SONG-LEVEL FEATURES AND SVMS FOR MUSIC CLASSIFICATION

Michael Mandel and Dan Ellis
LabROSA, Dept. of Elec. Eng., Columbia University, NY NY USA
{mim,dpwe}@ee.columbia.edu

ABSTRACT

Searching and organizing growing digital music collections requires automatic classification of music. Our system for artist and genre identification uses support vector machines to classify songs based on features calculated over their entire lengths. Since support vector machines are exemplar-based classifiers, training on and classifying entire songs instead of short-time features makes intuitive sense. We model songs as single Gaussians of MFCCs and use a KL divergence-based kernel to measure the distance between songs. This system placed first in both the audio genre and artist identification competitions at MIREX with classification accuracies of 72.45% and 78.81%, respectively.

1 MIREX Submission

All of our features are based on mel-frequency cepstral coefficients (MFCCs). MFCCs are a short-time spectral decomposition of an audio signal that conveys the general frequency characteristics important to human hearing. While originally developed to decouple vocal excitation from vocal tract shape for automatic speech recognition (Oppenheim, 1969), they have found applications in other auditory domains including music retrieval (Logan, 2000; Foote, 1997). At the recommendation of Aucouturier and Pachet (2004), we used 20-coefficient MFCCs.

Our features are most accurately described as *timbral* because they only model music's spectral characteristics on timescales of tens of milliseconds. We make the strong assumption that songs with the same MFCC frames in a different order are identical. Some authors call this type of modeling a "bag of frames", after the "bag of words" models used in text retrieval, which are based on the idea that each word is an independent, identically distributed (IID) sample from a bag containing many words in varying proportions.

Once we have extracted the MFCCs for a particular song, we describe that song as the Gaussian distribution with the maximum likelihood of fitting those MFCC frames. This Gaussian's mean and covariance are the population mean and covariance of the MFCC frames, a 20-dimensional vector and a 20×20 matrix, respectively.

To compare two songs, we measure the Kullback Leibler (KL) divergence between the two songs' Gaussians, with lower divergence indicating higher similarity. The KL divergence between two probability distributions, $p(x)$ and $q(x)$ is defined as

$$KL(p \,||\, q) \equiv \int p(x) \log \frac{p(x)}{q(x)} \, dx. \tag{1}$$

It should be noted that this is a divergence and not a distance because $KL(p \,||\, q) \geq 0$, with $KL(p \,||\, q) = 0 \iff p = q$, but $KL(p \,||\, q) \neq KL(q \,||\, p)$.

SVMs, however, require symmetric measures of similarity, so the KL divergence cannot be used directly and must be transformed appropriately. For d-dimensional Gaussians, $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu_p, \Sigma_p)$ and $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu_q, \Sigma_q)$, there is a closed form for the symmetrized KL divergence (Penny, 2001),

$$\begin{aligned} 2KL_s(p \,||\, q) = {} & Tr(\Sigma_q^{-1}\Sigma_p + \Sigma_p^{-1}\Sigma_q) - 2d \\ & + (\mu_p - \mu_q)^T(\Sigma_q^{-1} + \Sigma_p^{-1})(\mu_p - \mu_q). \end{aligned} \tag{2}$$

The symmetrized KL divergence is still a distance measure, not a similarity measure, i.e. the distance matrix is not positive semidefinite. We must convert distances to similarities in order to satisfy the Mercer conditions, which we accomplish by exponentiating the elements of this matrix. The final gram matrix has elements

$$K(X_i, X_j) = e^{-\gamma KL_s(X_i \,||\, X_j)}, \tag{3}$$

where γ is a positive constant that can be tuned to maximize classification accuracy. For this submission, $\gamma = 0.068$ was used, as it was found to yield the highest artist classification accuracy on a validation set of 100 songs taken from the *uspop2002* dataset.

Since the problems of artist and genre classification assign every song to exactly one class, we perform multiclass learning and classification with a DAGSVM (Platt et al., 2000). While n-class learning with a DAGSVM requires $O(n^2)$ binary SVMs like other multiclass SVM approaches, training and classification take only $O(n)$ time.

2 Comparison to Other Submissions

Logan's submission uses features most similar to those used in our system. She uses probabilistic models of MFCCs over the entire length of each song, specifically Gaussian mixture models, and then compares songs to one another using the earth mover's distance approximation to

Table 1: Classification accuracy for MIREX '05 competitions. Score for magnatune genre is after partial credit assignment for confusion with similar genres.

	Dataset	Raw	Normalized
Artist	magnatune	76.60%	76.62%
Artist	*uspop2002*	68.30%	67.96%
Artist	overall	72.45%	72.29%
Genre	magnatune	71.96%	69.63%
Genre	*uspop2002*	85.65%	76.91%
Genre	overall	78.81%	73.27%

KL divergence. The use of song-level features and a KL divergence distance measure between probability distributions is very similar to our approach. Unlike our system, however, she uses a simple nearest neighbor classifier.

West and Li's submission uses a classifier most similar to our system's. While it is unclear from their proposal what sort of multiclass SVM they use for classification, the idea of wrapping an SVM around a kernel designed for music is quite similar to ours. Whereas we chose a probabilistic model-based kernel, they chose one more related to document retrieval.

Three other systems also use SVMs for classification, mostly on audio chunks a few seconds long. Their general paradigm is to extract some audio features and then, presumably, use a standard radial basis function kernel to compare feature vectors.

3 Results

This system was ranked first in the artist classification task and second in the genre classification task, with accuracies of 72.45% and 78.81%, respectively. See Table 1 for a breakdown of those scores by test database, contest, and scoring metric. The two databases both contain around 1500 songs, *uspop2002* consists of mostly well-known rock and pop artists, magnatune includes independent artists in a wider variety of genres and styles. The two metrics, raw and normalized, differ in how they combine scores from different categories. Raw accuracy simply divides the number of songs correctly identified by the total number of songs. Normalized accuracy is the mean of each category's accuracy.

3.1 Running time

The running time of this algorithm is approximately linear in the number of songs, as computations are dominated by the feature extraction step. KL divergence calculations are relatively inexpensive, but involve matrix inversions and multiplications that can add up when multiplied by the million or so pairs of songs to be compared. Using a DAG-SVM requires time only linear in the number of classes, although space requirements are quadratic. A contest machine (a 3GHz Pentium 4) took approximately 5 seconds per song, for a total running time of around 2.5 hours per database per competition.

References

Jean-Julien Aucouturier and Francois Pachet. Improving timbre similarity : How high's the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

Jonathan T. Foote. Content-based retrieval of music and audio. In C.-C. J. Kuo, Shih-Fu Chang, and Venkat N. Gudivada, editors, *Proc. SPIE Vol. 3229, p. 138-147, Multimedia Storage and Archiving Systems II*, pages 138–147, October 1997.

Beth Logan. Mel Frequency Cepstral Coefficients for music modelling. In *International Symposium on Music Information Retrieval*, 2000.

Alan V. Oppenheim. A speech analysis-synthesis system based on homomorphic filtering. *Journal of the Acostical Society of America*, 45:458–465, February 1969.

William D. Penny. Kullback-Liebler divergences of normal, gamma, Dirichlet and Wishart densities. Technical report, Wellcome Department of Cognitive Neurology, 2001.

John C. Platt, Nello Cristianini, and John Shawe-Taylor. Large margin DAGs for multiclass classification. In S.A. Solla, T.K. Leen, and K.-R. Mueller, editors, *Advances in Neural Information Processing Systems 12*, pages 547–553, 2000.