SEMANTIC SIMILARITY FOR MUSIC RETRIEVAL

Luke Barrington†, Douglas Turnbull⋆, David Torres⋆, Gert Lanckriet†
†Dept. of Electrical and Computer Engineering ⋆Dept. of Computer Science and Engineering
Computer Audition Laboratory
University of California, San Diego
lbarring@ucsd.edu, dturnbul@cs.ucsd.edu, datorres@cs.ucsd.edu, gert@ece.ucsd.edu

ABSTRACT

We present a query-by-example system for content-based music 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 the CAL-500 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 similarity and retrieval is then based on ranking the database tracks by their similarity to the query in the semantic space.

1 MODELING AUDIO AND SEMANTICS

Our query-by-example music information retrieval (MIR) system takes an audio track as a query and retrieves new audio tracks that have similar *semantic descriptions* to the query track. For example, given a piece of music that a listener might describe as "crazy guitar rock with a screaming female singer that makes me want to get up and dance", our system ranks all retrievable songs by how well they fit this description.

The system is based on the models of [9, 3] which have shown promise in the domains of audio and image retrieval. Audio models are learned from a database 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, $\mathcal{V}$.

We train our system using the semantic labels from the CAL-500 data set [9] of 500 songs, each annotated by at least 3 humans using up to 200 words. We require that each word be positively associated with at least 10 songs, resulting in a vocabulary of 146 words ($|\mathcal{V}| = 146$).

1.1 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. For each 22050Hz-sampled, monaural audio track, 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 [2], resulting in about 5000 39-dimensional feature vectors per 30 seconds of audio content.

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 [5].

1.2 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 [8, 11] and animal sound effects [7], we follow the work of [8, 3] 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:

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

Learning the semantic distribution directly from all the feature vectors in $\mathcal{T}_i$ can be computationally intensive. Hence, we adopt the strategy of [3] and use an extension of EM, the hierarchical EM algorithm [10], to efficiently and robustly learn word-level distributions $P(\mathbf{a}|w_i)$ from 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.

2 QUERY BY SEMANTIC EXAMPLE

Query-by-semantic-example (QBSE) is an information retrieval method that has been applied to images [6], sound effects [1] and music [9]. QBSE uses semantic information to retrieve semantically meaningful audio from the database. In many cases, a semantic understanding of the audio signal enables retrieval of tracks that, while acoustically different, are semantically similar to the query. For example, given a query with a high pitched, electric guitar sound, a system based on acoustics alone might retrieve songs with other high-pitched, harmonic sounds like violins or a female vocalist. On the other hand, QBSE would retrieve acoustic guitars, distorted guitars or banjos.

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 song 'Heartbreak Hotel' by Elvis Presley might have high values in the "blues", "guitar" and "mournful" semantic dimensions, and low values for "electronica", "clarinet" and "jolly".

The semantic feature vector is computed using an annotation system that assigns a weight to 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 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 [4] 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, as is the case in systems based on comparison of audio features directly.

3 MIREX AUDIO MUSIC SIMILARITY

For the MIREX 2007 Audio Music Similarity competition, the UCSD Computer Audition Laboratory QBSE system has been packaged as a MATLAB function. The function reads a text file with a list of audio file names, extracts features from these files and annotates each file in 146 semantic dimensions including words that characterize the genre, instrumentation, vocals, emotion, rhythm and usages associated with the audio. By comparing the KL divergence between the semantic multinomials that represent the audio annotations, a distance matrix between songs is returned and a results file outputs the top 100 similar tracks for each track in the database as well as the distance from the query to the similar track.

4 REFERENCES

[1] L. Barrington, A. Chan, D. Turnbull, and G. Lanckriet. Audio information retrieval using semantic similarity. *ICASSP*, 2007.

[2] C. R. Buchanan. Semantic-based audio recognition and retrieval. Master's thesis, School of Informatics, University of Edinburgh, 2005.

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

[4] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[5] 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*, 39:1 – 38, 1977.

[6] N. Rasiwasia, N. Vasconcelos, and P.J. Moreno. Query by semantic example. *ICIVR*, 2006.

[7] M. Slaney. Mixtures of probability experts for audio retrieval and indexing. *IEEE Multimedia and Expo*, 2002.

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

[9] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Towards musical query-by-semantic-description using the cal500 data set. *SIGIR*, 2007.

[10] N. Vasconcelos. Image indexing with mixture hierarchies. *IEEE CVPR*, pages 3–10, 2001.

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