AUDIO INFORMATION RETRIEVAL USING SEMANTIC SIMILARITY

Luke Barrington[1]*, Antoni Chan*[1]*, Douglas Turnbull*[2] *& Gert Lanckriet*[1]

University of California, San Diego
[1]Electrical and Computer Engineering, [2]Computer Science and Engineering Department
9500 Gilman Drive, La Jolla, CA 92093

ABSTRACT

We improve upon query-by-example for content-based audio information retrieval by ranking items in a database based on *semantic* similarity, rather than acoustic similarity, to a query example. The retrieval system is based on semantic concept models that are learned from a training data set containing both audio examples and their text captions. Using the concept models, the audio tracks are mapped into a semantic feature space, where each dimension indicates the strength of the semantic concept. Audio retrieval is then based on ranking the database tracks by their similarity to the query in the semantic space. We experiment with both semantic- and acoustic-based retrieval systems on a sound effects database and show that the semantic-based system improves retrieval both quantitatively and qualitatively.

Index Terms— computer audition, audio retrieval, semantic similarity

1. INTRODUCTION

It is often joked that "writing about music is like dancing about architecture". Explaining the intangible qualities of an auditory experience using words is an ill-posed problem with many different solutions that might satisfy some, and few or none that are truly objective. Yet using semantics is a compact medium to describe what we have heard, and a natural way to describe content that we would like to hear from an audio database. An alternative approach is query-by-example (QBE), where the user provides an audio example instead of a semantic description and the system returns audio content that is similar to the query. The key to any QBE system is in the definition of audio *similarity*.

Many approaches to audio information retrieval consider similarity in the audio domain by comparing features extracted from the audio signals. In [1], songs are represented as HMM's trained on timbre- and rhythm-related features, and song similarity is defined as the likelihood of the query features under each song model. Similarly in [2], each song is represented as a probability distribution of timbre feature vectors, and the audio similarity is based on the Kullback-Leibler divergence between the query feature distribution and those of the database. Finally, state-of-the-art genre classification results [3], based on nearest-neighbor clustering of spectral features, suggest that the returns of purely acoustic approaches are reaching a ceiling and that a higher-level understanding of the audio content is required.

In many cases, semantic understanding of an audio query enables retrieval of audio information that, while *acoustically* different, is *semantically* similar to the query. For example, given a query of a high-pitched, warbling bird song, a system based on acoustics might retrieve other high-pitched, harmonic sounds such as a baby crying. On the other hand, the system based on semantics might retrieve sounds of different birds that hoot, squawk or quack.

Indeed, recent works based on semantic similarity have shown promise in improving the performance of retrieval systems over those based purely on acoustic similarity. For example, the acoustic similarity between pieces of music in [2] is combined with similarities based on meta-data, such as genre, mood, and year. In [4], the songs are mapped to a semantic feature space (based on musical genres) using a neural network, and songs are ranked using the divergence between the distribution of semantic features. In the image retrieval literature, [5] learns models of semantic keywords using training images with ground-truth annotations. The images are represented as semantic multinomials, where each feature represents the strength of the semantic concept in the image. Results from [5] show that this retrieval system returns more meaningful images than a system based on visual similarity. For example, a query of a red sunset image returned both red sunsets and orange sunsets, while the retrieval system based on visual similarity returned only red sunsets.

In this paper, we present a query-by-example retrieval system based on semantic similarity. While any semantic annotation method could be used, we base our work on the models of [6, 7] which have shown promise in the domains of audio and image retrieval. In Section 2, we present probabilistic models for the audio tracks and their semantic labels, and in Section 3, we discuss how to use the models for retrieval based acoustic similarity and semantic similarity. Finally, in Section 4 we compare the two retrieval methods using experiments on a sound effects database.

2. MODELING AUDIO AND SEMANTICS

Our audio models are learned from a database composed of audio tracks with associated text captions that describe the audio content:

$$\mathcal{D} = \{(\mathcal{A}^{(1)}, \mathbf{c}^{(1)}), ..., (\mathcal{A}^{(|\mathcal{D}|)}, \mathbf{c}^{(|\mathcal{D}|)})\} \quad (1)$$

where $\mathcal{A}^{(d)}$ and $\mathbf{c}^{(d)}$ represent the d-th audio track and the associated text caption, respectively. Each caption is a set of words from a fixed vocabulary, $\mathcal{V}$.

2.1. Modeling Audio Tracks

The audio data for a single track is represented as a *bag-of-feature-vectors*, i.e. an unordered set of feature vectors $\mathcal{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_{|\mathcal{A}|}\}$ that are extracted from the audio signal. Section 4.1 describes our particular feature extraction methods.

Each database track d is compactly represented as a probability distribution over the audio feature space, $P(\mathbf{a}|d)$. The track distribution is approximated as a K-component Gaussian mixture model (GMM);

$$P(\mathbf{a}|d) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{a}|\mu_k, \Sigma_k),$$

where $\mathcal{N}(\cdot|\mu, \Sigma)$ is a multivariate Gaussian distribution with mean μ and covariance matrix Σ, and π_k is the weight of component k in the mixture. In this work, we consider only diagonal covariance matrices since using full covariance matrices can cause models to overfit the training data, while scalar covariances do not provide adequate generalization. The parameters of the GMM are learned using the Expectation Maximization (EM) algorithm [8].

2.2. Modeling Semantic Labels

The semantic feature for a track, $\mathbf{c}$, is a *bag of words*, represented as a binary vector, where $\mathbf{c}_i = 1$ indicates the presence of word w_i in the text caption. While various methods have been proposed for annotation of music [6, 9] and animal sound effects [10], we follow the work of [6, 7] and learn a GMM distribution for each semantic concept w_i in the vocabulary. In particular, the distribution of audio features for word w_i is an R-component GMM;

$$P(\mathbf{a}|w_i) = \sum_{r=1}^{R} \pi_r \mathcal{N}(\mathbf{a}|\mu_r, \Sigma_r),$$

The parameters of the semantic-level distribution, $P(\mathbf{a}|w_i)$, are learned using the audio features from every track d, that has w_i in its caption $\mathbf{c}^{(d)}$. That is, the training set $\mathcal{T}_i$ for word w_i consists of only the *positive* examples:

$$\mathcal{T}_i = \{\mathcal{A}^{(d)} : \mathbf{c}_i^{(d)} = 1, d = 1, \ldots, |\mathcal{D}|\}$$

Learning the semantic distribution directly from all the feature vectors in $\mathcal{T}_i$ can be computationally intensive. Hence, we adopt one of the strategies of [7] and use naive model averaging to efficiently and robustly learn word-level distributions by combining all the track-level distributions $P(\mathbf{a}|d)$ associated with word w_i.

The final semantic model is a collection of word-level distributions $P(\mathbf{a}|w_i)$, that models the distribution of audio features associated with the semantic concept w_i.

3. AUDIO RETRIEVAL BY EXAMPLE

In this section, we describe two systems for retrieving audio by query example. While the first is based on retrieving audio that is *acoustically* similar to the query, the second utilizes the semantic word models to retrieve audio tracks that are *semantically* similar to the query track.

3.1. Query by acoustic example

The query-by-acoustic-example (QBAE) system is based on retrieving audio that is acoustically similar to the query. The score used to rank the similarity of database tracks to the query track is based on the likelihood of the audio features of the query under the database track distributions. Intuitively, the database tracks are ranked according to how likely the query features were generated from the particular database track. Formally, given the features from the query track, $\mathcal{A}^{(q)}$, the likelihoods are computed for each database track, $d = 1, \ldots, |\mathcal{D}|$,

$$\ell_d = P(\mathcal{A}^{(q)}|d) = \prod_{i=1}^{|\mathcal{A}^{(q)}|} P(\mathbf{a}_i^{(q)}|d).$$

We make the unrealistic naive Bayes assumption of conditional independence between audio feature vectors. Attempting to model the temporal dependencies between audio feature vectors may be infeasible due to computational complexity and data sparsity.

The database tracks are rank ordered by decreasing likelihood. Note that retrieval by acoustic example is computationally intensive because it requires computing the likelihood of a large set of features (on the order of tens of thousands) under the track models for each track in the database.

3.2. Query by semantic example

In contrast to QBAE, the query-by-semantic- example (QBSE) paradigm [5] utilizes semantic information to retrieve semantically meaningful audio from the database. QBSE is based on representing an audio track as a semantic feature vector, where each feature represents the strength of each semantic concept from a fixed vocabulary $\mathcal{V}$. For example, the semantic representation of the sound of a gun firing might have high

Table 1. Mean average precision for query-by-semantic-example (QBSE) and query-by-acoustic-example (QBAE).

	QBSE	QBAE
MAP	0.186±.003	0.165±.001

values in the "shot", "weapon" and "war" semantic dimensions, and low values for "quiet", "telephone" and "whistle".

The semantic feature vector is computed using an annotation system that assigns a weight for the presence of each semantic concept. Although any annotation system that outputs weighted labels could be used, when using the probabilistic word models described in the previous section, the semantic feature vectors are multinomial distributions with each feature equal to the posterior probability of that concept occurring given the audio features. Formally, given the audio features $\mathcal{A}$, the semantic multinomial is $\pi = \{\pi_1, \ldots, \pi_{|\mathrm{V}|}\}$ with each entry given by;

$$\pi_i = P(w_i|\mathcal{A}) = \frac{P(\mathcal{A}|w_i)P(w_i)}{\sum_{j=1}^{|\mathcal{V}|} P(\mathcal{A}|w_j)P(w_j)}$$

where we have applied Bayes' rule to compute the posterior.

The semantic multinomials are points in a probability simplex or *semantic space*. A natural measure of similarity in the semantic space is the Kullback-Leibler (KL) divergence [11] between the semantic multinomials;

$$\mathrm{KL}(\pi^{(q)} \| \pi^{(d)}) = \sum_{i=1}^{|\mathcal{V}|} \pi_i^{(q)} \log \left(\frac{\pi_i^{(q)}}{\pi_i^{(d)}} \right)$$

Query-by-semantic-example is performed by first representing the database tracks as semantic multinomials, and then, given a query track, retrieving the database tracks that minimize the KL divergence with the query. The bulk of QBSE computation lies in calculating the semantic distribution for the query track so that complexity grows with the size of the vocabulary rather than with the size of the database in QBAE.

In practice, some regularization must be applied to the semantic multinomials in order to avoid taking the log of zero. This regularization is achieved by adding a small positive constant (10^{-3} in this work) to all the multinomial dimensions and renormalizing. This is equivalent to assuming a uniform Dirichlet prior for the semantic multinomial.

4. EXPERIMENTS

4.1. Semantic and Audio Features

This work examines queries on a general sound effects corpus taken from 38 audio compact discs of the BBC Sound Effects library. Our data set comprises 1305 audio tracks (varying in length from 3 seconds to 10 minutes) with associated descriptive text captions up to 13 words long.

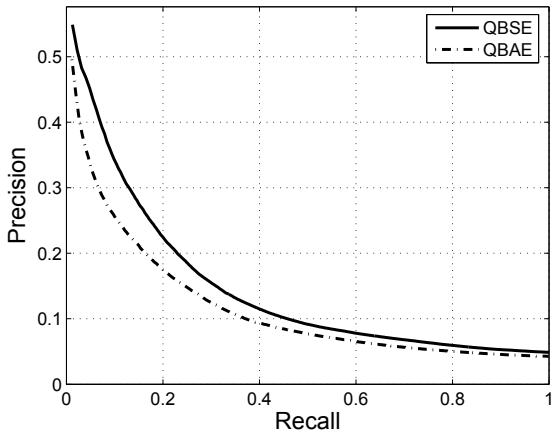

Fig. 1. Precision-Recall curves for query-by-semantic-example (QBSE) and query-by-acoustic-example (QBAE).

Each sound effect's caption, $\mathbf{c}$, is represented as a *bag of words*: a set of words that are found in both the track caption and our vocabulary $\mathcal{V}$. The vocabulary is composed of all terms which occur in the captions of at least 5 sound effects and does not include common stop words (e.g. 'the', 'into', 'a'). In addition, we preprocess the text with a custom stemming algorithm that alters suffixes so that semantically similar words (e.g., 'bicycle', 'bicycles', 'bike' and 'cycle') are mapped to the same semantic concept. The result is a vocabulary with $|\mathcal{V}| = 348$ semantic concepts. Each caption contains on average 3.7 words from the vocabulary.

For each 22050Hz-sampled, monaural audio track in the data set, we compute the first 13 Mel-frequency cepstral coefficients as well as their first and second instantaneous derivatives for each half-overlapping short-time (∼12 msec) segment [12], resulting in about 5000 39-dimensional feature vectors per 30 seconds of audio content.

4.2. Results

For each query track, our system orders all database tracks by their similarity to the query. Evaluation of this ranking (and of most auditory similarity systems) is difficult since acoustic and semantic similarity is a subjective concept. Rather than rely on qualitative evaluation, we divide the data into 29 disjoint categories (corresponding to the categories of the BBC sound effects CDs) and consider all audio tracks within the same category to be similar. This allows us to compute precision and recall for the database ranking due to each query track. Given a query track from category G, if there are $|G_T|$ total tracks from category G in the database and the system returns $|G_{auto}|$ tracks from that category, where $|G_C|$ are correct, recall and precision are given by: $recall = \frac{|G_C|}{|G_T|}$, $precision = \frac{|G_C|}{|G_{auto}|}$. Average precision is found by moving down this ranked list (incrementing $|G_{auto}|$) and averaging the precisions at every point where a new track is cor-

Table 2. Sample queries and retrieved database tracks using query-by-semantic-example (QBSE) and query-by-acoustic-example (QBAE). Words in *italics* are dimensions of our semantic vocabulary. Words in **bold** overlap with the query caption.

	BBC SFX Class	Caption
Query	Birds	willow *warbler singing*
QBSE	Birds	*birds* and *waterfowl*, roseate cockatoos, Australia
	Birds	*birds* and *waterfowl*, flamingoes, Caribbean
	Birds	*birds* and *waterfowl*, coot with geese and mallard
QBAE	Babies	*month* old *boy*, *screaming* tantrum
	Babies	*month* old *boy*, *words*, daddy
	Babies	*year* old *boy*, *screaming* temper
Query	Household	*electric drill*, *single* hole *drilled*
QBSE	Household	***electric drill***, series of holes in quick succession
	Sound Effects	at the *dentist*, *high speed* ***drilling***
	Bang	*quarrying*, *road* ***drill***, with compressor
	Sound Effects	at the *dentist*, *low speed* ***drilling***
QBAE	Household	***electric drill***, series of holes in quick succession
	Household	***electric*** circular saw
	Sports and Leisure	skiing *cross* country
	Babies	*week* old *boy*, hysterical *crying*
Query	Farm Machinery	*landrover safari diesel*, *horn* six *short blasts*, *exterior*
QBSE	Farm Machinery	***landrover safari diesel***, *door opened*
	Farm Machinery	***landrover safari diesel***, ***horn*** two long ***blasts***, ***exterior***
	Comedy Fantasy and Humor	***horn*** sounded twice
	Farm Machinery	***landrover safari diesel***, *door* slammed *shut*
QBAE	Transport	***diesel*** *lorry*, 10-*ton*, ***exterior***, *approach*, *stop*, *switch* off
	Household	*domestic chiming clock*, quarter-hour *chime*
	Sports and Leisure	rugby county *match* with *large crowd* with scrums
	Sound Effects	*footsteps*, group of young *people walking* in park

rectly identified. The mean average precision (the mean over all tracks) for QBSE and QBAE are shown in Table 1 and precision-recall curves are displayed in Figure 1. Results are averaged over 10-folds of cross-validation where 90% of the audio tracks are used to compute the word-level models and the remaining 10% are used as testing examples for querying the retrieval system.

The quantitative results show the difficulty of the audio query-by-example task. Sound effects from different BBC categories often have strong similarities (e.g., {""Livestock", Dogs" and "Horses"} or {"Cities", "Exterior Atmospheres" and "Human Crowds"}) and many tracks could easily fit in multiple categories. Without a reliable ground-truth, automatically evaluated results are bound to be poor. Though recall and precision scores are low, QBSE shows a significant improvement over QBAE (e.g., a 26% relative improvement in precision at 0.1 recall). Table 2 illustrates the results of both QBSE and QBAE for a number of example audio queries. It can be seen that, while tracks returned by QBAE could be expected to sound similar to the query, the results of QBSE have more semantic overlap and often return database tracks that might sound different (e.g., the low-pitched sound of a road drill in response to a high-pitched query of an electric drill) but have a strong semantic connection.

5. REFERENCES

[1] T. Zhang and C.-C. Jay Kuo, "Classification and retrieval of sound effects in audiovisual data management," *Asilomar Conference on Signals, Systems, and Computers*, 1999.

[2] F. Vignoli and S. Pauws, "A music retrieval system based on user-driven similarity and its evaluation," *ISMIR*, 2005.

[3] Elias Pampalk, Arthur Flexer, and Gerhard Widmer, "Improvements of audio-based music similarity and genre classification," *ISMIR*, 2005.

[4] Adam Berenzweig, Beth Logan, Daniel P.W. Ellis, and Brian Whitman, "A large-scale evalutation of acoustic and subjective music-similarity measures," *Computer Music Journal*, 2004.

[5] Nikhil Rasiwasia, Nuno Vasconcelos, and Pedro J Moreno, "Query by semantic example," *ICIVR*, 2006.

[6] D. Turnbull, L. Barrington, and G. Lanckriet, "Modelling music and words using a multi-class naïve bayes approach," *ISMIR*, 2006.

[7] G. Carneiro and N. Vasconcelos, "Formulating semantic image annotation as a supervised learning problem," *IEEE CVPR*, 2005.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society B*, vol. 39, pp. 1 – 38, 1977.

[9] B. Whitman and D. Ellis, "Automatic record reviews.," *ISMIR*, 2004.

[10] Malcolm Slaney, "Mixtures of probability experts for audio retrieval and indexing," *IEEE Multimedia and Expo*, 2002.

[11] Thomas Cover and Joy Thomas, *Elements of Information Theory*, Wiley-Interscience, 1991.

[12] C. R. Buchanan, "Semantic-based audio recognition and retrieval," M.S. thesis, School of Informatics, University of Edinburgh, 2005.