



A Saliency-Based Auditory Attention Model with Applications to Unsupervised Prominent Syllable Detection in Speech

Ozlem Kalinli and Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
 Department of Electrical Engineering-Systems
 University of Southern California
 3750 McClintock Avenue, EEB 400, Los Angeles, California 90089.
 kalinli@usc.edu, shri@sipi.usc.edu

Abstract

A bottom-up or saliency driven attention allows the brain to detect nonspecific conspicuous targets in cluttered scenes before fully processing and recognizing the targets. Here, a novel biologically plausible auditory saliency map is presented to model such saliency based auditory attention. Multi-scale auditory features are extracted based on the processing stages in the central auditory system, and they are combined into a single master saliency map. The usefulness of the proposed auditory saliency map in detecting the prominent syllable and word locations in speech is tested in an unsupervised manner. When evaluated with broadcast news-style read speech using the BU Radio News Corpus, the model achieves 75.9 % accuracy at the syllable level, and 78.1 % accuracy at word level. These results compare well to results reported on human performance.

Index Terms: auditory attention, auditory saliency map, prominent syllable detection, attention model.

1. Introduction

The brain is the most advanced information processing device, and a large portion of its computation power is devoted to sensory processing with vision and hearing being the two highly developed senses in humans. In [1], the common principles of visual and auditory processing are discussed, 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.

Our nervous system is exposed to tremendous amount of sensory stimuli, but our brain cannot fully process all stimuli at once. A neural mechanism exists that selects a subset of available sensory information before further processing it [2, 3, 4]. This selection is a combination of rapid bottom-up saliency-driven (task-independent) attention, as well as slower top-down cognitive (task dependent) attention [2]. First, stimulus-driven rapid bottom-up processing of the whole scene occurs that attract attention towards conspicuous or salient locations in an unconscious manner. Then, the top-down processing shifts the attention voluntarily towards locations of cognitive interest. Only the selectively attended location is allowed to progress through cortical hierarchy for high-level processing to analyze the details [2, 4, 5]. In vision, for example, for an observer a red circle in a gray background will be salient (bottom-up, saliency-driven analysis), but after consciously paying attention to the red spot it will be aware that the red spot is actually a traffic sign (top-down analysis). Similarly, in audition, one may hear people talking, music playing in a room (saliency-driven), but it won't be immediately apparent what people are saying or what type of instruments are producing the music. Only if the subject chooses to listen the music, s/he will be aware of what kinds of instruments are producing the music (top-down).

In [2, 6], the concept of saliency map was proposed to understand bottom-up visual attention in primates. A set of low-level features are extracted in parallel from the image in multi-scale 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 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 attention, i.e. detecting traffic signs, detecting colors etc. [2]. Analogous to visual saliency maps, a saliency map for audition was proposed in [7]. The structure of the saliency map was identical to the visual saliency map in [2]. This model was able to replicate basic properties of auditory scene perception, i.e. the relative salience of short, long, temporally modulated tones in noisy backgrounds [7]. These results clearly support the hypothesis that the mechanisms extracting conspicuous events from a sensory representation are identical in the central auditory and visual systems, and bottom-up human attention can be modelled with a saliency map.

In this paper, we propose a novel biologically plausible auditory saliency map that builds on the architectures proposed in [2, 7]. The contributions of this work are as follows: An auditory spectrum of the sound is first computed based on early stages of human auditory system. This two-dimensional (2D) spectrum is processed by the auditory saliency model. In addition to the intensity, temporal and frequency contrast features used in [7], orientation features extracted analogous to the dynamics of the auditory neuron responses to moving ripples in the primary auditory cortex, and pitch which is a fundamental percept of sound are included in the model as well. To integrate the different features into a single saliency map, a biologically inspired nonlinear local normalization algorithm is used. The normalization algorithm is adapted from the model proposed for vision in [8] to a plausible model for auditory system. The proposed auditory saliency map is tested in the context of a prominent syllable detection task in speech. The motivation behind choosing prominent syllable detection task is 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 the accent, and physical and emotional state of the talker [5]. This information encoded in the acoustical signal is perceived by the listeners, and we propose to detect these salient syllable locations using the proposed bottom-up auditory attention model. The experimental results show that the proposed auditory saliency map detects the prominent syllables and words in speech with 75.9% and 78.1% accuracy, respectively, in an unsupervised manner.

The paper is organized as follows: first the auditory saliency map model is explained in Section 2, followed by experimental results in Section 3. The conclusions drawn and possible future work presented in Section 4.

This research was supported by NSF, ONR and Army.

2. Auditory Saliency Model

The block diagram of the proposed auditory saliency model is given in Fig 1. First, the auditory spectrum of the sound is estimated based on the information processing stages in the early auditory (EA) system [1]. 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 [1]. The input signal is filtered with a bank of 128 overlapping constant-Q asymmetric band-pass filters with center frequencies that are uniformly distributed along a 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 [1]. Here, for analysis, audio frames of 20 ms with 10 ms shift are used, i.e. each 10 ms audio frame is represented by a 128 dimensional vector.

The output of the EA model is an *auditory spectrum* with time and frequency axes, which is analogous to the input image in visual saliency map. In the next stage, this spectrum is analyzed by extracting a set of multi-scale features that are similar to the information processing stages in the central auditory system. Intensity, frequency contrast, temporal contrast and orientation features are extracted using spectro-temporal receptive filters mimicking the analysis stages in the primary auditory cortex [1, 9]. Pitch is included in our model, because it is an important property of sound and recent functional imaging studies showed that the neurons of the auditory cortex also respond to pitch [10].

2.1. Features

All the receptive filters 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 (RF) in the auditory cortex with only an excitatory phase selective for a particular region [9], and can be implemented with a Gaussian kernel [11]. The multi-scale intensity features $I(\sigma)$ are created using a dyadic pyramid: the input spectrum is filtered, and decimated by a factor of two, and this is repeated. Finally, eight scales $\sigma = \{1, \dots, 8\}$ are created, yielding size reduction factors ranging from 1:1 (scale 1) to 1:128 (scale 8).

The frequency contrast filters correspond to RF with an excitatory phase and simultaneous symmetric inhibitory side bands and the temporal contrast filters correspond to RF with an inhibitory phase and a subsequent excitatory phase as described in [7, 9], and they are shown in Fig 2. These filters are implemented using a 2D Gabor filter (product of a cosine function with 2D Gaussian envelope [11]) with orientation $\theta = 0^\circ$ for frequency contrast $F(\sigma)$ and $\theta = 90^\circ$ 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 30 ms excitation phase flanked by 20 ms inhibition phase. The orientation filters mimic the dynamics of the auditory neuron responses to moving ripples [1, 9]. To extract orientation features $O_\theta(\sigma)$, 2D Gabor filters with $\theta = \{45^\circ, 135^\circ\}$ 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 it can manifest the lateral inhibition structure, i.e. an excitatory phase with simultaneous symmetric inhibitory sidebands [12]. Similar to $I(\sigma)$, the multi-scale $F(\sigma)$, $T(\sigma)$, $O_\theta(\sigma)$ features are extracted using the filters described above on eight scales each being a resampled version (factor 2) of the previous.

In general, there are two hypotheses for the encoding of pitch in the auditory system: temporal and spectral [1]. We extract pitch based on the temporal hypothesis which assumes that the brain estimates the periodicity of the waveform in each au-

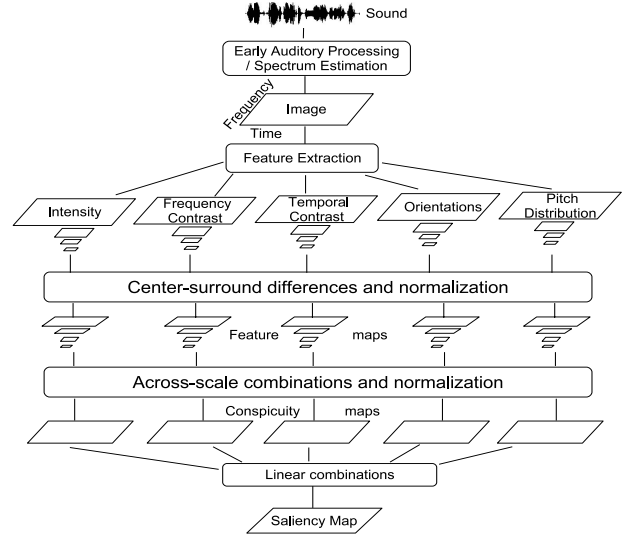


Figure 1: Auditory saliency map structure adapted from [2]

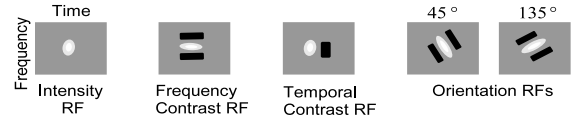


Figure 2: Receptive Filters

ditary nerve fiber by autocorrelation [1]. 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 multi-scale pitch features $P(\sigma)$ are created using a dyadic Gaussian pyramid identical to the one used for extracting intensity features.

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 [2, 13]:

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

Here, $c = \{2, 3, 4\}$, $s = c + \delta$ with $\delta \in \{3, 4\}$ are used. In total we have, 36 feature maps computed: six for each intensity, frequency contrast, temporal contrast, pitch and twelve for orientation since it has two angles $\theta = \{45^\circ, 135^\circ\}$.

The feature maps are combined to provide bottom-up input to the saliency map. However, the maps have to be normalized since they represent non-comparable modalities, i.e. different dynamic ranges and feature extraction mechanisms. An iterative nonlinear normalization algorithm $\mathcal{N}(\cdot)$ is used to normalize the feature maps (discussed in detail in Section 2.2). The normalized feature maps are combined into “conspicuity maps” at scale $\sigma = 3$ using across scale addition \oplus [2]:

$$\bar{\mathcal{M}} = \bigoplus_{c=2}^4 \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{M}(c, s)) \quad \mathcal{M} \in \{I, F, T, P\} \quad \text{and} \quad (2)$$

$$\bar{O} = \sum_{\theta \in \{45^\circ, 135^\circ\}} \mathcal{N} \left(\bigoplus_{c=2}^4 \bigoplus_{s=c+3}^{c+4} \mathcal{N}(O(c, s)) \right) \quad (3)$$

Finally, the auditory saliency map is computed by combining the normalized conspicuity maps with equal weights:

$$S = \frac{1}{5} (\mathcal{N}(\bar{I}) + \mathcal{N}(\bar{F}) + \mathcal{N}(\bar{T}) + \mathcal{N}(\bar{O}) + \mathcal{N}(\bar{P})) \quad (4)$$

The maximum of the saliency map defines the most salient location in 2-D time-frequency auditory spectrum.

2.2. Iterative Nonlinear Normalization

The normalization algorithm used in the auditory saliency map proposed in [7] is a global nonlinear amplification algorithm. This operation strongly promotes the maps with small number of strong peaks of activity, while globally suppressing maps with many comparable peaks. It, however, has several drawbacks: first, it has a strong bias towards enhancing the feature maps which have a unique location that is significantly more conspicuous or salient than the others [8]. For example, this normalization algorithm would suppress a map with two equally strong conspicuous locations, and otherwise no activity, while a human would usually report that both locations are salient [8]. Second, this normalization algorithm is not robust to noise [8].

The normalization $\mathcal{N}(\cdot)$ algorithm used here is an iterative, nonlinear operation simulating competition between the neighboring salient locations. It was originally proposed in [8] for visual saliency map, and is adapted to the auditory system here. Each feature map is first scaled to the range $[0, 1]$ to eliminate the dynamic-range modality. Then, each iteration step consists of a self-excitation and inhibition induced by neighbors. This is implemented by convolving each map with a large 2D difference of Gaussians (DoG) filter, and clamping the negative values to zero [8]. A feature map \mathcal{M} is transformed in each iteration step as follows:

$$\mathcal{M} \leftarrow |\mathcal{M} + \mathcal{M} * \text{DoG} - C_{\text{inh}}| \geq 0 \quad (5)$$

where, C_{inh} is 2% of the global maximum of the map. The details of DoG filter parameters can be found in [8], except that the filter size is modified as follows.

The visual saliency model operates only on spatial domain, while the auditory saliency map consists of temporal and frequency domain. Therefore, this requires a different normalization process for auditory model especially for temporal domain. The normalization algorithm in [8] uses the same filters for both (x, y) axes since they are both spatial domains. First, we modify this part and design the temporal and frequency filters separately. The cortical neurons along the cochlea are connected locally [12], hence only neighboring basilar membrane filter outputs can inhibit each other. Based on this fact, a DoG filter that operates on the frequency domain (y axes) is designed such that a single frequency channel output is self-excited, and inhibited by the two lower and upper channel outputs next to it.

In [7], the auditory saliency model uses a 0.45 s analysis window based on the temporal masking facts in the auditory system. Also, it is shown that for the prominent syllable detection task an analysis window centered in the syllable nuclei encompassing the neighboring syllables (approximately 0.5 s window size) performs well [14]. To design the temporal DoG filter, we derived the statistics of the database used for prominent syllable task to get an estimate. It is found that the mean syllable duration is approximately 0.2 s (μ) with 0.1 s standard deviation (std). Hence, the temporal DoG filter used for normalization is implemented such that it comprises an excitation phase of approximately 0.2 s, followed and preceded by 0.2 s inhibitory regions (considering neighboring syllables), yielding a 0.6 s analysis window. There are also syllables with duration much larger than 0.2 s. For instance, the maximum syllable duration is 1.4 s in the database. Hence, C_{inh} is computed over a 3 s audio stream during normalization to take into account longer syllables as well. The normalization filter duration is further analyzed in Section 3.

3. Experiments and Results

To test our auditory saliency model, the Boston University Radio News Corpus (BU-RNC) database [15] was used in the experiments. The BU-RNC is a broadcast news-style read speech corpus that consists of speech from 7 speakers (3 females and 4 males), totaling about 3 hours of acoustic data. A significant portion of the data has been manually labelled with prosodic tags. The database also contains the orthography corresponding

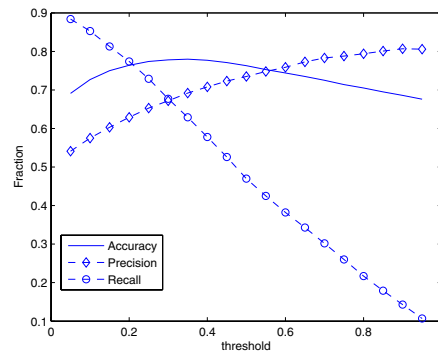


Figure 3: Prominent syllable detection performance vs. threshold (with IFTO)

to each spoken utterance together with time alignment information at the syllable and word level. We mapped all the pitch accent types (H*, L*, L*+H, etc..) to a single stress label since it is not our interest to detect what type of pitch accent produces the stress. Hence, the syllables annotated with any type of pitch accent was labelled “prominent”, and otherwise “non-prominent”. Also, we derived word level prominence tags from the syllable level prominence tags. The words that contain one or more prominent syllables are labelled as prominent, non-prominent otherwise. The prominent syllable fraction in the BU-RNC corpus is 34.3% (chance level), and 54.3% (chance level) is the prominent word fraction. We chose this database for two main reasons: i) syllables are stress labelled based on human perception ii) it allows an easy, concrete evaluation of our algorithm since stress labels and time alignment information are available.

The maximum of the saliency map defines the most salient location in 2-D auditory spectrum. In vision, the visual saliency map is scanned sequentially to find the locations in the order of decreasing saliency. However, there is neither available saliency ranking for prominent syllables, nor is there information regarding the frequency location that makes the syllable prominent at that time point. Here, we assume that saliency combines additively across frequency channels. The saliency map for the given sound frame is first scaled back to the original size (scale 1), and then summed across frequency channels for each time point, and normalized to $[0, 1]$ range, yielding a saliency score $S(t)$ for each time point t . The local maxima of $S(t)$ which are above a threshold (th) are found, and the syllable at the corresponding time point is marked as prominent. The threshold is varied from 0.05 to 0.95, and Fig. 3 shows the variation of prominent syllable detection performance as a function of the threshold. As expected, for increasing threshold, the precision rate increases, whereas the recall rate decreases. It is clear that the performance is not sensitive to selection of threshold value, i.e. accuracy doesn’t change dramatically for varying thresholds, and it is well above chance level for all threshold values. Any threshold value between 0.1 and 0.3 can be considered reasonable. It is also important to note that more than 80% of the “most salient” locations, i.e. locations marked salient with $th > 0.9$, correspond to an actual prominent syllable (for $th > 0.9$, precision > 0.80). This supports the observation that prominent syllables attract auditory attention.

The contribution of each feature to the prominent syllable detection task is examined, and accuracy (Acc), precision (Pr), recall (Re), and F-score (Fs) are reported in Table 1. The reliability measures in Table 1-2 are obtained after averaging across the threshold values between 0.1 and 0.3. The initial letter of the feature names is used in Tables and Figures to denote the corresponding conspicuity map, i.e. I=Intensity, F=Feature contrast, T=temporal contrast, O=Orientation, P=Pitch. The combina-

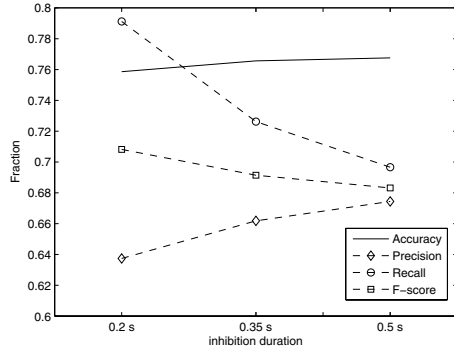


Figure 4: Prominent syllable detection performance vs. inhibition window size (with IFTO features)

Table 1: Prominent Syllable Detection Performance

I:Intensity, T:Temporal contrast, F:Frequency contrast, O:Orientation, P:Pitch

Features	Acc.	Pr	Re	Fs
I	74.7%	0.63	0.74	0.68
P	65.9%	0.51	0.85	0.66
IFT	75.2%	0.63	0.77	0.69
IFTO	75.9%	0.64	0.79	0.71
IFTOP	73.0%	0.59	0.82	0.69

Table 2: Prominent Word Detection Performance

Features	Acc.	Pr	Re	Fs
IFTO	78.1%	0.78	0.86	0.82

tion of letters indicates the conspicuity maps that contribute to the saliency map in Eq 4. The best performance is 75.9% accuracy with an F-score=0.71, and obtained when the auditory saliency map consisted of I, F, T and O features (IFTO), for the prominent syllable detection task. Even though pitch is an important prosodic cue, the performance obtained with only pitch feature (P) is low (Acc=65.9%), and when it is combined with the rest of the features, it also causes performance degradation (IFTOP performs worse than IFTO). This can be due to two reasons: i) even though the auditory experiments show that human perceive pitch, where/how in the brain pitch is computed is ambiguous [1], so the pitch feature may not be modelled correctly in the proposed framework ii) as the findings of study in [14], loudness (or intensity here) predicts the syllable prominence, and pitch does not contribute much for syllable prominence task. The word prominence performance is also evaluated similarly, and it is summarized in Table 2. We achieved 78.1% accuracy with an F-score=0.82 for the word prominence task.

The inhibition duration used in the iterative normalization step of the method was investigated. The inhibition duration is varied from 0.2 s (μ) to 0.5 s ($\mu + 3 \times std$) resulting a normalization window size from 0.6 s to 1.2 s (0.2 s excitation phase is preceded and followed by inhibition phase considering the previous and next syllable). The results are presented in Fig. 4. While the accuracy does not change significantly for inhibition duration, precision increases and recall decreases with increasing inhibition duration. This shows that when inhibition duration increases, the normalization algorithm promotes the most salient location and suppresses less conspicuous locations (i.e. precision increases).

These results are encouraging given that the average inter-transcriber agreement for manual annotators is 80-85% for stress labelling [15]. The results also compare well against the previously reported performance levels for unsupervised prosody labeling with the BU-RNC database, i.e. in [16], the unsupervised method obtained 77% accuracy at the syllable level using acoustical, lexical, and syntactic features.

4. Conclusion and Future Work

In this paper, an auditory saliency map based on a model of bottom-up stimuli driven auditory attention is presented. A set of auditory features are extracted in parallel from the auditory spectrum of the sound, and fed into a master saliency map in a bottom-up manner. This structure provides fast selection of conspicuous events in an acoustical scene, which can be further analyzed by more complex and time-consuming processes. The model could successfully detect the prominent syllable and word locations in read speech with 75.9% and 78.1% accuracy, respectively. One advantage of this attention model is that it is language independent, and can detect the prominent syllables in an unsupervised manner. The auditory saliency model proposed here is not only limited to prosody labelling. For example, it can be used in general computational auditory scene analysis (CASA) applications to select conspicuous events rapidly. Similar to the selective attention in humans [4], after a conspicuous location is selected (focused), it can be analyzed further to recognize the details of the object.

In this work, features are combined with equal weights to create the saliency map. As part of our future work, the weights will be learned in a supervised fashion for different types of auditory tasks, i.e. general audio scene analysis, spoken language processing etc. This can provide insights into what types of cues human brain uses while pre-attending to events in a given task.

5. References

- [1] S. Shamma, "On the role of space and time in auditory processing," *Trends Cogn. Sci.*, vol. 5, pp. 340–348, 2001.
- [2] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [3] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sounds*. London: The MIT Press, 1990.
- [4] C. Alain and S. R. Arnott, "Selectively attending to auditory objects," *Front. Biosci.*, vol. 5, pp. d202–212, 2000.
- [5] S. Harding, M. P. Cooke, and P. Koenig, "Auditory gist perception: An alternative to attentional selection of auditory streams," in *WAPCV2007*, India, 2007.
- [6] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying circuitry," *Hum. Neurobiol.*, vol. 4, pp. 219–227, 1985.
- [7] C. Kayser, C. Petkov, M. Lippert, and N. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Current Biology*, vol. 15, no. 8, pp. 1943–1947, 2005.
- [8] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *J. of Electronic Imaging*, vol. 10, pp. 161–169, January 2001.
- [9] R. C. deCharms, D. T. Blake, and M. M. Merzenich, "Optimizing sound features for cortical neurons," *Science*, vol. 280, pp. 1439–1443, 1998.
- [10] D. Bendor and X. Wang, "The neuronal representation of pitch in primate auditory cortex," *Nature*, vol. 436, pp. 1161–1165, 2005.
- [11] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, pp. 1395–1407, 2006.
- [12] P. Ru, "Multiscale multirate spectro-temporal auditory model," *Ph.D Dissertation*, 2001.
- [13] 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.
- [14] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: fundamental frequency lends little," *J. Acoust. Soc. Am.*, vol. 118, 2005.
- [15] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, *The Boston University Radio Corpus*, 1995.
- [16] S. Ananthakrishnan and S. Narayanan, "Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling," in *ICSLP*, September 2006.