Music Genre Classification Using Explicit Semantic Analysis

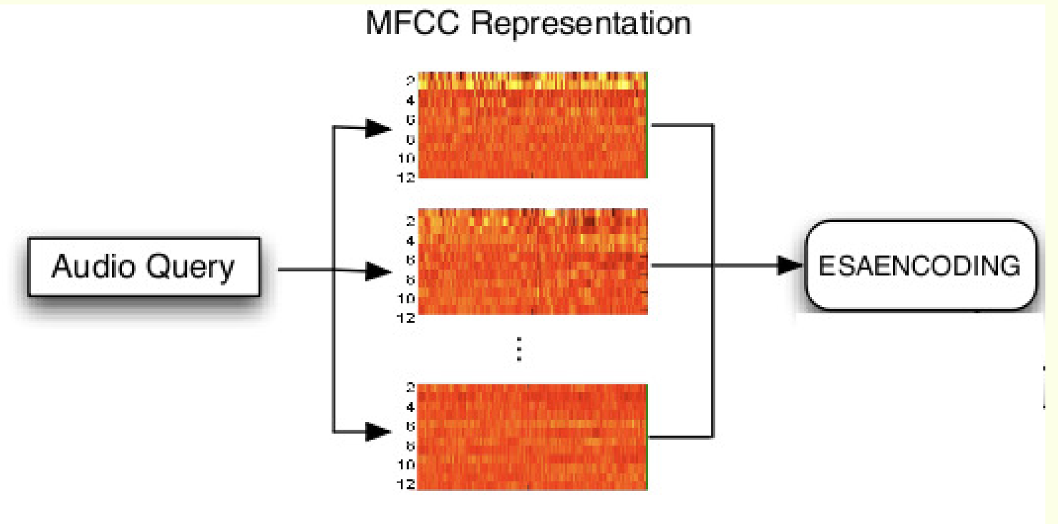
Kamelia Aryafar Ali Shokoufandeh

Department of Computer Science, Drexel University, Philadelphia, PA, USA

Motivation

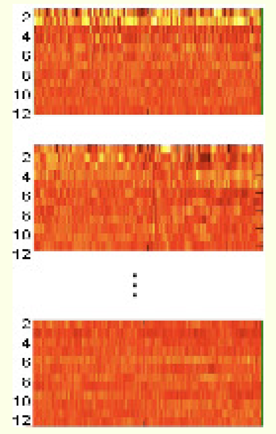
- We are interested in automatically finding **genre labels** for a music data set.
- The collection of user protocols is common in human factors research, but analyzing the large data sets produced can be tedious.
- Recent work has proposed the use of an automated method based on **explicit semantic analysis** to identify the most representative genre patterns in a large data set.
- The method only uses signal-based mel frequency cepstral coefficients (MFCCs) as audio feature representation and improves upon previous methods that use the same set of features for music genre classification.

Mel frequency cepstral coefficients (MFCCs)



The approximation algorithm of Denton, et al. (2008) expresses the problem in unoriented graphs as an optimization problem which simultaneously minimizes the weights of the **intra edges**, maximizes the weights of the **cut edges**, and maximizes the weights of the vertices in the canonical set.

Canonical Sets in Oriented Topologies



In oriented topologies, we note that there are two types of cut edges — those that lead into the canonical set (**cut-in edges**) and those that lead out of the canonical set (**cut-out edges**). We maximize the weights of the cut-in edges while minimizing the weights of the cut-out edges. The constraints on the intra edges and stability remain the same as in the unoriented formulation.

Definition of Terms

- set of features $\{f_1, ..., f_\ell\}$
- code-book of features $\mathcal{D} = \{\delta_1, ..., \delta_k\}$
- set of audio signals in the data set $C = \langle (f_1, w_1), ..., (f_\ell, w_\ell) \rangle$ where $\mathcal{C} = \{C_1, ..., C_M\}$
- $M \times k$ ESA matrix $\mathcal{E}_{\mathcal{C},\mathcal{D}}$, where
  - ▷ $tf(C, x) = \frac{\sum_{i=1}^{\ell} w_i \times d(f_i, x)}{\sum_{i=1}^{\ell} w_i}$ is the term frequency.
  - ▷ $idf_\delta = \log \frac{M}{\sum_{i=1}^{M} \chi(\delta, C_i)}$ is the inverse document frequency.
  - ▷ $tfidf(C, \delta) = tf(C, \delta) \times idf_\delta$.
  - ▷ $\mathcal{E}_{\mathcal{C},\mathcal{D}}[i, j] = tfidf(C_i, \delta_j)$.
- set of t ordered pairs $\mathcal{T} = \{(A_1, L_1), ..., (A_t, L_t)\}$ of audio sequences, A_i, and their corresponding genre labels, L_i

ESA encoding

Without loss of generality, assume that a set of t ordered pairs $\mathcal{T} = \{(A_1, L_1), ..., (A_t, L_t)\}$ of audio sequences, A_i, and their corresponding genre labels, L_i, are provided as the training data. Our hypothesis is that one can build a classifier with reasonable performance simply based on the ESA vectors and their labels in $\mathcal{T}$ for genre detection.

Algorithm: ESAENCODING$(A, \mathcal{D}, \mathcal{E})$
INPUT: A: INPUT AUDIO, $\mathcal{D}$: CODE-BOOK, $\mathcal{E}$: ESA MATRIX
RESULT: $\mathcal{E}(A)$: ESA-REPRESENTATION OF A
$\{f_1, ..., f_\ell\} \leftarrow \mathbf{MFCC}(A)$;
$\mathcal{E}(A) \leftarrow \bar{\mathbf{0}}$;
FOREACH $f \in \{f_1, ..., f_\ell\}$ **DO**
  $\delta^* = \max_{\delta \in \mathcal{D}} d(f, \delta)$;
  $\mathcal{E}(A) = \mathcal{E}(A) + \mathcal{E}(\delta^*)$;
END
RETURN $\mathcal{E}(A)$

ALGORITHM 1: CONSTRUCTION OF THE ESA VECTOR OF AN AUDIO SEQUENCE.

To this end, we form the set $\mathcal{E}(\mathcal{T}) = \{(\mathcal{E}(A_1), L_1), ..., (\mathcal{E}(A_t), L_t)\}$, where $\mathcal{E}(A_i)$is the ESA encoding of $(A_i, \mathcal{D}, \mathcal{E}),\ i = 1, ..., t$. The set $\mathcal{E}(\mathcal{T})$ of (ESA-vector, label) pairs will be provided as the training data to a supervised classifier algorithm such as a support vector machine (SVM) , building a model that assigns samples (points in the SVM feature space) to their genre categories.

Experiment: Music genre classification

In our experiments, we use the publicly available benchmark dataset for audio classification and clustering proposed by Homburg et al.

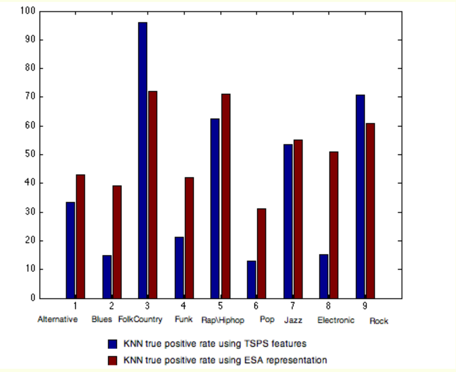
Genre	Number of Samples
alternative	145
blues	120
electronic	113
folk-country	222
funk soul/R&B	47
jazz	319
pop	116
rap/hip-hop	300
rock	504

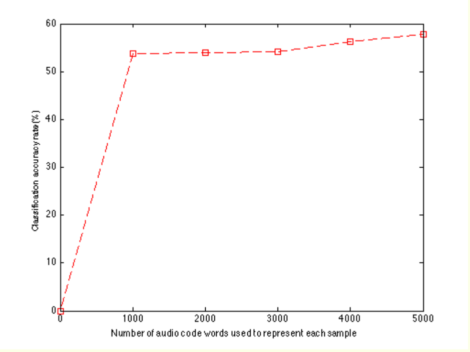
Table: Number of songs per genre.

- 10-fold cross validation to evaluate the accuracy of the classification
- *classification accuracy* as our performance measure
- similarity measure: cosine distance

Classifier	AM	TSPS	ESA	
			$k = 1000$	$k = 5000$
Random	22.39	21.68	29.51	25.40
k-NN	35.83	47.40	48.59	51.88
SVM	40.81	51.81	53.76	57.81

Table: Genre classification accuracy on the benchmark data set





Future Work

- incorporate other audio features
  - ▷ chroma features
  - ▷ textual metadata
- music artist identification in specific genre
- lyrics retrieval using an extended ESA model

Acknowledgments

This work was funded in part by Office of Naval Research (ONR) grant N00014-04-1-0363 and United States National Science Foundation grant N0803670.