# Saliency-Driven Unstructured Acoustic Scene Classification Using Latent Perceptual Indexing

Ozlem Kalinli [#1], Shiva Sundaram [*2], Shrikanth Narayanan [#3]
[#] *Speech Analysis and Interpretation Laboratory (SAIL), Dept. of Electrical Engineering-Systems*
*University of Southern California, Los Angeles, CA, USA.*
[1] kalinli@usc.edu [3] shri@sipi.usc.edu
[*] *Deutsche Telekom Laboratories, Quality and Usability Lab, TU-Berlin, Berlin, Germany.*
[2] shiva.sundaram@telekom.de

***Abstract*—Automatic acoustic scene classification of real life, complex and unstructured acoustic scenes is a challenging task as the number of acoustic sources present in the audio stream are unknown and overlapping in time. In this work, we present a novel approach to classification such unstructured acoustic scenes. Motivated by the bottom-up attention model of the human auditory system, salient events of an audio clip are extracted in an unsupervised manner and presented to the classification system. Similar to latent semantic indexing of text documents, the classification system uses unit-document frequency measure to index the clip in a continuous, latent space. This allows for developing a completely class-independent approach to audio classification. Our results on the BBC sound effects library indicates that using the saliency-driven attention selection approach presented in this paper, 17.5% relative improvement can be obtained in frame-based classification and 25% relative improvement can be obtained using the latent audio indexing approach.**

## I. Introduction

Automatic categorization of complex, unstructured [1] acoustic scenes is a difficult task as the appropriate label eventually associated with the audio clip of an unknown scene depends on the key acoustic event present in it. For example, an audio clip labeled *crash* may have human conversation and/or other related sources in it but the highlight "crash" of the vehicle is used to categorize the acoustic scene. When any (unknown) number of acoustic sources can be present in a clip but only one or a handful of them are relevant, the approach adopted by conventional audio classification systems would entail classifying every source in the scene and then implement an application specific post processing. This approach has two major drawbacks that can be immediately identified. (1) A large amount of computational resources are committed to processing feature-level information that is subsequently marginalized and (2) it is not possible to train for every possible acoustic source a priori. Motivated by this, in this paper, we present a novel approach that addresses the two issues by combining models of human attention-driven processing with class-independent representation of audio clips. We test our system on the BBC sound effects library [1] that consists of a large variety of audio clips belonging to acoustic scene categories such as *household, military, office* etcetera. This data set is particularly challenging for machine-based classification task because almost all the clips contains multiple, unknown number of unique acoustic sources present in it.

Axiomatically, it is known that in comparison to machine-based systems, humans can precisely process and interpret complex scenes rapidly despite the tremendous amount of stimuli impinging our senses. One of the key enablers of this capability is a neural mechanism, called "attention", that selects a subset of available sensory information before fully processing all stimuli at once [2], [3]. Attention can be thought as a spotlight that is directed towards a target of interest in a scene to enhance the processing in the related area while ignoring the stimuli that fall outside of the spotlighted area [4]. In a scene, some stimuli are inherently salient within the context, and they attract attention. For example, a red flower among yellow flowers or the sound of a siren in a street immediately and unconsciously attracts attention. Saliency-driven attention is a rapid, bottom-up, task-independent process, and it detects the objects that perceptually pop-out of a scene by significantly differing from their neighbors [2]. The second form of attention is a top-down task-dependent process which uses prior knowledge and learned past expertise to focus attention on the target locations in a scene to enhance information processing [2], [3]. For example in vision, it was shown that gaze patterns depend on the task performed while viewing the same scene [5]. The gaze of the observer fell on faces when estimating the peoples age, but fell on clothing when estimating the peoples material conditions. Similarly in audition, it is the selective attention that allows a listener to extract a particular persons speech in the presence of others (the cocktail party phenomenon) by focusing on a variety of acoustic cues such as pitch, timbre, spatial location, etc. [3].

As stated previously, one caveat of the conventional audio content processing approaches is that they process the entire signal or acoustical scene fully and equally in details (i.e. recognizing each and every source/event in an acoustic scene). This issue can be alleviated by taking advantage of a selective attention mechanism similar to what humans perform. Thus, here we propose a novel method that emulates human auditory attention for acoustic scene recognition. The algorithm first detects the salient audio events in a cluttered auditory scene in an unsupervised manner, and then processes only the selected events with a previously learned representation

[1] where an acoustic scene may consist of any number of unknown sources which may also overlap in time.

*MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.*
*978-1-4244-4464-9/09/$25.00* 

for acoustic scene recognition. It is important to note that saliency (and its definition) does not depend on any individual acoustic source; hence we are interested in class-independent representation approach for a classification framework. For class-independent representation of audio clips, we use latent perceptual indexing (LPI) [6], [7], which seeks a single vector representation of an audio clip within a collection by using unit-document frequency measures. The main advantage of this approach is that it allows for comparison of arbitrary audio clips through vector similarity measure that also embodies both semantic and perceptual similarities [7]. By combining this with saliency-based attention model, called *latent indexing using saliency* (LISA), the work presented here allows us to process only a subset of meaningful information in a complex acoustic scene. This has the potential to improve classification accuracy of unstructured audio clips and additionally, reduce the computational bandwidth[2] required to process audio content.

The paper is organized as follows: first a comprehensive discussion of related work in audio content processing is present in Section II, then auditory saliency map model is explained in Section III-A followed by latent perceptual indexing in Section III-C. The experimental results and conclusions are presented in Section IV and V, respectively.

## II. Related Work

Starting from [8], typical examples of audio classification systems use category based modeling for a selection of audio clips [8], [9], [10]. In [8], [9] the system is evaluated on a variety of categories such as *animals, bells, crowds, female, laughter, machines, male voices, percussion instruments, telephone, water sounds etcetera*. While the performance of these systems is notable, they were trained and tested on homogenous[3] clips. Examples of similar approaches that deal with more complex acoustic scenes include sports highlighting [11], context-aware listening for robots [12] and also in background/foreground audio tracking [13]. While these methods target more complex acoustic scenes they are still based on the typical classification approaches of category-based modeling and therefore they are difficult to generalize to clips of unstructured acoustic scenes.

Examples of other approaches that deal with clips of unstructured acoustic scenes are [14], [15]. In [14] the author improves on the naive labeling scheme by creating a mapping from each node of a hierarchical model in the abstract semantic space to the acoustic feature space. The nodes in the hierarchical model (represented probabilistically as words) are mapped onto their corresponding acoustic models. In [15], the authors have adopted a similar approach of modeling features with text labels in the captions. In such cases, however, the focus has mainly been on relating the audio clips to its language-level descriptions.

In contrast to these approaches and the typical class-based training approaches presented earlier, in this work we focus on selectively processing the audio events similar to the way humans detect important segments in a cluttered acoustic scene and subsequently use them to classify the given clip. In the literature, computational attention models have been mostly explored for vision. For example in [2], a concept of saliency map was proposed to understand bottom-up visual attention in primates, and it was shown that the model could replicate several properties of human attention, i.e. detecting traffic signs, detecting colors etc. Inspired by the visual saliency map, a bottom-up auditory attention model was proposed by us in [16], and it was shown that the model could detect prominent syllables in speech. Here, we use the bottom-up attention model to detect salient audio events present in an acoustic scene. As far as we know, there has been no work in this area that applies saliency-based attention models to recognition of unstructured acoustic scenes.

[2]In terms of amount of training data and/or runtime memory requirements.

[3]Audio clips or segments that contain only one acoustic source in it, for example an instance of *laughter*.

## III. Proposed Method

The block diagram of the proposed method is shown in Fig. 1. First, audio signal is fed into a salient event detector which is described in Section III-A. The output of the salient event detector is a one dimensional saliency score time-aligned with the original acoustic signal. As explained in section IV, the audio events for subsequent classification are selected in a decreasing order of saliency. To capture the audio event corresponding to a salient point, the sound around each salient point is extracted using a window of duration $W$ that centers on that time point. In other words, we assume that an audio segment of duration $W$ that centers on a salient point (in time) corresponds to a salient audio event. The perceptually motivated features are extracted from the detected salient audio events and indexed into the latent space ('Learner' in training) as explained in Section III-C. Classification of an unknown test clip (the 'Predictor') is performed by comparing it with a collection of labeled audio clips ('Learner' in training) in the latent space. It is important to note that the collection of labeled training clips are used only to assess the performance of the approach presented here, the actual information used to derive the latent representation are derived in a class-independent manner. Details of the experiments are given in Section IV. Next, the salient audio event detector is explained.

### A. Audio Saliency Map

At the core of the proposed method is our previously described bottom-up auditory attention model which computes an auditory saliency map from the input sound [16]. The block diagram of the auditory attention model is given in Fig 2. First, an auditory spectrum of sound is estimated using an early auditory (EA) system model. The EA model 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 [17]. 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.

Next, the auditory spectrum is analyzed by extracting a set of multi-scale features which consist of *intensity ($I$), frequency contrast ($F$), temporal contrast ($T$)* and *orientation ($O$)* feature channels. They are extracted using 2D spectro-temporal receptive filters mimicking the analysis stages in the primary auditory cortex. Each of the receptive filters (RF) simulated for feature extraction is illustrated with gray scaled

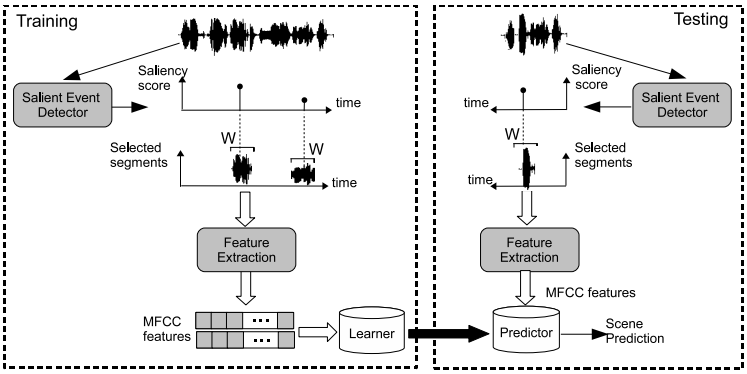


Fig. 1. Salient Audio Event Detection. W is the duration of the window that centers on the detected salient time point to extract salient audio event.

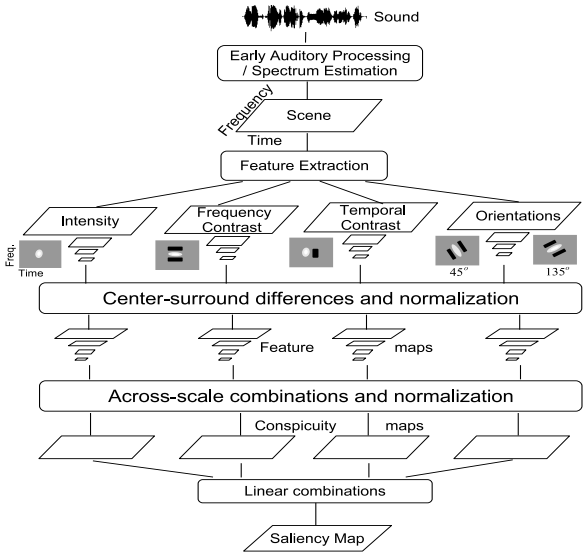


Fig. 2. Auditory saliency map model.

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 receptive fields in the auditory cortex 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 ($\theta$) $0^o$, $90^o$, $\{45^o, 135^o\}$, 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 audio segment duration $W \geq 1.28$ s; otherwise there are fewer scales), yielding size reduction factors ranging from 1:1 (scale 1) to 1:128 (scale 8).

As shown in Fig 2, 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 detects the local temporal and spatial discontinuities in feature channels. Center-surround differences are computed as point wise differences across scales using three center scales $c = \{2, 3, 4\}$ and two surround scales $s = c + \delta$ with $\delta\epsilon\{3, 4\}$ resulting in six feature maps for each of the feature channels. In total, there are 30 features maps computed: six for each intensity, frequency contrast, temporal contrast and twelve for orientation since it has two angles $\theta = \{45^o, 135^o\}$. Each feature map is normalized in the order of within-scale, across-scale, and across-features. The normalization algorithm is an iterative, nonlinear operation simulating competition between the neighboring salient locations using a large 2D difference of Gaussians filter [16]. As a result of normalization, possible noisy feature maps are reduced to sparse representations of only those locations which strongly stand-out from their surroundings [2], [16]. All normalized maps are then summed to provide bottom-up input to the saliency map.

The saliency map holds non-negative values and its maximum defines the most salient location in 2D auditory spectrum. It is assumed that saliency combines additively across frequency channels. The saliency map is summed across frequency channels for each time point, and normalized to $[0, 1]$ range for each audio clip, yielding a saliency score $S(t)$ for each time point $t$. Then, the local maxima of $S(t)$ are found and the audio event at the corresponding time point is marked as salient together with its saliency score. Later, these salient points are selected in the order of decreasing saliency score as discussed in Section IV.

### B. Discussion

The bottom-up attention model is capable of detecting only the salient audio events represented in at least one of the four implemented features; i.e., intensity, frequency contrast, temporal contrast and orientation features. In Fig. 3, a sample sound clip tagged with "goat_machine_milked" is shown. In the figure, the first and second tiers show the waveform and the spectrum of the clip, respectively. The third tier shows the transcription where M represents machine noise and G represents goat voice. The fourth tier shows the saliency score results obtained from the bottom-up auditory attention model. For this clip, the model detects location of all goat voices in the sound clip. Although the third goat event from the left is drowned in the machine sound in the background, the model could successfully detect this event as well. This alludes to the fact that the model is not limited to intensity feature. This scene can be summarized as follows; in the scene the voice of goat pops out perceptually while the machine noise becomes less prominent as in the figure-ground phenomenon in visual

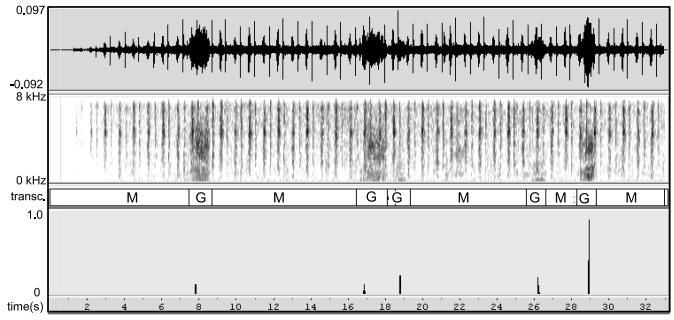


Fig. 3. Results of a sample sound clip tagged as "goat_machine_milked". The tiers shows i) waveform of sound, ii) spectrum, iii) transcription where M represents machine noise and G represents goat voice iv) saliency score.

perception.

Similarly, another example from the database we used is a sound clip tagged with "wigeon_at_pool". For this clip, the auditory attention model detects the locations of the bird tweets and suppresses the background water sound.

### C. Latent Perceptual Indexing

In this work, latent perceptual indexing (LPI) [6] is used for class-independent representation of audio clips. An entire audio clip from a collection of audio clips is represented as a single vector in a latent perceptual space; this is similar to latent semantic indexing/mapping (LSI) [18], [19] for text documents. First, a *bag of feature-vectors* is extracted from a given audio clip. Then, this clip is characterized by calculating the number of feature-vectors that are quantized into each of the *reference clusters* of signal features (analogous to the term-document frequency counts in information retrieval). By applying this procedure to the whole collection of clips, it results in a sparse matrix where each row represents a quantitative characterization of a complete clip in terms of the reference clusters. The reference clusters are obtained by unsupervised clustering of the whole collection of features extracted from the clips in the library, and assumed to represent distinct perceptual qualities. A reduced rank approximation of this sparse representation is obtained by singular-value decomposition resulting in mapping audio clips to points in a latent perceptual space. Thus each audio clip is represented as a single vector. The LPI approach is similar to LSI of text documents [19]; the units or reference clusters in LPI are taken to be equivalent to terms (or words) in LSI and the audio clips in LPI are equivalent to text documents in LSI.

This method is implemented as follows. Let us assume that a collection of $M$ audio clips is available in a database with the $i^{th}$ clip having $T_i$ feature-vectors. Then, the procedure involved in obtaining a representation in the latent perceptual space listed below:

**STEP 1.** The collection of all the feature-vectors obtained from all the clips in the database is clustered using the *k-means* clustering algorithm. This results in $C$ *reference clusters*.

**STEP 2.** Let the $i^{th}$ audio clip have a total of $T_i$ frames.
FOR audio clip $A_i$ where, $i \in \{1, \ldots, M\}$, DO:

i. Calculate : $f_{i,j} = \frac{\sum_{t=1}^{t=T_i} I(lab(t)=j)}{T_i} . \forall j \in 1, \ldots, C$. Here $I(\cdot) \in \{0, 1\}$ is an indicator function.
$I\,(lab(t) = j) = 1$ if the $t^{th}$ frame is labeled to be in the $j^{th}$ cluster, otherwise $I(\cdot) = 0$.

ii. Assign $F(i, j) = f_{i,j}$ the $(i, j)^{th}$ element of the sparse matrix $F_{M \times C}$.

**STEP 3.** END FOR loop;

**STEP 4.** Obtain $F_{M \times C} = U_{M \times M} \cdot S_{M \times C} \cdot (V_{C \times C})^T$ by SVD.

**STEP 5.** Obtain the approximation of $F$ as $\tilde{F}_{M \times C} = \tilde{U}_{M \times R} \cdot \tilde{S}_{R \times R} \cdot (\tilde{V}_{C \times R})^T$ by retaining the $R$ largest singular values.

In addition to the $F$ matrix obtained at the end of step 3, an entropy-based weighting term also weighs each column [19]. The approximation $\tilde{F}$ is obtained by the span of basis vectors that have significant singular values. By retaining only the significant singular values, the randomness in quantization is eliminated. The similarity measure between a given test audio clip and the audio clips in the training set in the latent space is computed using cosine vector similarity function [6], [19]. Using this measure, the k-nearest neighbor (KNN) is used for classification of an unknown test audio clip.

In LPI, all segments or feature-vectors of an audio clip are used for indexing. Here, we propose a modified method called *latent indexing using saliency* (LISA) which combines saliency based audio event selection with class independent LPI method for audio scene recognition. In other words, LISA uses only selected salient segments of an audio clip whereas the original LPI uses the whole audio clip for scene recognition. The information processed by SVD of the term-document matrix in LPI is different from the segments selected by the saliency map. In LPI, one attempts to derive the underlying perceptual structure by eliminating randomness caused by different recording conditions or realizations of the same acoustic source. However, auditory saliency model selects salient events in an audio clip while ignoring parts that would typically constitute 'background' in an acoustic scene. As our result illustrates, by combining the two in LISA, we attempt to use only a subset of meaningful acoustic information in an unsupervised manner to classify a given acoustic scene. It is important to note that in the next section, for LPI and LISA, the category labels are only used to asses and compare the performance of the different approaches. The latent representation derived is obtained using only unsupervised $k-$means algorithm for reference clusters and SVD of the term-document matrix.

## IV. Experiments and Results

For the experiments in this paper 2,491 whole audio clips from the BBC Sound Effects Library [1] were used. The sound clips consist of natural unconstrained audio recorded in real environments that is composed of many mixed audio events and sources. The duration of clips varies from 1 second to 9.5 minutes. The database is available pre-organized according to high-level semantic categories and their corresponding subcategories. Each clip in the library is labeled with a semantically high-level category that best describes the acoustic properties of the scene. There are twenty one categories with varying number of sound clips under each category as in Table I.

The twelve dimensional Mel-frequency cepstral coefficients (MFCCs) ($C0$ energy feature was excluded) were extracted from each audio clip for sound classification experiments. The MFCCs are based on the early auditory system of humans and successfully used in generic audio classification task in the literature [20]. Instead of the features extracted from the front-end of auditory saliency model, we preferred to use the standard MFCC features here since the focus of this work is

TABLE I
DISTRIBUTION OF CLIPS UNDER EACH CATEGORY

| Category | No. of files | Category | No. of files |
|---|---|---|---|
| IMPACT | 16 | NATURE | 85 |
| OPEN | 8 | SPORTS | 151 |
| TRANSPORTATION | 295 | HUMAN | 357 |
| AMBIENCES | 311 | EXPLOSIONS | 18 |
| MILITARY | 102 | MACHINERY | 117 |
| ANIMALS | 359 | SCI-FI | 121 |
| OFFICE | 144 | POLICE | 96 |
| HORROR | 98 | PUBLIC | 44 |
| AUTOMOBILES | 53 | DOORS | 4 |
| MUSIC | 25 | HOUSEHOLD | 38 |
| ELECTRONICS | 49 | | |

acoustic scene classification based on salient acoustic events rather than definition and presentation of new features. The MFCC features were extracted every 10 ms with a Hamming window of 20 ms length. The length of audio segment for audio classification task was analyzed empirically in [20], and the best audio classification accuracy was obtained using 1 second (sec) window. Thus, mean and standard deviation of the MFCC features were estimated over 1 sec window resulting in 24 dimensional feature vector representing each 1 sec audio segment.

All classification performances are evaluated by ten-fold cross-validation. In this, 10% of the whole database is chosen as the test set and the remaining were retained as the train set. This is repeated ten times (without replacement) and the final result is the average of these repetitions. Chance-level performance, which is dependent on data distribution amongst the categories, was estimated to be 14.4%.

First, we establish a baseline system based on the conventional approach of creating category-based models. A 3-layer neural network is used for baseline classification experiments. The neural network had $D_{in}$ inputs, $(D_{in} + D_{out})/2$ hidden nodes and $D_{out}$ output nodes, where $D_{in} = 24$ is the length of feature vector, and $D_{out} = 21$ since there are twenty one classes. The neural network is used together with 1 sec audio segments. Later, the output of 1 sec frame classification results were combined by majority voting to obtain the sound clip classification result since some clips are longer than 1 sec. The baseline result is obtained by using all the frames in all the clips (i.e., without using the auditory saliency model) and it was 40.0% accuracy.

For the first experiment, the saliency model is used to scan and detect salient audio events in a scene as explained in Section III-A and only these salient segments are used for classification. The saliency score takes values between 0.0 and 1.0. A value close to 0.0 indicates no saliency and 1.0 indicates the most salient point in an audio clip. The score for each clip is sorted, and the top $N$ locations are marked as salient. Then, $W = 1$ sec window centered on a marked time location is used to extract the corresponding salient audio event. 24 dimensional MFCC features are extracted from these segments as explained previously. The reduction in data gained by keeping only top $N$ salient events is illustrated in Fig. 4. The number of retained salient audio events are varied starting from $N = 1$ to $N = all\_sal$ (all the detected salient points are used irrespective of their saliency score). Retaining only the top salient point provides 98.8% data reduction. Retaining all the salient audio events still provides more than 40% data reduction.

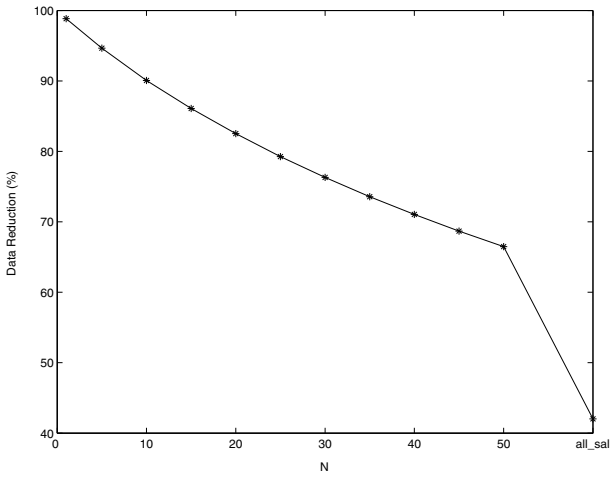


Fig. 4. Amount of data reduction as a function of the number of retained salient points ($N$).

For classification, the 'learner' in Fig. 1 is implemented using a 3-layer neural network to test the effectiveness of salient audio event detection for frame-based classification. The frame-based classification results after applying the saliency model are shown as a function of $N$ in Fig. 5. Additionally, Fig. 5 presents the baseline result using all the frames (40% accuracy) and the chance level (14.4%) for comparison purpose. It can be observed that the performance obtained by retaining only the top salient location ($N = 1$) is better than using all of the frames (the whole sound clip). The best result is 46.9% clip accuracy obtained with $N = 15$. This provides approximately 7% absolute improvement over the baseline while reducing the amount of data processed by the classifier by more than 85% as shown in Fig. 4. A reduction in performance is observed when *all* the salient locations are used for classification. This however, is still above the baseline result.

Finally, the results obtained using LISA are illustrated in Fig 6. In the experiments the number of reference clusters in LISA are varied starting from $C = 200$ to $C = 2000$ with a step size of 100. For KNN, $K = 7$ nearest neighbors were found to have the best performance. In Fig. 6, we present the best accuracy results obtained with $C$ reference clusters by retaining top $N$ salient points. The best performance of 49.7% was obtained with LISA by retaining the top 35 salient points and using $C = 1700$ clusters. LISA provides approximately 10% absolute improvement over the baseline frame-based classification and 3% absolute improvement over frame-based classification using only the salient segments. Results obtained using only LPI (i.e. using the whole audio clip without salient event selection) is also shown in Fig. 6. LPI achieves 50.4% classification accuracy using $C = 2000$ clusters. As a result, it can be seen that comparable results to LPI using all the feature vectors can be obtained by using only top 35 salient points (data reduction of approximately 74%). Consequently, we can also say that in most cases, the salient segments of an audio clip are the defining moments of the audio clip of an unstructured acoustic scene.

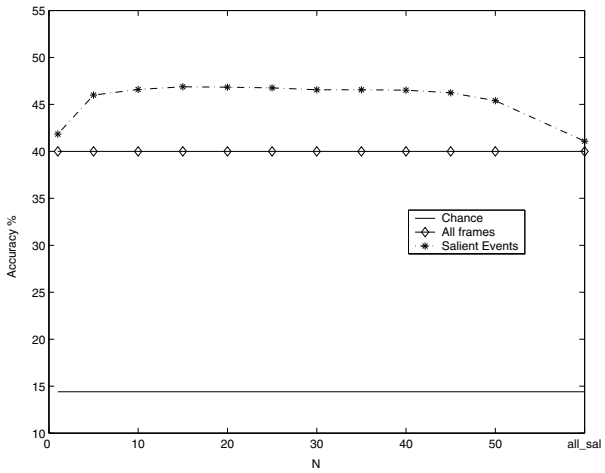


Fig. 5. Clip accuracy results obtained using all frames and top N salient audio events using neural network classifier.

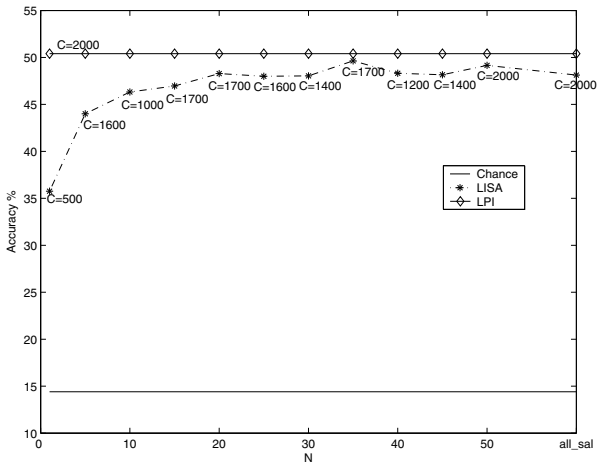


Fig. 6. Clip accuracy results obtained with LISA and LPI methods.

## V. Conclusion and Discussion

In this paper, a novel method called LISA that mimics human auditory attention for acoustic scene recognition was presented. LISA first detects the salient audio events present in an unstructured audio clip using a bottom-up auditory attention model, and then processes only the selected salient events for acoustic scene recognition using latent perceptual indexing. The salient event detection algorithm is completely unsupervised; it can be used to obtain salient, defining audio events in an acoustic scene cluttered with different (unknown) acoustic sources. This allows us to categorize unstructured audio clips of acoustic scenes without processing the whole clip. Additionally for such scenes, using the term-document frequency measures to derive a representation is desirable as it makes no assumptions about the individual sources present in it. This makes this approach applicable to a variety of audio content processing problems. The performance of the method is tested using the BBC sounds effect library, and it is shown that LISA provides 10% absolute (25% relative) improvement over the baseline by retaining only top 35 salient points, and reduces the amount of data approximately 74%. It is shown that LPI and LISA perform approximately the same however LISA uses less number of data points and reference clusters since it only uses selected salient events in an audio clip.

The auditory saliency model behaves as a highlighting mechanism that selects only the events that pop-out of an acoustic scene while ignoring segments or sources that are part of the background. For example, in the previously discussed example in Section III-B, the saliency model detects the locations of goat sound and ignores the machine sound in the clip tagged with "goat machine milked". Hence, the predicted label for this clip would be "Animal" when only salient segments are considered for classification. However, the high level semantic label for this clip provided with the database was"Machine". As it can be seen from the detailed tag, this is not completely incorrect since the clip description includes an animal sound. Relating semantic descriptions (with multiple tags) to ranked salient segments of acoustic scenes is an interesting avenue to explore with many applications in audio content processing. This is a part of our planned future work for this framework.

## References


[1] "The BBC Sound Effects Library Original Series," *http://www.sound-ideas.com*, May 2006.

[2] 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, 1998.

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

[4] E. Weichselgartner and G. Sperling, "Dynamics of automatic and controlled visual attention," *Science*, vol. 238, pp. 778–780, 1987.

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

[6] S. Sundaram and S. Narayanan, "Audio Retrieval by Latent Perceptual Indexing," *Proc. of ICASSP, Las Vegas*, 2008.

[7] ——, "Classification of sound clips by two schemes: Using Onomatopoeia and Semantic labels," *Proc. of ICME, Hannover, Germany*, June 2008.

[8] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-Based Classification, Search, and Retrieval of Audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, Fall 1996.

[9] G.Guo and S. Z. Li, "Content-Based Audio Classification and Retrieval by Support Vector Machines," *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 209–215, January 2003.

[10] L. Liu, H. J. Zhang, and H. Jiang, "Content Analysis for Audio Classification and Segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, October 2002.

[11] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Audio Events Detection Based Highlights Extraction from Baseball, Golf and Soccer Games in a Unified Framework," *Proc. of ICASSP, Hong Kong.*, April 2003.

[12] S. Chu, S. Narayanan, C. C. Kuo, and M. J. Mataric, "Where am I? Scene Recognition for Mobile Robots using Audio Features," *Proc. of ICME*, July 2006.

[13] R. Radhakrishnan and A. Divakaran, "Generative Process Tracking for Audio Analysis," *Proc. of ICASSP*, May 2006.

[14] M. Slaney, "Semantic Audio Retrieval," *Proc. of ICASSP, Orlando*, May 2002.

[15] L. Barrington, A. Chan, and D. T. land G. Lanckriet, " Audio Information Retrieval using Semantic Similarity," *Proc. of ICASSP, Honolulu, Hawaii*, 2007.

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

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

[18] S. Deerwester, S. T. Dumais, G. W. Furnas, T. Landaurer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, vol. 6, no. 41, pp. 391–407, 1990.

[19] J. Bellagarda, "Latent Semantic Mapping: A Data driven Framework for Modeling Global Relationships Implicit in Large Volumes of Data," *IEEE Signal Processing Magazine*, vol. 22, pp. 70–80, September 2005.

[20] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Proc. of ICASSP*, 2002.