AUDIO RETRIEVAL BY LATENT PERCEPTUAL INDEXING

Shiva Sundaram and Shrikanth Narayanan

Speech Analysis and Interpretation Lab. (SAIL), Dept. of Electrical Engineering-Systems,
University of Southern California (USC), Los Angeles, California, USA.
email: ssundara@usc.edu, shri@sipi.usc.edu

ABSTRACT

We present a query-by-example audio retrieval framework by indexing audio clips in a generic database as points in a latent perceptual space. First, feature-vectors extracted from the clips in the database are grouped into *reference clusters* using an unsupervised clustering technique. An audio clip-to-cluster matrix is constructed by keeping count of the number of features that are quantized into each of the reference clusters. By singular-value decomposition of this matrix, each audio clip of the database is mapped into a a point in the latent perceptual space. This is used for indexing the retrieval system. Since each of the initial reference clusters represents a specific perceptual quality in a perceptual space (similar to words that represent specific concepts in the semantic space), querying-by-example results in clips that have similar perceptual qualities. Subjective human evaluation indicates about 75% retrieval performance. Evaluation on semantic categories reveals that the system performance is comparable to other proposed methods.

Index Terms— audio retrieval, query by example, audio indexing, audio representation, audio clustering, similarity measure.

1. INTRODUCTION

The Web 2.0 platform and the proliferation of consumer devices has lead to online availability of large amount of multimedia content on the World Wide Web. To provide efficient access to both the user and the back-end retrieval system the content needs to be organized and indexed. This is usually implemented by searching through the short text caption/tags included with the media file. Examples of commercial systems for image/video that rely on indexing text are YouTube and Flickr.

In the case of audio, the human auditory system predominantly relies on perception [1]. The recognition of the actual event generating the acoustic signal is an additional context dependent process. Since text captions typically describe only the higher level event/ content, it is not possible to derive perceptual similarity between two acoustic events. For example, (from the sound effects database [12]) the events "PAPER WRAPPING GIFT PAPER" and "MANUAL TOOTHBRUSH BRUSHING TEETH" have unrelated descriptions, yet considering only the underlying acoustics, they are similar sounds. To have full duplexity in the retrieval system, it needs to be able to compute both in terms of text description and signal-level measures. Text-only based indexing (albeit its usefulness) solves only one part of the retrieval problem. In this respect, the work presented in this paper addresses the other aspect of retrieval and focuses on an example-based audio retrieval that is perceptually more relevant.

A retrieval system must ideally handle the content in a way that is relevant to both its perception and its description. Therefore, it is desirable to move away from methods that implement naive signal-based labelling/modelling and similarity measure (implemented in classical content-based audio analysis methods [2, 3, 7]) to a more perceptually and semantically meaningful measure. Examples of methods that deal with semantic aspects of audio retrieval are [4, 5, 6]. In [4] 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 [5], the authors have a similar approach of modelling features with text labels in the captions. Other techniques for retrieval using semantic relations in language include [6]. 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 this paper, a query-by-example audio retrieval system that addresses the perceptual issues mentioned earlier is presented. The main contributions of this work are as follows. First, in contrast to methods such as in [5, 7], the framework presented here does not deal with training models for explicit class definitions (such as *music, stationary noise, speech* etc.). Instead, the framework is implemented and evaluated using a generic sound effects audio database [12] that covers a wide variety of domains. Here, a whole audio clip is represented as a single vector in a latent perceptual space (LPS). This makes the computationally intensive signal-based similarity measure manageable. The method also brings out an underlying latent perceptual structure of audio clips and measures similarity based on this. The performance of the retrieval system is measured in two ways. First, subjective human evaluation (by listening) is performed on a set of test audio clips. For this case, however, since no explicit categories are defined the performance for the top 5 matching retrieved clips is estimated. Human evaluation by listening is a more stringent and conservative, but it is a better performance metric for the perceptually motivated framework proposed here. Then, for semantic evaluation the available audio database is split into 20 mutually exclusive high-level categories (such as *airplane, crowd, construction, industry* etc.). The categories were derived using the available text captions. Although semantic evaluation does not address perceptual aspects motivated here, it is presented here to allow comparison with contemporary retrieval by example methods. Additionally, examples of query clips and retrieved audio clips are also illustrated.

In the proposed system, a *bag of feature-vectors* is extracted from an audio clip. Then it 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 leads to 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

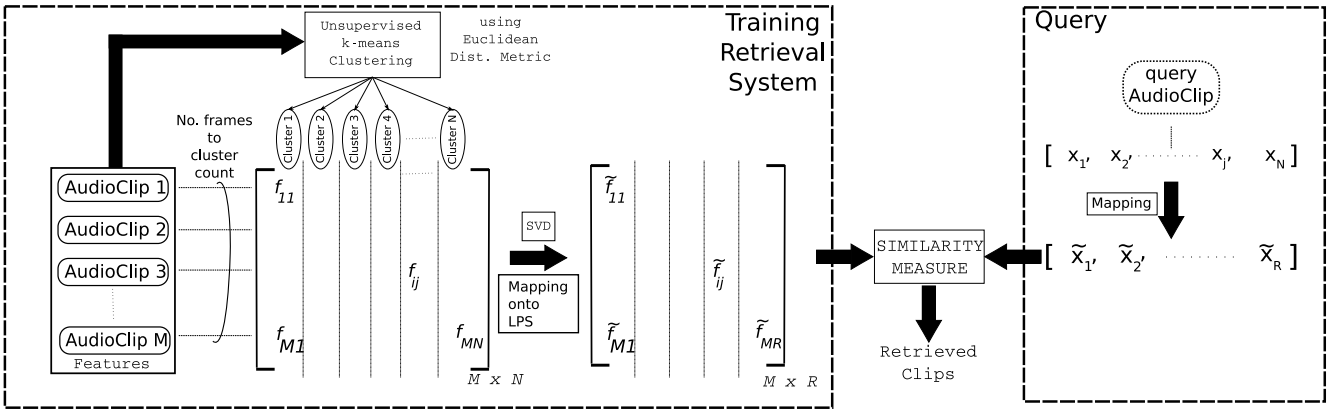

Fig. 1. Indexing and retrieval of clips in latent perceptual space.

extracted from the clips in the database. By singular-value decomposition (SVD), this sparse representation is mapped to points in the LPS. Thus each audio clip is represented as a single vector in the the perceptual space. We demonstrate the framework by applying it to a large generic sound effects database (≈9000 audio clips) using commonly used perceptual signal features. This approach is similar to latent semantic analysis (LSA) of text documents [8], where words represent conceptual entities that occupy distinct volumes in the latent semantic space. Here, we suppose that the volume occupied by clusters of signal features represent specific perceptual qualities in the LPS. This method is illustrated in figure 1 and its implementation is further explained in the next section.

2. IMPLEMENTATION

2.1. Algorithm

Let's 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 labelled 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. This is analogous to text document representation by LSA with term-document frequency. Therefore, ideas of similarity measurement and representation of a query can be re-applied here.

2.2. Similarity Measure

As mentioned in [8], the row vectors (corresponding to audio clips) are projected on to the basis formed by the columns of the matrix V. Thus, 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. Using a cosine metric, the similarity between k and i audio clip can be expressed as the angle between the vectors, i.e.:

$$Similarity(\tilde{f}_k, \tilde{f}_i) = cos^{-1}\left(\frac{(\tilde{u_k} \times \tilde{S}) \cdot (\tilde{u_i} \times \tilde{S})}{\| \tilde{u_k} \times \tilde{S} \| \cdot \| \tilde{u_i} \times \tilde{S} \|}\right)$$

Here, $\times$ is the vector-matrix product, $(\cdot)$ is the dot product between two vectors and $\| \quad \|$ is the vector length.

2.3. Query Representation

To represent a query audio clip in LPS (not part of the initial collection), 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 the similarity measure in section 2.2, it is possible to retrieve the set of $\{R_x\}$ that are close to the query x.

Since the similarity is not calculated directly in the feature space, it makes comparison of two audio clips significantly more manageable. After the initial clustering and SVD (can be performed offline), since $N < T_i \forall i \in \{1, \ldots, M\}$, retrieving clips based on this similarity measure is easily tractable.

2.4. Relationship with the LSA framework

While the method presented here is similar to the LSA framework there are some differences which are discussed here for clarity. As stated in [8], LSA 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 features extracted from a given database, the method presented here seeks distinct acoustic clusters in the perceptual space. The significance of these acoustic clusters in the perceptual space are analogous to concepts in the semantic space. Therefore, the resulting similarity is a measure of closeness in the perceptual structure between two clips.

In the next section, the details of the experiments are provided. Followed by the results obtained. Finally, an interpretation of the results, and the methodology is provided.

3. EXPERIMENTS

A collection of 9100 whole audio clips (average length: 47.62 seconds, minumum: 1.47 seconds maximum: 370.16 seconds) was obtained from the General 6000 sound effects library [12]. For subjective human evaluation, 100 clips were randomly selected and retained as query files and the remaining $M = 9000$ were used for the back-end. For semantic evaluation the clips were grouped into 20 high-level categories. These are then split into 90% training (for the back-end) and 10% test (for the queries) clips. All the clips (available in 44.1kHz, stereo) were first converted to 16.0 kHz mono-channel tracks. For each clip, a set of 14 dimensional feature-vectors was extracted every 10 milliseconds using a Hamming window of 20 milliseconds size. The 14 dimensions comprised of 12 Mel-frequency cepstral coefficients (MFCC), Spectral Centroid (SC) and Spectral Roll-off frequency (SRF). MFCCs model the front-end of the human auditory system and SC and SRF measure the spectral content such as *brightness* of the audio clip. These features are popular in generic audio classification tasks. The silence frames were eliminated by appropriately thresholding the root-mean-squared energy measure. The details of the feature-vectors and their performance in audio classification tasks can be found in [9].

As mentioned in section 2.1 STEP 1, the extracted features are first grouped into N reference clusters using the k-means algorithm. The value of N was experimentally determined to be $N = 1450$, a value that maximizes the Bayesian information criterion [10]. This results in a $N = 1450$ dimension vector for each audio clip, resulting in a 9000×1450 sparse matrix F. By SVD, and retaining the largest singular-values that contain $> 90\%$ of the variance, we obtain a reduced dimensional representation of $R = 792$ for the similarity measure.

Subjective Evaluation: By the steps described in section 2.1 and 2.3, the selection of 100 clips is first represented in the LPS. Then using the similarity measure in section 2.2, for each query clip a list of 5 closest matching clips is obtained. Examples of query clips (the text caption only) and the 5 best matching retrieved clips are presented in table 1.

Seven subjects evaluated the retrieval system. The 100 query clips and the ordered list of top 5 best matching clips was presented to them one by one using a web-page interface. For each query they were instructed to select the retrieved clips that sounded similar to it. Alternatively, they could also choose a "None of them similar" option if they determined that none of the retrieved clips sounded similar to the query.

Semantic Evaluation: Although *semantic* evaluation requires detailed analysis of the concomitant text captions, the scope of this evaluation is restricted here by only considering 20 high-level categories: *airplane, animal, applause, auto, bird, boat, construction, crowd, electronic, explosion, fire, footsteps, gun, industry, metal, motorcycle, music, telephone, traffic, water*. For the evaluation, about 100 clips were randomly selected for each category from the complete database. The selection was made by choosing the appropriate category label (such as *industry*) from a clip's caption.

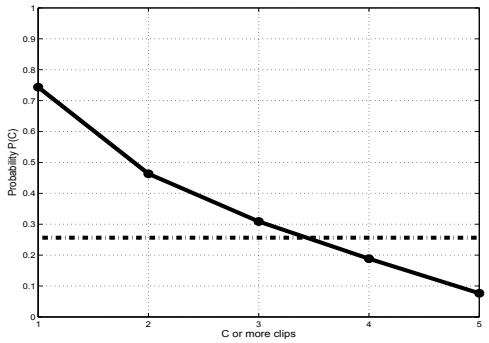

Fig. 2. **Subjective Evaluation:** Probability of retrieving $\geq C$ relevant clips. Dotted line represents the probability of retrieving 0 relevant clips.

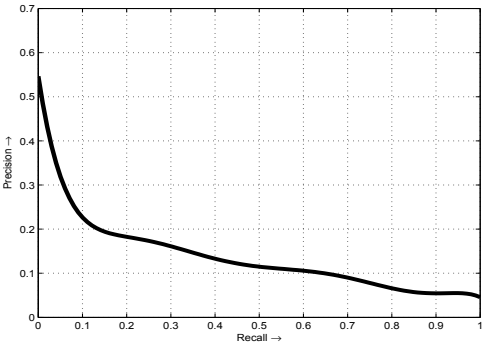

Fig. 3. **Semantic Evaluation:** Precision v/s Recall for high-level categories.

The results of subjective human evaluation and semantic evaluation is presented next.

4. RESULTS

Table 1 lists descriptions of four query clip examples and the corresponding retrieved clips. From the descriptions, it can be seen that the retrieved clips are perceptually related to the query. For example, in query clip 4, metallic sound of chain dropping on wooden surface is indeed similar to metallic *clank* sound and in query 3, the bird sounds are related to the tonal "siren" sound presented in the query. Additional examples, and audio clips can be found at [`http://sail.usc.edu/AIRdemo.html`].

Subjective Evaluation: Seven users evaluated 100 query clips by listening to 5-best retrieved clips for each query. This is taken as $7 \times 100 = 700$ samples for estimating the performance. As mentioned previously, no explicit categories can be defined for groups of perceptually similar clips. Therefore, instead of the conventional precision and recall rates for retrieval, the performance is measured in terms of probability of retrieving C or more relevant clips in the $5-$ best matching clips. Figure 2 shows the estimated probability as a function of C. It can be seen that the P(retrieving $\geq$ 3 relevant clips) $>$ P(retrieving 0 clips) by 20%. Also, the probability of retrieving *at least* 1 relevant clip in 5-best list is ≈ 0.75. This essentially is a worst-case measure of the retrieval system.

Semantic Evaluation: For the high-level categories, the precision and recall rates obtained is illustrated in figure 3. The retrieval performance is much better than the 5% chance level for the 20 cat-

Query clip 1: **AUTO RACE PIT AMBIENCE BEFORE RACE** 1. SUBWAY EXTERIOR PULL INTO STATION STOP EXIT STATION TRAIN 2. ORCHESTRA WARMING UP IN CONCERT HALL 3. AUTO RACE INDY TIER AMBIENCE OVERALL PERSPECTIVE 4. BAR PUB SMALL CROWD AMBIENCE 5. RESTAURANT LARGE CROWD . . .	Query clip 2: **CHAIR WOOD SIT DOWN IN WOODEN CHAIR** 1. FOOTSTEPS METAL MALE LEATHER SOLE WALK 2. CHAIR LAWN SIT DOWN IN LAWN CHAIR 3. BRIEFCASE UNLOCK CLASP OFFICE 4. CHAIR KITCHEN SIT DOWN IN KITCHEN CHAIR 5. BOTTLE SOFT DRINK REMOVE SCREW LID SODA . . .
Query clip 3: **SIREN SIREN POLICE AMBULANCE FIRE TRUCK** 1. SIREN CONSTANT YELPING FOR EMERGENCY VEHICLE 2. SIREN FIVE SIRENS SIMULTANEOUSLY WAILS AND YELPS 3. SIREN WAIL SIREN POLICE AMBULANCE FIRE TRUCK 4. SIREN CONSTANT WAILS AND YELPS 5. FOREST VENEZUELA VENEZUELA DAYTIME AMBIENCE BIRDS . . .	Query clip 4: **TWO SWORDS CLANKING TOGETHER SINGLE HIT** 1. TWO SWORDS CLANKING TOGETHER SINGLE HIT 2. SWORD SLIDE INTO SHEATH 3. SWORD REMOVE FROM SHEATH 4. CHAIN DROP SMALL CHAIN DROP TO WOOD SURFACE 5. SWORD TWO SWORDS SCRAPING . . .

Table 1. Query examples and corresponding 5 best matching retrieved clips.

egories (about 100 files as test sample from each category).
The average precision and recall rates were calculated by using the formula $Precision = \left(\frac{R_{correct}}{R_{retrieved}}\right)$ and $Recall = \left(\frac{R_{correct}}{R_{category}}\right)$. Here $R_{retrieved}$ is the number of clips retrieved for a query, $R_{correct}$ is the number of correctly retrieved clips for the query, and $R_{category}$ is the number of clips in the database that belong to the same category as the query. Recall is varied from 0 to 1.0 by considering more and more clips from the ordered list of retrieved clips. The average precision is calculated by averaging the precision values at every instance of correctly retrieved clip in the list. The average precision values of each test files were again averaged over 10-fold cross-validation of the 90/10 train/test split data.

5. CONCLUSION AND DISCUSSION

In this paper, a framework for a query-by-example audio retrieval system is presented. The system first characterizes a given clip in terms of reference clusters. Then, using the basis derived through singular value decomposition, it maps the clip into a latent perceptual space (LPS). It is able to retrieve matching clips using a vector similarity measure in this space. Since the reference clusters and the basis from SVD are derived only once as an offline procedure, the clip-to-clip similarity measure is made more manageable as a vector dot product. Also, since the initial reference clusters have distinct perceptual characteristics, the resulting vector representation of audio clips are indexed according to their perceptual qualification in the LPS. We obtained encouraging results. Semantic evaluation using high-level categories results in performance that is comparable to other methods [5]. The proposed method has the additional advantage of being computationally less demanding. Underperformance as compared to content-based methods can be attributed to perceptually overlapping categories (e.g. *boat*, *motorcycle*, *auto* and *industry*, *construction*). Additionally, the subjective human evaluation results indicate that the system is indeed able to retrieve relevant clips that sounds similar to a given query.

For evaluating system performance, it should be noted that text based retrieval is fundamentally different from example based retrieval. Text descriptions relate to semantic concepts that occupy dense volumes and are sparsely spread in the continuous semantic space, i.e, text descriptions are more specific and less overlapping. This makes it is easier to resolve particular aspects of descriptions of audio by including appropriate keywords in the query. Example based queries, however, are loose, less dense and more evenly spread in the acoustic/perceptual space. This is especially true in the case of complicated acoustic events as queries, because they posses a variety of perceptual qualities over time and this results in a query that is more *spread* in the perceptual space. Based on this reasoning, subjective evaluations, in general are stringent and conservative estimates of system performance.

Additionally, each subject has a different definition of similarity. Our experience indicate that in some cases subjects specifically look for content similarity, in spite of similarities in perceptual qualities with other clips.

The proposed framework can be extended in many ways. Vector representation of an audio clip, and and its text description can be derived. Since both vectors represent the same clip in different spaces, a one-to-one mapping of an audio clip in LPS to the semantic space can be established [11]. Using the proposed similarity measures a fully-duplex retrieval system can be implemented where it is possible to retrieve audio clips from a database using both text queries and example queries. Other applications include audio clustering, auditory scene classification, and even audio-based video indexing. These extensions are part of our ongoing and future work.

6. REFERENCES

[1] M. S. Lewicki, "Efficient coding of natural sounds", in Nature Neuroscience, Vol. 5,No. 4, pp 356-363, 2002.

[2] 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, Jan. 2003.

[3] L. Liu, H.J. Zhang and H. Jiang,"Content Analysis for Audio Classification and Segmentation," IEEE Trans. on Speech and Audio Processing, Vol.10, No.7, October, 2002.

[4] M. Slaney, "Semantic-Audio Retrieval," Presented at the ICASSP, Orlando, Florida, USA. May 13-17,2002.

[5] L. Barrington, et.al. ,"Audio Information Retrieval using Semantic Similarity," Presented at the ICASSP, Hawaii, USA. 2007.

[6] P. Cano, M. Koppenberger, et.al. "Nearest-neighbor generic sound classification with a WordNet-based taxonomy",In Proc. 116^{th} Audio Eng. Society (AES) Convention, Germany, 2004.

[7] M. Helen, T Viratnen,"Audio Information Retrieval using Semantic Similarity," Presented at ICASSP, Hawaii, USA. 2007.

[8] J. Bellegarda, "Latent semantic mapping", IEEE Signal Processing Magazine Vol. 22, Issue. 5, September 2005.

[9] D. Li, et. al ,"Classification of general audio data for content-based retrieval," Pattern Recognition Letters, Vol.22, 533-544, 2001.

[10] S. .S Chen and P. S. Gopalakrishnan ,"Clustering Via the Bayesian Information Criterion with applications in Speech Recognition". In Proc. of the ICASSP Vol.2 ,12-15 May 1998.

[11] S. Sundaram and S. Narayanan ,"Analysis of Audio Clustering using Words". Presented at ICASSP, Hawaii, USA. 2007.

[12] "The Series 6000 General Sound Effect Library." http://www.sound-ideas.com/6000.html