A TOP-DOWN AUDITORY ATTENTION MODEL FOR LEARNING TASK DEPENDENT INFLUENCES ON PROMINENCE DETECTION IN SPEECH

Ozlem Kalinli and Shrikanth Narayanan

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

ABSTRACT

A top-down task-dependent model guides attention to likely target locations in cluttered scenes. Here, a novel biologically plausible top-down auditory attention model is presented to model such task-dependent influences on a given task. First, multi-scale features are extracted based on the processing stages in the central auditory system, and converted to low-level auditory "gist" features. These features capture rough information about the overall scene. Then, the top-down model learns the mapping between auditory gist features and the scene categories. The proposed top-down attention model is tested with prominent syllable detection task in speech. When tested on broadcast news-style read speech using the BU Radio News Corpus, the model achieves 85.8% prominence detection accuracy at syllable level. The results compare well to the reported human performance on this task.

Index Terms— auditory attention, auditory gist, prominence detection, stress detection, accent.

1. INTRODUCTION

The nervous system is exposed to tremendous amount of sensory stimuli, but the brain cannot fully process all stimuli at once. A neural mechanism exists that selects a subset of available sensory information before fully processing it [1, 2]. This selection is a combination of rapid bottom-up (task-independent) attention, as well as slower top-down (task dependent) attention [1]. First, a stimulus-driven bottom-up process of the whole scene attracts attention towards conspicuous or salient locations in an unconscious manner. Then, the top-down processing shifts the attention voluntarily towards locations of cognitive interest. Only the selected location is allowed to progress through the cortical hierarchy for high-level processing to analyze the details [2].

The bottom-up saliency-driven attention detects the objects that perceptually stand out of a scene by significantly differing from their neighbors. For instance, consider detection of a short tone burst in silent/noisy background. However, the top-down task-relevant process uses 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 the same scene [3]. The gaze of the observer fell on faces when estimating the people's age, but fell on clothing when estimating the people's material conditions. Since vision and audition have similar neural processing stages and perceptual behavior [4, 5], one can expect similar task-dependent influences in audition, as well. For example, in a cocktail party problem setting, the attention of the subject may shift to the speech sound if the task is "who is speaking, what?", while the attention may shift to the music if the task is "which instruments are being played?".

As stated previously, the task-independent bottom-up attention finds the locations where there is a target/source that pops-out perceptually. For example, in our previous work [6], the proposed bottom-up auditory attention model could detect prominent syllables in speech. However, when humans are asked to find the prominent (stressed) syllable, they also use their prior task-relevant knowledge to pick among the conspicuous locations. The motivation for this work is to analyze the effect of top-down *task-dependent* influences on the auditory attention for a given task.

The top-down model proposed here is based on the "gist" phenomenon commonly studied in vision. Gist processing is a pre-attentive process and guides attention to focus into particular subset of stimuli locations to analyze the details of the target locations [2, 5]. 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. For example, after very brief exposure of a scene, a subject can report general attributes of the scene, i.e., whether it was indoors, outdoors, kitchen, street traffic etc. In [7], a computational model that captures the gist of an image into a low-level signature vector is proposed, and used for classification of outdoor scenes. In [5], a review of gist perception is presented, and it is argued that gist perception also exists in audition.

In this paper, we propose a novel biologically plausible top-down model which guides attention during acoustical search for a target. The feature extraction is accomplished by sharing the same front-end with the bottom-up auditory attention model proposed in [6], since it is based on the processing stages in the primary auditory cortex. First, an auditory spectrum of the sound is computed based on early stages of human auditory system. This two-dimensional (2D) time-frequency spectrum is akin to an image of a scene in vision. Then, multi-scale features are extracted from the spectrum based on the processing stages in the central auditory system, and converted to low-level auditory gist features. Finally, by accumulating the statistics of the gist features, the top-down model learns to associate a given gist feature set with likely scene categories, i.e., for the current task, scene categories are prominent vs. non-prominent syllables. It should be noted that the proposed top-down auditory attention model is a generic model with a variety of applications, i.e., speaker recognition, scene change detection, context recognition etc. Here, we apply it to the prominent syllable detection problem, and the experimental results show that the proposed model detects prominent syllables in speech with 85.8% accuracy, and provides approximately 10% absolute improvement over using just the bottom-up attention model.

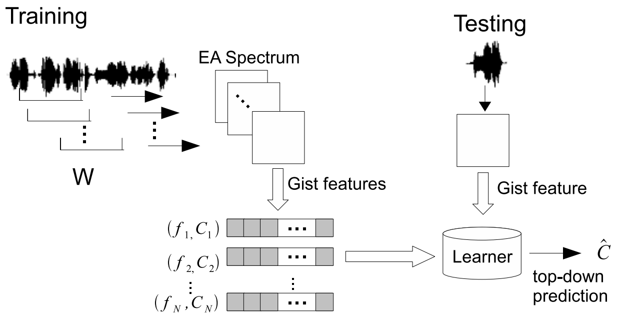

Fig. 1. Top-down task-dependent model structure

The paper is organized as follows: the top-down auditory attention model with gist feature extraction is explained in Section 2. This is followed by the details of experiments in Section 3, and the results in Section 4. The conclusions and future work are presented in Section 5.

2. TOP-DOWN TASK-DEPENDENT MODEL

The top-down model with gist features is illustrated in Fig. 1. To learn top-down task-dependent influences on a given task, we split the data 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 class categories C_i. Here, the term "*scene*" represents an audio segment of duration W. The details of selecting an appropriate window duration W will be discussed later. The features are stacked and passed through a "learner" (a machine learning algorithm) to discover the mapping between gist feature vectors and class categories. 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 sample, the gist of scene is extracted, and passed to the learned map to generate its top-down prediction class category $\hat{C}$. In Sections 2.1 and 2.2, the gist feature extraction is explained in detail.

2.1. Multi-Scale Feature Extraction

The structure of the gist feature extraction is presented in Fig. 2. As stated earlier, the starting point of this model is our previously proposed bottom-up auditory attention model [6]. Hence, the bottom-up and top-down models share a common front-end: multi-scale feature extraction module and center-surround operation which finally yields feature maps. This also saves in some computational cost in case the bottom-up and top-down models are combined in the future. The multi-scale feature extraction will be explained briefly here, one may refer [6] for details.

The feature extraction is biologically inspired, and it mimics the processing stages in the early and central auditory systems. First, the auditory spectrum of the sound is estimated using an early auditory (EA) system model. The EA model used here consists of cochlear filtering, inner hair cell, and lateral inhibitory stages mimicking the process from basilar membrane to the cochlear nucleus in the auditory system [4]. The cochlear filtering is implemented using a bank of 128 overlapping constant-Q asymmetric band-pass filters. For analysis, audio frames of 20 milliseconds (ms) with 10 ms shift are used, i.e. each 10 ms audio frame is represented by a 128 dimensional vector.

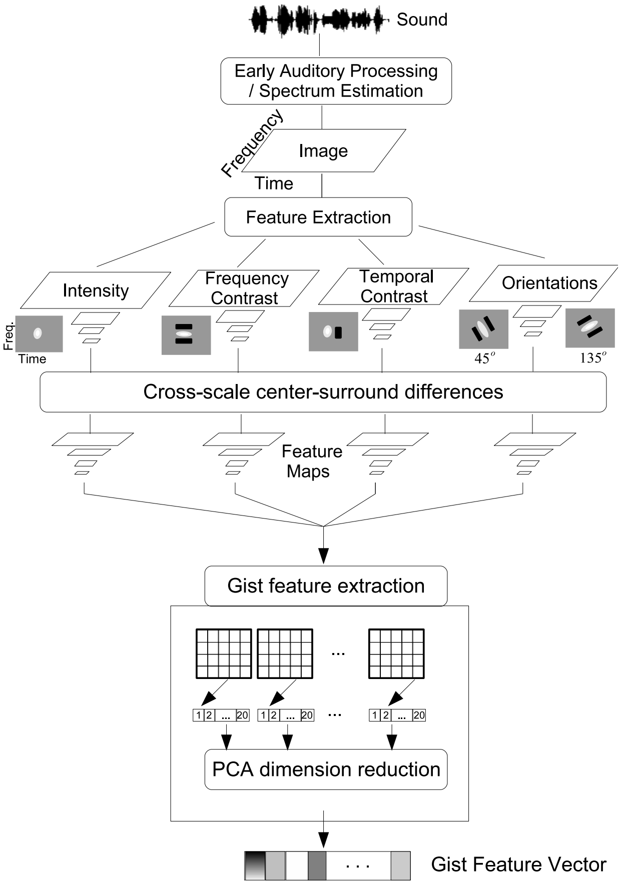

Fig. 2. Auditory gist feature extraction

The output of the EA model is an *auditory spectrum* with time and frequency axes.The spectrum is analyzed by extracting a set of multi-scale features which consists of *intensity (I), frequency contrast (F), temporal contrast (T)* and *orientation (O)* feature channels. They are extracted using 2D spectro-temporal receptive filters mimicking the analysis stages in the primary auditory cortex. Each of the receptive filters (RF) simulated for feature extraction are illustrated with gray scaled images in Fig. 2 next to its corresponding feature channel. The excitation phase, and inhibition phase are shown with white and black color, respectively. For example, the frequency contrast filter corresponds to RF with an excitatory phase and simultaneous symmetric inhibitory side bands. The RF for generating frequency contrast, temporal contrast and orientation features are implemented using 2D Gabor filters with angles (θ) 0^o, 90^o, $\{45^o, 135^o\}$, respectively. The RF for intensity feature is implemented using a 2D Gaussian kernel. The multi-scale features are obtained using a dyadic pyramid: the input spectrum is filtered, and decimated by a factor of two, and this is repeated. Finally, eight scales are created (if the scene duration W is large enough; otherwise there are fewer scales), yielding size reduction factors ranging from 1:1 (scale 1) to 1:128 (scale 8). For details of the feature extraction and filters used, one may refer to [6].

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

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

The across scale subtraction between two scales is computed by interpolation to the finer scale and pointwise subtraction. Here, $c = \{2,3,4\}$, $s = c+\delta$ with $\delta\epsilon\{3,4\}$ are used, which results in six feature maps for each feature channel, except orientation, when features are extracted at eight scales. The orientational channel has twelve feature maps since two angles of $\theta = \{45^o, 135^o\}$ are used.

2.2. Gist Features

Processing of the gist is rapid, and the gist of a scene describes the overall properties of the scene [7, 5]. The gist extraction algorithm is similar to the one proposed in [7] for vision. A gist vector is extracted from the feature maps of I, F, T, O_θ such that it covers the whole scene at low resolution. A feature map is divided into m by n grid of subregions, and the mean of each subregion is computed to capture rough information about the region, which results in a gist vector with length $m \times n$. For a feature map $\mathcal{M}_i$ with height h and width w, the computation of gist features can be written as:

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

An example of gist feature extraction with $m = 4, n = 5$ is shown in Fig. 2. After extracting a gist vector from each feature map, we obtain the cumulative gist vector by augmenting them. Then, principal component analysis (PCA) is used to reduce the dimension to make the machine learning more practical.

Averaging operation is the simplest neuron computation. Other second-order statistics such as variance may also provide additional information, but for our application we found that there was no appreciable benefit of using it.

3. EXPERIMENTS

To test our top-down task dependent model with gist features, the Boston University Radio News Corpus (BU-RNC) database [8] was used in the experiments. The BU-RNC is a broadcast news-style read speech corpus that consists of speech from 3 female and 3 male speakers, totaling about 3 hours of acoustic data. A significant portion of the data has been manually labelled with prosodic tags. The database also contains the orthography corresponding to each spoken utterance together with time alignment information at the phone and word level. To obtain the syllable level time-alignment information, the orthographic transcriptions are syllabified using the rules of English phonology [9]. We mapped all pitch accent types (H*, L*, L*+H, etc..) to a single stress label, reducing the task to a two-class problem. Hence, the syllables annotated with any type of pitch accent were labelled “prominent”, and otherwise “non-prominent”. The database consists of approximately 49,000 syllables, and the prominent syllable fraction is 34.3% (chance level). We chose this database for two main reasons: i) syllables are stress labelled based on human perception ii) since it carries labelled data, it helps us to train the top-down model in a supervised fashion.

The learner in Fig. 1 is implemented using a 3-layer neural network (MLP) 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. 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 prediction. The reason for using neural network classifier is that they are biologically well motivated. In addition to the neural network learner, a nonparametric k-nearest neighbor (k-NN) classifier (k = 5) was also used. All of the experimental results presented here are estimated using the average of 10-fold cross-validation. For each cross-validation, 90% of the data were retained for the training set and remaining 10% was used for testing. Also, during MLP training 10% of the training data was separated as validation set to be used as stopping criteria during training.

3.1. Defining the “Scene” and Its Duration

A scene is generated for each syllable in the database. The scenes are produced by extracting the sound around each syllable with an analysis window that centers on the syllable. To design the scene window duration W, we derived the statistics of the database to get an estimate. It is 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 BU-RNC database.

The role of scene duration W was investigated in the experiments. The duration is varied starting from 0.2 s, considering only the syllable itself, up to 1.4 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 $(m,n) = (1,w)$ grids, where w is the width of the feature map. The dimension of the gist feature vector, D, changes with varying scene duration. For instance, when $W = 0.6$ s, the EA model outputs a 128×60 dimensional image. Then, we can extract features up to 6 scales (instead of 8 scales), which enables the center-surround operation at scales $(c-s) = \{(2-5),(2-6),(3-6)\}$. When $(m,n) = (1,w)$, the dimension of the gist vector for each feature is $(30+30+15) = 75$ (since w is 30 at scale-2 and 15 at scale-3), finally resulting in a cumulative gist vector of $(75*5) = 375$ dimension (one feature set for each I, F, T and two sets for O_θ since $\theta = \{45^o, 135^o\}$, total 5 sets). Finally, the dimension is reduced to 68 with PCA while retaining 99% of the variance.

4. RESULTS

The performance of k-NN and MLP classifiers as a function of scene duration is shown in Fig. 3. We can conclude that the performance depends on scene duration when k-NN algorithm is used as learner. It performs poorly for both short ($W < 0.4$ s) and long scene durations ($W > 0.8$ s). The best performance achieved with k-NN classifier is when $W = 0.5$ s. In contrast to the k-NN algorithm, the MLP still learns the mapping between the scenes and the prominence classes even when the scene duration is large, at the expense of computation (when the scene duration is large, gist feature dimension gets larger requiring a larger neural network for training). It can be observed from Fig. 3 that the accuracy does not change significantly for varying scene durations with MLP classifier, except for $W = 0.2$ s. The best performance achieved with MLP is when $W = 0.8$ s. Both of the MLP and the k-NN classifiers perform poorly for short scenes ($W < 0.5$ s). This essentially indicates that the prominence of a syllable is affected by its neighboring syllables.

In Table 1, some of the results are detailed with accuracy (Acc.), precision (Pr), recall (Re) and F-score (F-sc) values, together with scene duration W and gist feature dimension D after PCA. The re-

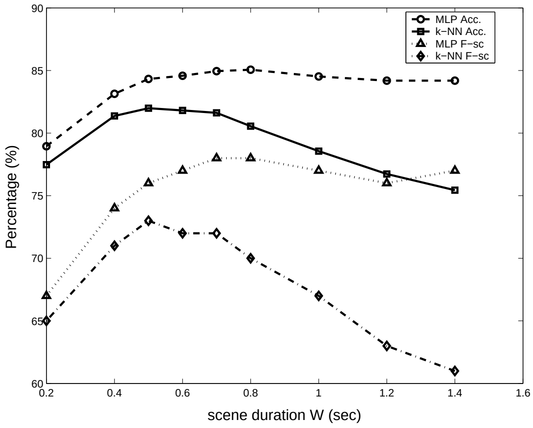

Fig. 3. Performance of the MLP and the k-NN learners for varying scene durations (Acc: Accuracy, F-sc: F-score)

sults obtained with unsupervised bottom-up (BU) attention model from [6] are also summarized in Table 1 for comparison purpose. In the prominence syllable detection task, the best performance achieved is 85.1% accuracy with an F-score= 0.78, and obtained with the MLP when $W = 0.8$ s.

The effect of the grid size on the 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 $(m, n) = (4, 5)$ grids, resulting in a $(4 * 5) = 20$ dimensional gist vector. As in the previous example, when $W = 0.6$ s, it generates a 300 dimensional cumulative gist feature vector. The dimension is reduced to 122 with PCA while still retaining 99% of the variance. For all scene durations, it results in a larger dimensional gist feature vector compared to $(m, n) = (1, w)$. This indicates that the gist features obtained with $(m, n) = (4, 5)$ carries more diverse information about the scene compared to the one obtained with $(m, n) = (1, w)$. In Table 2, results obtained with $(m, n) = (4, 5)$ for varying scene durations are reported using the MLP. The best performance achieved with $(m, n) = (4, 5)$ is 85.8% accuracy with an F-score= 0.79, and obtained when $W = 0.8$ s using the MLP learner. The top-down model provides approximately 10% absolute improvement over the bottom-up model. The results also compare well against the previously reported performance levels with the BU-RNC database, e.g. a supervised model obtained 76.6% accuracy using only acoustical features, and 83.9% accuracy using acoustical and syntactical features in [10].

5. CONCLUSION AND FUTURE WORK

In this paper, a task-dependent top-down auditory attention model is presented. A set of multi-scale auditory features are extracted in parallel from the auditory spectrum of the sound, and converted into low-level auditory gist features that capture the essence of a scene. Using a machine learning algorithm, the model learns the mapping between the gist features and the acoustical scene. The model was demonstrated to successfully detect prominent syllables in read speech with up to 85.8% accuracy. These results are encouraging given that the average inter-transcriber agreement for manual annotators is 80-85% for stress labelling [8].

Table 1. Prominent Syllable Detection Performance with $(m, n) = (1, w)$

Learner	W (s)	D	Acc.	Pr.	Re.	F-sc
BU	0.6	NA	75.9%	0.64	0.79	0.71
k-NN	0.5	56	82.0%	0.76	0.7	0.73
MLP	0.5	56	84.3%	0.80	0.73	0.76
MLP	**0.8**	**116**	**85.1%**	**0.80**	**0.75**	**0.78**

Table 2. Prominent Syllable Detection Performance with $(m, n) = (4, 5)$ using MLP

W (s)	D	Acc.	Pr.	Re.	F-sc.
0.4	110	84.9%	0.81	0.74	0.77
0.6	122	85.6%	0.81	0.76	0.79
0.8	**190**	**85.8%**	**0.81**	**0.77**	**0.79**

It has been experimentally seen that the prominence of syllables is affected by the neighboring syllables. Considering the performance and the computational cost, it may be reasonable to have an analysis window duration of 0.5-0.6 s for the prominent syllable detection task. A finer grid at the gist feature extraction stage increases the resolution and the computational cost, since it produces larger dimensional feature vectors. A suitable balance between resolution and the cost based on the chosen application needs to be found.

The top-down information that comes with language has not been considered in this work. As a part of our future work, we would like to incorporate top-down influences of lexical and syntactic knowledge into the proposed model. Also, the presented top-down model can be combined with the bottom-up auditory attention model such that the combined model makes a selection among the perceptually salient locations obtained from the bottom-up model.

The top-down auditory attention model proposed here is not limited to prosody labelling. It can be used in other spoken language processing tasks and general computational auditory scene analysis applications to classify ambient scenes, as well.

6. REFERENCES

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

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

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

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

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

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

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

[9] D. Kahn, *Syllable-based generalizations in English phonology.*, Massachusetts Institute of Technology, 1976.

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