CLASSIFICATION OF SOUND CLIPS BY TWO SCHEMES: USING ONOMATOPOEIA AND SEMANTIC LABELS

Shiva Sundaram and Shrikanth Narayanan

Speech Analysis and Interpretation Lab. (SAIL),
Dept. of Electrical Engineering-Systems, University of Southern California (USC)
3740 McClintock Ave, EEB.400, Los Angeles,CA 90089. USA.
shiva.sundaram@usc.edu,shri@sipi.usc.edu

ABSTRACT

Using the recently proposed framework for latent perceptual indexing of audio clips, we present classification of whole clips categorized by two schemes: high-level semantic labels and the mid-level perceptually motivated onomatopoeia labels. First, feature-vectors extracted from the clips in the database are grouped into *reference clusters* using an unsupervised clustering technique. A unit-document co-occurrence matrix is then obtained by quantizing the feature-vectors extracted from the audio clips into the reference clusters. The audio clips are then mapped to a latent perceptual space by the reduced rank approximation of this matrix. The classification experiments are performed in this representation space using corresponding semantic and onomatopoeic labels of the clips. Using the proposed method, classification accuracy of about sixty percent was obtained when tested on the BBC sound effects library using over twenty categories. Having the two labeling schemes together in a single framework makes the classification system more flexible as each scheme addresses the limitation of the other. These aspects are the main motivation of the work presented here.

Index Terms— audio classification, audio representation, indexing, semantic audio, onomatopoeia, latent document analysis.

1. INTRODUCTION

A variety of problems in multimedia processing are based on segmentation, classification and clustering of audio. These include applications such as robust automatic speech recognition (ASR), video stream segmentation, context recognition, browsing and audio/video retrieval. To accomplish this in a human readable way, especially for end-user applications, audio data is usually organized and indexed using words as tags or labels. Consequently, the objective of audio classification systems, in general, is to distinguish amongst these categories of audio.

Contemporary work in audio classification differ in its categorization and classification scheme according to the proposed application. The work presented in [1, 2] focuses on real-time speech/music discrimination of audio stream for robust ASR. A more sophisticated extension to this is to organize clips in a hierarchy. For example in [3] the authors present hierarchical categories such as silence, with music components and without music components. They also present results on grouping as pure speech, pure music, speech plus music background, sound effects plus music background, harmonic and non-harmonic sound effects. A similar hierarchical scheme by grouping sources into speech, music or environmental sounds is used in [4]. They first determined if a given segment is speech or non speech, followed by a speaker change point detection in the case of speech segments and for non-speech segments, the samples are classified as music, environment and silence. These systems are designed to be scalable, and used for applications such as online audio/visual segmentation and classification.

In the content-based approach for audio information retrieval presented in [5, 6], the authors use high-level specific audio categories such as *animals, bells, crowds, female, laughter, machines, telephone, water* sounds etc. Other examples of methods that deal with semantic labeling of audio are [7, 8, 9]. In [7], the author improves on the 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 [8], the authors have a similar approach of modeling features with semantic text labels in the captions. In [9], their systems use semantic relations in language. Here the authors have used WordNet to generate words for a given audio clip using acoustic feature similarities, and then retrieve clips that are similar to the tags.

In [10], the authors use onomatopoeic descriptions for retrieval of drum loops. Here, the query to the retrieval system is spoken form of onomatopoeia. The authors focus on eight basic drum categories and their corresponding onomatopoeia labels (such as *bom, ta, ti, do*). Spoken queries were in the form of short sequence of onomatopoeia words. Since they use onomatopoeia descriptions of acoustic properties, their approach is most relevant to the work presented here. Particularly, we focus on both semantic and onomatopoeic labeling of generic sound clips, and its classification. This has the descriptive advantage of content-based and the flexibility of hierarchical methods.

As mentioned, in this work we present classification of whole audio clips using two methods of categorization: high-level lexical-semantic labels and the mid-level perceptually motivated onomatopoeia word labels. Semantic labels are useful for audio retrieval since high-level description of acoustic sources typically denote a semantic event. However they can be perceptually vague, especially in the context of audio. For example, the label "*Ambiences*" could mean sounds in a cafeteria with sounds of cutlery and conversation, or even sounds in a forest containing tweeting birds and flowing water. The main motivation for using onomatopoeia labels, on the other hand, is the close and direct relation between these word-level entities and the perceptual characteristics of the target acoustic events. Many acoustic qualifications of events are intuitively expressed using onomatopoeia words [11]: for the event "*knocking on the door*", the words "*tap-tap-tap*" describe the acoustic properties well. Communicating acoustic events in such a manner is possible because of a two-way mapping between the acoustic space and language or semantic space. Existence of such a mapping is language dependent and a result of common understanding of familiar acoustic properties [11, 12]. This form of labeling is perceptually meaningful but can be semantically vague. For instance, *Buzz* can describe sound of either a machine, or bees. In summary, each scheme has its advantages and limitations, but having them together in a single framework makes the system more flexible as each labeling scheme addresses the limitation of the other. These aspects are the main motivation of the work presented here.

This paper is organized as follows. In the next section a technique based on unit-document co-occurrence measure to represent audio clips in a latent perceptual space is presented. Then, the data-set and the signal features used in this work is described. The experiments and the results obtained are also described, and finally a discussion of the results is presented.

The work presented in this paper was funded by the US Army, Office of Naval Research (ONR) and the Department of Homeland Security (DHS)

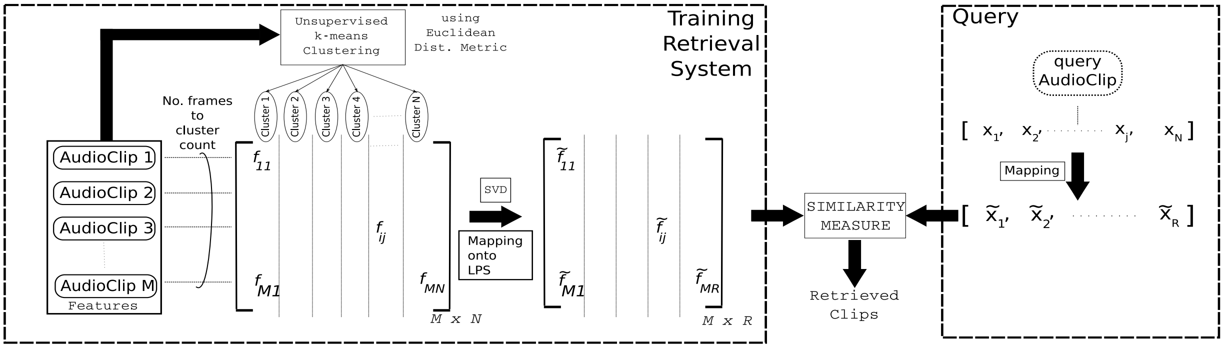

Fig. 1. Indexing and classification of sound clips in latent perceptual space.

2. PROPOSED METHOD

In this work, an entire audio clip from a sound library [13] is represented as a single vector in a latent perceptual space; this is similar to latent semantic mapping (LSM) [14] used for text document indexing. First, a *bag of feature-vectors* is extracted from a given audio clip. The feature-vectors can be signal-level measures such as the Mel-frequency cepstral coefficients (MFCCs) extracted using frame-based analysis. 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). This 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. 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 (LPS). Thus each audio clip is represented as a single vector. This representation of audio clips along with its semantic and onomatopoeic labels is used for the classification experiments. The algorithm is illustrated in figure 1 and its details are as follows:

Lets 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 N *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 : $f_{i,j} = \frac{\sum_{t=1}^{t=T_i} I(lab(t)=j)}{T_i}.\forall j \in 1, \ldots, N$. Here $I(\cdot) \in \{0,1\}$ is an indicator function.
$I\left(lab(t) = j\right) = 1$ 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 $F_{M \times N}$.

STEP 3. END FOR loop;

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

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

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. Since the initial matrix representation F was obtained from clusters of signal feature-vectors, the columns of $\tilde{U}$ and $\tilde{V}$ essentially span the LPS. Therefore, the given set of audio clips are indexed in the LPS. While the method presented here is similar to the LSM framework for text document indexing using term-document frequency, there are some differences which are discussed here for clarity. As stated in [14], LSM tries to uncover the underlying semantic structure in data by eliminating the *randomness* that arises due to variations in expressing the same concept with different choice of words. It maps discrete objects such as words and documents onto a continuous space. The words and documents occupy specific volumes in the semantic space as concepts, which is used in measuring "closeness" between documents. The present work, attempts to derive the underlying perceptual structure notwithstanding the randomness caused by temporal variations. Based on the feature-vectors extracted from a given database, the method presented here seeks distinct acoustic clusters in the perceptual space. These acoustic clusters in the perceptual space are analogous to concepts in the semantic space. Therefore, ideas of similarity measure between audio clips and representation of a test clip can be re-applied here. This, and a description of the database and the feature set used in this work is presented next.

3. EXPERIMENTS

3.1. Database

For the experiments in this paper, $M = 2,140$ whole audio clips from the BBC Sound Effects Library [13] were used. Each clip in the library is labeled with a semantically high-level category and a perceptually descriptive onomatopoeia tag that best describes the acoustic properties of the source. The database is available pre-organized according to high-level semantic categories and their corresponding subcategories. Onomatopoeia labeling of the audio clips was done manually. Choosing an onomatopoeic label that best describes the acoustic properties of the source in an audio clip is completely based on subjective perception, and this can only be achieved by manually listening to each clip. Such subjective categorization of clips for a reasonably large database is tedious and prone to inconsistencies. In this work, to keep errors to a minimum, the assignment of onomatopoeic labels was done in three passes and a final consolidation step. First a small set of clips were manually tagged by subjects. Then, other clips having the same file-name were assigned the same tags as the corresponding ones in the initial set. Details of these two steps are given in [12]. Finally by manual comparison with the sets obtained from the first two passes, the remaining files in the database were also appropriately labeled. As a final post labeling step using automatic clustering methods and word-based similarity measure, files with similar onomatopoeic labels were grouped together. This was done according to the automatic onomatopoeia word clusters obtained from the procedure described in [12]. This resulted in forming twenty two onomatopoeic categories for the set of M clips. At the end of the labeling procedure a semantic and onomatopoeia label is available for each clip in the database. The final distribution of clips for both semantic and onomatopoeia categories is shown in table 1.

3.2. Features

The fourteen dimensional feature-vectors extracted from each audio clip are comprised of twelve Mel-frequency cepstral coefficients (MFCCs), spectral centroid measure and the spectral roll-off frequency. The MFCCs are based on the early auditory system of hu-

	BANG	BEEP	BLEAT	BURR	BUZZ	CLATTER	CRACKLE	CROW	CRUNCH	DONG	GABBLE	GROWL	HONK	HUM	MEOW	SPLASH	SQUEAK	TAP	THUD	TICK	TWEET	WHOOSH
AMBIENCES	0	1	0	17	17	7	0	0	1	7	56	0	3	15	0	1	8	0	0	2	50	2
ANIMALS	1	0	14	0	13	26	1	6	10	3	0	62	2	2	60	3	54	33	3	1	53	2
AUTOMOBILES	0	1	0	24	3	0	0	0	0	1	0	0	5	4	0	0	9	0	2	2	0	0
DOORS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
ELECTRONICS	1	17	0	0	3	8	0	0	0	5	0	0	0	6	0	0	1	3	1	3	0	0
EXPLOSIONS	4	1	0	0	4	0	3	0	1	0	0	0	0	1	0	1	1	0	0	0	0	1
HORROR	4	2	0	0	10	0	2	1	7	6	2	1	0	6	0	2	12	0	11	1	1	11
HOUSEHOLD	0	0	0	0	5	1	1	0	0	9	0	0	0	7	0	8	1	0	0	4	0	1
HUMAN	0	3	0	4	3	4	8	1	4	6	33	1	6	1	0	3	10	177	17	0	1	3
IMPACT	0	0	0	0	1	1	0	0	7	4	0	0	1	0	0	0	0	0	2	0	0	0
MACHINERY	0	0	1	24	18	13	1	0	2	3	0	0	2	44	0	0	2	0	7	0	0	0
MILITARY	26	0	0	17	5	17	3	0	0	0	0	1	3	22	0	0	0	0	2	1	0	4
MUSIC	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	1	0	0
NATURE	14	0	0	0	11	3	14	0	1	1	0	0	4	3	0	10	4	0	1	0	3	14
OFFICE	3	5	2	5	18	20	0	4	5	4	15	5	0	16	6	0	7	10	1	0	0	0
OPEN	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	2	2	0	0	0
POLICE	3	46	0	8	1	1	0	0	1	5	0	0	21	5	0	0	1	0	1	0	0	0
PUBLIC	0	0	0	0	2	0	0	0	0	3	9	0	1	4	0	2	1	1	0	0	0	0
SCI-FI	9	30	0	0	0	0	1	0	1	14	0	1	0	20	1	0	1	0	1	0	0	37
SPORTS	2	0	0	1	11	8	0	0	3	6	13	1	1	12	0	10	0	13	12	0	0	9
TRANSPORTATION	1	5	0	43	26	47	0	0	5	8	0	1	15	72	0	26	6	1	14	2	0	2

Table 1. The distribution of 2,140 clips by its semantic (row) and onomatopoeia categories (column).

mans, the spectral centroid and the roll-off frequency measure are a measure of perceptual *brightness* of the audio signal. These features are popular in generic audio classification task [15]. By frame-based analysis the features were extracted every 10 millisecond with a Hamming window of 20 millisecond length.

3.3. Classifiers

For the audio classification task, two data-driven algorithm are used: Support Vector Machines (SVM) with radial basis kernel function (RBF) and $k-$ nearest neighbor (KNN) [16]. In SVM classification, the algorithm *learns* a separating hyperplane between the categories in a high-dimensional representation space. The algorithm proceeds by optimizing generalized error bounds [17]. The binary learning/classification procedure is extended to the multi-class problem by using the one-against-all scheme and optimization procedure as explained in [18]. The KNN classifies a given sample according to the label of its k nearest neighbors. The similarity measure between a given sample and the samples in the training set in the representation space is given by the vector dot product in the latent perceptual space as [19]:

$$Similarity(\tilde{f_k}, \tilde{f_i}) = cos^{-1}\left(\frac{(\tilde{u_k} \times \tilde{S}) \cdot (\tilde{u_i} \times \tilde{S})}{\parallel \tilde{u_k} \times \tilde{S} \parallel \cdot \parallel \tilde{u_i} \times \tilde{S} \parallel}\right)$$

Here, $\times$ is the vector-matrix product, $(\cdot)$ is the dot product between two vectors and $\parallel \quad \parallel$ is the vector length. The vector characterizing the i^{th} audio clip in the database f_i (the i^{th} row of F) is represented by $\tilde{f_i}$ the i^{th} row of $\tilde{U} \cdot \tilde{S}$ in LPS.

The classification is 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. This procedure is also used to determine the parameter value of the RBF kernel and the error-tradeoff. Additionally, an average baseline chance-level performance is also estimated. This is dependent on data distribution amongst the categories. It indicates the performance of the system if the labels of the given test clips are randomly assigned from the train set instead of predicting using the proposed approach. To represent an audio clip in the test set (not part of the initial collection of training set), the number of feature-vectors of the query in each of the N reference clusters is first estimated. This results in a N dimensional vector x similar to a row of F. This can be seen as an additional row of F, and assuming S and V remain the same, we can express:

$$x = u_x \times S \cdot V^T$$

Here u_x is the additional row in U corresponding to x. For similarity measurement we need to estimate $u_x \cdot S$. From the above equation we get the representation of the query audio clip as:

$$\tilde{x} = \tilde{u_x} \times \tilde{S} = x \times \tilde{V}$$

By using this and the similarity measure, it is possible to predict the label of an unknown clip x by the set $\{R_x\}$ that is the closest.

For each classifier, four types of evaluation were conducted according to the labeling scheme. The predicted class for a given sample from the test set is assumed to be correct if: (1) the onomatopoeia label is correct [O only], (2) the semantic label is correct [S only], (3) both onomatopoeia and semantic label are correct [O $\times$ S], (4) either onomatopoeia or semantic label is correct.[O $+$ S]. For each case the classification accuracy is calculated as:

$$\text{A} = \frac{\text{No. of correctly classified instances}}{\text{Total no. of instances in test set}} \qquad (1)$$

$$\text{F-score} = \frac{2 \cdot P \cdot R}{(P + R)} \qquad (2)$$

where,

$$P = avg\left(\frac{\text{No. of correctly classified instances in class C}}{\text{Total instances in test set}}\right) \qquad (3)$$

$$R = avg\left(\frac{\text{No. of correctly classified instances in class C}}{\text{Total instances in test set from class C}}\right) \qquad (4)$$

Here, P and R are precision and recall rates respectively. The values are averaged over the different categories. Calculating P and R for the first two cases is straightforward. In the third case, a given sample is correctly classified if its onomatopoeia and semantic labels are correctly classified. In the fourth case, a given sample is correctly classified if either its onomatopoeia or its semantic label is correctly classified. In the next section, the results, its interpretation, and further discussions are presented.

4. RESULTS AND DISCUSSION

The change in average classification accuracy as a function of number of reference clusters (value of N) is shown in figure 2. It can be seen that for both the onomatopoeia and the semantic labeling schemes the classification accuracy is well above baseline chance-level. The plot to the left shows the performance of KNN classifier. For this, $k = 3$ was experimentally determined to give the best performance. The plot to the right shows the performance of SVM using RBF kernel. For both classifiers, it can be seen that increasing N results in better classification accuracy, however the gain in performance is not significant after $N = 2000$. For the KNN classifier, $N = 8000$ was found to give best performance, and for SVM, $N = 2000$ was best. The corresponding F-score is also indicated. For SVM, a slight performance degradation can be observed for $N > 2000$, this may be specific to the classifier and the choice of the values of the parameter of the RBF kernel. The classification results for onomatopoeia categories are perceptually meaningful. Table 2 lists some examples of onomatopoeia categories and its respective most confusing categories. It can be seen that the classification is perceptually meaningful implying that proposed framework

KNN Classifier		SVM-RBF Classifier	
Category	**Confused with:**	**Category**	**Confused with:**
TAP	Clatter, Thud	TAP	Clatter, Crunch
TWEET	Squeak, Gabble	TWEET	Squeak, Hum
MEOW	Burr, Squeak	BURR	Hum, Clatter
HONK	Gabble, Beep	DONG	Beep, Buzz
THUD	Clatter, Tap	CLATTER	Tap, Hum

Table 2. Onomatopoeic categories and their most confusion categories.

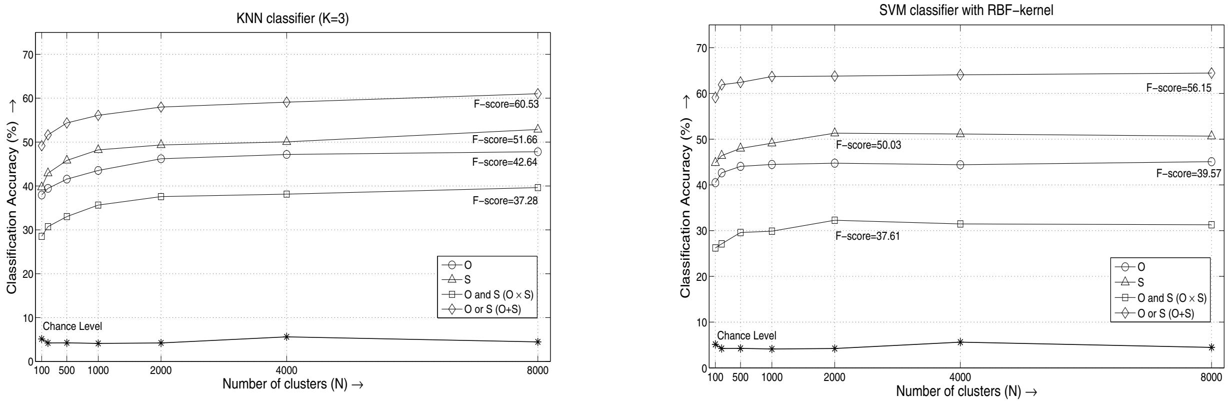

Fig. 2. Performance of classifiers as a function of number of reference clusters (N). Left- K=3 NN classifier, Right- SVM with RBF kernel

is able to handle perceptual categories of audio appropriately.

In figure 2, for both classifiers, the classification performance of the semantic labeling scheme is superior to the onomatopoeia labeling scheme by about $3 - 5\%$. This can be attributed to the difficulty in mapping specific acoustical properties of sources to onomatopoeia labels. The primary reason for this is the loose subjective definitions of onomatopoeia words. For example, how different is *Tap* from *Thud* or *Hum* from *Buzz*? Making such distinctions when manually identifying the onomatopoeia category of audio clips is difficult. In the case of semantic labels, however, the labeling can be more specific and less prone to confusion. For both classifiers, the [O+S] scheme has the highest performance (about 61% average accuracy) , and the [O×S] has the least (about 40%). This implies predicting both the onomatopoeia label and the semantic label for a given test sample is more challenging than predicting either one of them correctly. This is because for the predicted label to be correct in the [O×S] case, both the onomatopoeia label and the semantic label need to be correct (only one correct possibility for each sample), however in the [O+S] case, if either one of them is correct (three possibilities), the prediction is assumed to be correct. The performance of [O+S] > [O only] and [S only], indicating the proposed method can appropriately capture perceptual qualities of audio along with its semantic categorization. Overall, the performance of SVM and KNN classifiers are comparable except for the [O×S] case. This again may be specific to the classifier and the values of the parameters used for training.

5. CONCLUSION

In this paper, a framework for audio classification system using unit-document co-occurrence measure is presented. In the proposed method, a given clip is first characterized by quantizing the feature-vectors extracted from the audio clip into reference clusters. These clusters are derived by unsupervised clustering of the whole set of feature-vectors extracted from the library of clips. Then, using the basis derived through singular value decomposition, it maps the clip into a latent perceptual space (LPS). Classification of an unknown audio clip is performed by mapping it to this representation space. Since the initial reference clusters have distinct perceptual characteristics, the resulting vector representation of audio clips is according to their perceptual qualification in the LPS. The performance of the system was tested for two categorization schemes of the clips: by its semantic labels and by its onomatopoeic labels. While the semantic labels describe the high-level category of the audio event, the onomatopoeia labels describe the acoustic properties of the clip. The representation scheme and the classification by the dual-labeling of audio clips remain the main contributions of this work. Each scheme has its advantages and limitations, but, having them together in a single framework makes the system more flexible as each labeling scheme addresses the limitation of the other. The obtained results are encouraging and suggest that the proposed framework can handle both semantic and onomatopoeic categorization of audio. It also indicates that it appropriately captures perceptual similarities between acoustic sources.

The approach presented here can be applied to audio information retrieval, event detection, context recognition and genre classification. These ideas are a part of our ongoing and planned future work.

6. REFERENCES

[1] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," *IEEE International Conf. on Acoustic Speech and Signal Processing (ICASSP). Munich, Germany.*, vol. 2, April 1997.

[2] C. Panagiotakis and G. Tziritas, "A Speech/Music Discriminator Based on RMS and Zero-Crossings," *IEEE Trans. on Multimedia*, vol. 7, no. 1, February 2005.

[3] T. Zhang and J. Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," *IEEE Trans. on Speech and Audio Proc.*, vol. 9, no. 4, May 2001.

[4] H.J. Zhang L. Liu and H. Jiang, "Content Analysis for Audio Classification and Segmentation," *IEEE Trans. on Speech and Audio Proc.*, vol. 10, no. 7, October 2002.

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

[6] G. Guo and S. Z. Li, "Content-Based Audio Classification and Retrieval by Support Vector Machines," *IEEE Trans on Neural Nets.*, vol. 14, no. 1, January 2003.

[7] M. Slaney, "Semantic-Audio Retrieval," *International Conf. on Acoustic Speech and Signal Proc. (ICASSP), Orlando, USA.*, pp. 13–17, May 2002.

[8] L. Barrington, A. Chan, D. Turnbull, and G. Lanckriet, " Audio Information Retrieval using Semantic Similarity," *IEEE International Conf. on Acoustics Speech and Signal Proc. (ICASSP), Honolulu, Hawaii, USA.*, vol. 2, April 2007.

[9] P. Cano, M. Koppenberger, S. Le Groux, J. Ricard, P. Herrera, and N. Wack, "Nearest-Neighbor Generic Sound Classification with a WordNet-based Taxonomy," *In Proc.* 116^{th} *Audio Engineering Society (AES) Convention, Berlin, Germany.*, 2004.

[10] O. Gillet and G. Richard, "Drum Loops Retrieval from Spoken Queries," *Journal of Intelligent Info. Systems*, vol. 24, no. 2, pp. 159–177, 2005.

[11] Hugh. Bredin, "Onomatopoeia as a Figure and a Linguistic Principle," *New Literary History*, vol. 27, no. 3, pp. 555–569, 1996.

[12] S. Sundaram and S. Narayanan, " Analysis os Audio Clustering using Word Descriptions," *IEEE International Conf. on Acoustics Speech and Signal Proc. (ICASSP), Honolulu, Hawaii, USA.*, vol. 2, April 2007.

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

[14] J. R. 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.

[15] D. Li, I.K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, vol. 22, pp. 533–544, 2001.

[16] R. O. Duda, P. E. Hart, and D.G. Stork, "Pattern Classification," *Wiley-Interscience*, vol. 2nd edition, October 2000.

[17] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods," *Cambridge University Press*, 2000.

[18] H. Chih-Wei and L. Chih-Jen, "A comparison of methods for multi-class support vector machines," *IEEE Trans. on Neural Nets.*, vol. 13, no. 2, pp. 415–425, 2002.

[19] S. Sundaram and S. Narayanan, "Audio Retrieval by Latent Perceptual Indexing," *Proceedings of IEEE International Conf. on Acoustics Speech and Signal Proc. (ICASSP), Las Vegas, USA.*, 2008.