves 88.33% accuracy with an F-score of 0.83 as listed in Table IX.

In addition to these experiments, we also investigated the combination of auditory features together with lexical information and the combination of auditory features together with syntactic information. The results are summarized in Table IX. Incorporating syllable token information into the top-down prediction of auditory features leads to 2.56% ($p \leq 0.001$) accuracy improvement over the auditory features only model (88.01% versus 85.45%). Also, the improvement of 0.78% over the acoustical model prediction accuracy due to syntactic information is significant (86.23% versus 85.45%, $p \leq 0.001$). The performance difference between the model that includes all three models and the one that does not include the syntactic model is also significant at $p \leq 0.005$ (88.33% versus 88.01%). The best prominence detection performance accuracy is achieved with using all three information streams: auditory features and lexical and syntactic evidence. Finally, the combined model achieves 85.71% prominence detection accuracy at the word level with an $\mathrm{F-sc} = 0.87$.

TABLE X
PREVIOUSLY REPORTED RESULTS ON PROMINENCE DETECTION TASK USING THE BU-RNC

Previous Work	Features	Acc	d	Level
Wightman et al [57]	Acoustic + Prosodic LM	84.0%	12	syllable
Ross et al. [27] (single speaker)	Lex + Syn	87.7%	NA	syllable
Hirschberg [28]	Syn	82.4%	NA	word
Chen et al. [53]	Acoustic only	77.3 %	15	word
Chen et al. [53]	Acoustic + Syn	84.2 %	15	word
Ananthakrishnan et al. [50]	Acoustic only	74.1%	9	syllable
Ananthakrishnan et al. [50]	Acoustic+ Lex + Syn	86.1%	9	syllable
Our method	Acoustic only	85.45%	80	syllable
Our method	Acoustic + Lex + Syn	88.33%	80	syllable
Our method	Acoustic only	83.11%	80	word
Our method	Acoustic + Lex + Syn	85.71%	80	word

VI. CONCLUSION, DISCUSSIONS AND FUTURE WORK

In this paper, a novel model that combines bio-inspired auditory attention cues with higher level task-dependent lexical and syntactic cues was presented. The model was demonstrated to detect prominent syllables successfully in read speech with 85.67% accuracy using only acoustic cues, and 88.33% accuracy using acoustic, lexical and syntactic cues. The results compare well with human performance on stress labeling reported with BU-RNC: the average inter-transcriber agreement for manual annotators was 85–90% for presence versus absence of stress labeling [33].

It has been experimentally demonstrated with the auditory attention model that the prominence of syllables is affected by the neighboring syllables. The performance obtained with scenes that include approximately only the syllable itself was poor, i.e., short scenes with $D = 0.2$ s. Considering the performance and the computational cost, it is concluded that it is reasonable to have an analysis window duration of 0.6 s for the prominent syllable detection task.

The influence of higher level task-dependent rules due to lexical and syntactic knowledge was incorporated into the model. We can conclude from the experimental results summarized in Table IX that the contribution of lexical information is significant in the prominence detection task. The contribution of the syntactic cues captured with POS tags is small compared to that from lexical cues for the prominence detection at the syllable level. This might be mainly due to the fact that the POS tags used to represent syntactic information are associated with words. Hence, the syntactic model is accurate for the word level prominence detection, and its contribution to prominent syllable detection is limited to only detecting nonprominent words, and hence the nonprominent syllables the word contains. On the other hand, when a multisyllabic word is prominent, we don't have any information regarding which syllables within the word are prominent.

In Table X, we compare the performance of our model with results presented by other authors for this task using the BU-RNC database. The table shows the prominence detection accuracy obtained in these previously published papers, names of the features which were used for the experiments, the acoustic feature dimension (denoted as d in Table X), and the level at which the prominence detection experiments were performed (syllable or word). The set of possible features used in the literature are syntactic features (referred as Syn in Table X), lexical features (referred as Lex in Table X) and acoustic features (referred as Acoustic in Table X). Also, in [57] a prosodic bigram language model (LM) was used together with acoustic features. In the literature, all the previous work which uses acoustic cues for prominence detection utilizes prosodic features which consist of pitch, energy and duration features. Those acoustic feature dimensions vary between 9–15, whereas our model has a much higher acoustic feature dimension (80). However, the proposed acoustic only model performs significantly better than all the previously reported methods used only acoustic features, and it provides approximately 8%–10% absolute improvement over the previous results. In summary, we achieve significant performance gain but at the cost of computation. These results are especially beneficial for the cases where the utterance transcripts are not available in which syntactic and lexical cues cannot be easily extracted. The proposed auditory attention cues, however, perform sufficiently well even without text-derived features. Finally, when we combined all three lexical, syntactic and acoustic cues, our model performs 2.32% better than the previously reported results in [50]. Although, [27] reports 87.7% accuracy on this task, our results are not directly comparable since their experiments are limited to a single speaker, whereas we used the entire data set (six speakers) for the experiments.

The combined top-down task-dependent auditory attention model is used in this paper to detect prominent regions of speech. However, the prominence itself can actually be a feature that may attract human attention in a bottom-up manner. Incorporating prominence as a bottom-up attention cue into the current machine speech processing systems can be beneficial. As part of our future work, we are planning to integrate it into an automatic speech recognizer to improve the speech recognition accuracy.

One of the strengths of the proposed auditory attention model proposed here is that the model is a generic model, and it can be used in other spoken language processing tasks and general computational auditory scene analysis applications such as scene understanding, context recognition, and speaker recognition. Based on the selected application, first, optimal scene duration should be set. A finer grid size selection at the auditory gist feature extraction stage increases resolution and the computational cost. In the current paper, two grid sizes; i.e., 1-by-v and 4-by-5, are selected in an ad-hoc manner for the prominence detection task, and they performed sufficiently well. However, for new applications, an appropriate grid size needs to be found

considering the balance between resolution (hence the task performance) and the computational cost. Finding an optimal grid size is an open problem that we are planning to address as part of our future work.

The performance expected from the auditory attention model is limited by the five features used in the model. In other words, the model will fail to perform the tasks that require features which are not implemented here. For example, the current model uses mono signal, and hence spatial cues are not implemented in the model. As a result, while the model will successfully work for the tasks which are represented by at least one of the features of the model (i.e., intensity, frequency contrast, temporal contrast, and pitch), it will fail in the tasks which require spatial cues, such as localization and source separation.

It was shown with mutual information measurements that the raw auditory gist features extracted from intensity, temporal contrast, frequency contrast, orientations, and pitch features have redundancies. Hence, we applied PCA to the gist features to reduce redundancy and also to reduce feature dimension. PCA is optimal in least squares terms. PCA retains the components of the data set that contribute most to its variance, and it assumes that these components carry the most important aspects of the data. However, this might not be the case always. In the literature, there are examples of information maximization in neural codes [58]. Thus, as part of our future work, we plan to investigate information maximization criteria to select features.

REFERENCES

[1] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annu. Rev. Neurosci.*, vol. 18, no. 1, pp. 193–222, 1995.

[2] E. Weichselgartner and G. Sperling, "Dynamics of automatic and controlled visual attention," *Science*, vol. 238, no. 4828, pp. 778–780, 1987.

[3] S. Hochstein and M. Ahissar, "View from the top hierarchies and reverse hierarchies in the visual system," *Neuron*, vol. 36, no. 5, pp. 791–804, 2002.

[4] J. Jonides and S. Yantis, "Uniqueness of abrupt visual onset in capturing attention," *Percept Psychophys.*, vol. 43, no. 4, pp. 346–354, 1988.

[5] H. J. Muller and P. M. A. Rabbitt, "Reflexive and voluntary orienting of visual attention: Time course of activation and resistance to interruption," *J. Exper. Psychol.*, vol. 15, no. 2, pp. 315–330, 1989.

[6] W. James, *The Principles of Psychology*. New York: Dover, 1950, vol. 1.

[7] C. Alain and S. R. Arnott, "Selectively attending to auditory objects," *Front. Biosci.*, vol. 5, pp. d202–d212, 2000.

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

[9] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: Bottom-up versus top-down," *Current Biol.*, vol. 14, no. 19, pp. 850–852, 2004.

[10] A. Yarbus, "Eye movements during perception of complex objects," *Eye Movements and Vision*, pp. 171–196, 1967.

[11] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, pp. 975–979, 1953.

[12] E. R. Hafter, A. Sarampalis, and P. Loui, , W. Yost, Ed., "Auditory Attention and Filters," in *Auditory Perception of Sound Sources*. New York: Springer, 2007, vol. 29.

[13] D. H. Hubel, C. O. Henson, A. Rupert, and R. Galambos, "Attention units in the auditory cortex," *Science*, vol. 129, no. 3358, pp. 1279–1280, 1959.

[14] J. Moran and R. Desimone, "Selective attention gates visual processing in the extrastriate cortex.," *Science*, vol. 229, no. 4715, pp. 782–784, 1985.

[15] B. C. Motter, "Neural correlates of attentive selection for color or luminance in extrastriate area V4," *J. Neurosci.*, vol. 14, no. 4, pp. 2178–2189, 1994.

[16] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention-focusing the searchlight on sound," *Current Opinion Neurobiol.*, vol. 17, no. 4, pp. 437–455, 2007.

[17] C. Kayser and N. Logothetis, "Vision: Stimulating your attention," *Current Biol.*, vol. 16, no. 15, pp. R581–R583, 2006.

[18] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying circuitry," *Hum. Neurobiol.*, vol. 4, pp. 219–227, 1985.

[19] J. M. Wolfe, "Guided search 2.0: A revised model of guided search," *Psychonomic Bull. Rev.*, vol. 1, no. 2, pp. 202–238, 1994.

[20] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimal object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition (CVPR)*, New York, 2006, pp. 2049–2056.

[21] R. J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition (CVPR)*, Minneapolis, MN, 2007, pp. 1–8.

[22] S. Shamma, "On the role of space and time in auditory processing," *Trends Cogn. Sci.*, vol. 5, pp. 340–348, 2001.

[23] A. Johnson and R. W. Proctor, *Attention: Theory and Practice*. Newbury Park, CA: Sage, 2004.

[24] C. Kayser, C. Petkov, M. Lippert, and N. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Current Biol.*, vol. 15, no. 8, pp. 1943–1947, 2005.

[25] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 1941–1944.

[26] B. Lindblom, S. Brownlee, B. Davis, and S. J. Moon, "Speech transforms," *Speech Commun.*, vol. 11, no. 4–5, pp. 357–368, 1992.

[27] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Comput. Speech Lang.*, vol. 10, no. 3, pp. 155–185, 1996.

[28] J. Hirschberg, "Pitch accent in context: Predicting intonational prominence from text," *Artif. Intell.*, vol. 63, no. 1-2, pp. 305–340, 1993.

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

[30] I. Bulyko and M. Ostendorf, "Joint prosody prediction and unit selection for concatenative speech synthesis," in *Proc. ICASSP*, 2001, vol. 2, pp. 781–784.

[31] M. Ostendorf, I. Shafran, and R. Bates, "Prosody models for conversational speech recognition," in *Proc. 2nd Plenary Meeting Symp. Prosody Speech Process.*, 2003, pp. 147–154.

[32] M. Hasegawa-Johnson *et al.*, "Simultaneous recognition of words and prosody in the Boston university radio speech corpus," *Speech Commun.*, vol. 46, no. 3–4, pp. 418–439, 2005.

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

[34] O. Kalinli and S. Narayanan, "A top-down auditory attention model for learning task dependent influences on prominence detection in speech," in *Proc. ICASSP*, Las Vegas, NV, Apr. 2008, pp. 3981–3984.

[35] O. Kalinli and S. Narayanan, "Combining task-dependent information with auditory attention cues for prominence detection in speech," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, pp. 1064–1067.

[36] K. Silverman, M. Beckman, J. Pierrehumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg, "ToBI: A standard scheme for labeling prosody," in *Proc. Int. Conf. Spoken Lang. Process.*, 1992, pp. 867–870.

[37] D. Kahn, "Syllable-based generalizations in English phonology," Ph.D. dissertation, Univ. Mass., Boston, MA, 1976.

[38] X. Yang, K. Wang, and S. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 824–839, Feb. 1992.

[39] R. C. deCharms, D. T. Blake, and M. M. Merzenich, "Optimizing sound features for cortical neurons," *Science*, vol. 280, pp. 1439–1443, 1998.

[40] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, pp. 1395–1407, 2006.
[41] P. Ru, "Multiscale multirate spectro-temporal auditory model," Ph.D. dissertation, Univ. of Maryland, College Park, 2001.
[42] D. Bendor and X. Wang, "The neuronal representation of pitch in primate auditory cortex," *Nature*, vol. 436, pp. 1161–1165, 2005.
[43] M. Slaney and R. F. Lyon, "On the importance of time-a temporal representation of sound," *Visual Representations of Speech Signals*, pp. 95–116, 1993.
[44] C. E. Schreiner, H. L. Read, and M. L. Sutter, "Modular organization of frequency integration in primary auditory cortex," *Annu. Rev. Neurosci.*, vol. 23, pp. 501–529, 2000.
[45] A. Oliva, , L. Itti, G. Rees, and J. K. Tsotsos, Eds., "Gist of a scene," in *Neurobiology of Attention*. New York: Elsevier, 2005, pp. 251–256.
[46] D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cogn. Psychol.*, vol. 9, pp. 353–383, 1977.
[47] S. Harding, M. P. Cooke, and P. Koenig, "Auditory gist perception: An alternative to attentional selection of auditory streams," in *Proc. WAPCV*, Hyderabad, India, 2007, pp. 399–416.
[48] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 300–312, Feb. 2007.
[49] N. Moray, "Attention in dichotic listening: Affective cues and the influence of instructions," *Quart. J. Exper. Psychol.*, vol. 11, no. 1, pp. 56–60, 1959.
[50] S. Ananthakrishnan and S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 216–228, 2008.
[51] K. Kirchhoff, J. Bilmes, and K. Duh, "Factored language models tutorial," Dept. of Elect. Eng., Univ. of Washington, Tech. Rep. UWEETR-2007-0003, Jun. 2007.
[52] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Proc. Int. Conf. Spoken Lang. Process.*, Sep. 2002, pp. 901–904.
[53] K. Chen, M. Hasegawa-Johnson, A. Cohen, and J. Cole, "A maximum likelihood prosody recognizer," in *Proc. Speech Prosody*, Nara, Japan, 2004, pp. 509–512.
[54] R. Lowry, "Vassarstats: Wilcoxon Signed-Rank Test," [Online]. Available: http://faculty.vassar.edu/lowry/wilcoxon.html
[55] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, no. 6, p. 66138, 2004.
[56] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," *Comput. Ling.*, vol. 19, no. 2, pp. 313–330, 1994.
[57] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 469–481, Jul. 1994.
[58] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.*, vol. 24, no. 1, pp. 1193–1216, 2001.

Ozlem Kalinli (S'08) received the B.S. degree (summa cum laude and with honors) in electronics and communication engineering from Istanbul Technical University (ITU), Istanbul, Turkey, in 2001, and the M.S. degree (with outstanding academic achievement) in electrical engineering from Illinois Institute of Technology (IIT), Chicago, in 2003. She is currently pursuing the Ph.D. degree in electrical engineering at the University of Southern California (USC), Los Angeles.

She was a Member of the Immersive Audio Laboratory at USC from 2003 to 2005. Since 2005, she has been a Member of the Signal Analysis and Interpretation Laboratory (SAIL), USC. Her current research interests include bio-inspired signal processing for speech and audio applications, auditory attention, auditory perception, speech recognition, speech analysis, and machine learning for speech and audio applications.

Shrikanth Narayanan (S'88-M'95-SM'02–F'09) received the Ph.D. degree in electrical engineering from the University of California (USC), Los Angeles, in 1995.

He is Andrew J. Viterbi Professor of Engineering at USC, Los Angeles, where he holds appointments as Professor in electrical engineering and jointly in computer science, linguistics, and psychology. Prior to joining USC, he was with AT&T Bell Labs and AT&T Research, first as a Senior Member, and later as a Principal Member of its Technical Staff from 1995 to 2000. At USC, he is a member of the Signal and Image Processing Institute and directs the Signal Analysis and Interpretation Laboratory (SAIL). He has published over 300 papers and has seven U.S. patents.

Dr. Narayanan is a recipient of an NSF CAREER award, a USC Engineering Junior Research Award, a USC Electrical Engineering Northrop Grumman Research Award, a Provost fellowship from the USC Center for Interdisciplinary research, a Mellon Award for Excellence in Mentoring, an IBM Faculty award, an Okawa Research award, and a 2005 Best Paper award from the IEEE Signal Processing society (with Alex Potamianos). Papers with his students have won best paper awards at ICSLP'02, ICASSP'05, MMSP'06, and MMSP'07. He is an Editor for the *Computer Speech and Language Journal* (2007-present) and an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA. He was also an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (2000–2004) and the *IEEE Signal Processing Magazine* (2005–2008). He served on the Speech Processing technical committee (2005–2008) and Multimedia Signal Processing technical committees (2004–2008) of the IEEE Signal Processing Society and presently serves on the Speech Communication committee of the Acoustical Society of America and the Advisory Council of the International Speech Communication Association. He is a Fellow of the Acoustical Society of America and a member of Tau-Beta-Pi, Phi Kappa Phi, and Eta-Kappa-Nu.