Exploration of Point-Distribution Models for Similarity-based Classification and Indexing of Polyphonic Music.

Abstract

Similarity is an intuitive criterion for indexing and classification of digital audio files in music information retrieval systems. While significant work has been done on similarity-based approaches to monophonic music, methods for reliably dealing with databases of arbitrary polyphonic music remain elusive. In this paper we describe our ongoing research in exploring the use of high-order multivariate statistical techniques for similarity-based classification of polyphonic music in digital audio files. The statistical techniques we employ, known as *point distribution models* (PDMs), have recently proven to be of surprising value in computer vision research for rating visual similarity; here we are attempting to apply PDMs to musical similarity. This involves creating neural networks that approximate the statistical processing, to save on potentially explosive storage and processor requirements. This paper reports on work in progress: our results to date are inconclusive and somewhat negative. We describe our rationale for exploring PDMs in polyphonic music similarity-rating and discuss the problems we have encountered so far, with the intention of encouraging other members of the music information retrieval community to explore this and related approaches.

Overview

The approach we are currently exploring is, in essence, to perform sophisticated statistical analysis on the results of applying standard signal processing techniques to digital audio polyphonic music data. The key difference between our approach and much prior research is that we are explicitly rejecting the need for an intermediate symbolic representation (such as audio-to-MIDI) of the content of digital-audio music files.

Rather, we view individual digital audio files as individual data-points in an ultra-high dimensional space (the space of all possible digital audio files for a given duration, sampling rate, bit depth, and number of channels). We are exploring the use of multivariate statistical analysis techniques to find transformations that map from this input space (which could have many millions of dimensions) onto new spaces with several orders of magnitude fewer dimensions, where the new spaces have similarity-like distance metrics and classification-like partitions.

This might seem like a hopelessly naïve approach, had not a similar approach recently been shown to be a successful foundation for similarity-based indexing in other ultra-high-dimensional spaces: namely those found in computer vision.

In the past decade, new statistical approaches have been developed for computer vision in medical imaging (Wolfson, 1999) and automated surveillance of pedestrians (Leeds, 1998). In essence, these new statistical methods involve computing a characterization of the mean shape, along with a characterization of the *primary modes of variation* in the shape (i.e., the main ways in which the samples of the shape differ from the mean of the samples). Surprisingly, these primary modes often correspond to intuitive visual notions such as variation in viewing angle of a particular object (such as a human face), or the variation seen between two classes of object (such between female faces and male faces).

The approach we have explored in music classification is, in essence, to apply these vision techniques to visual objects that are representations of portions of digital audio. The visual objects are created by taking the spectrograms (amplitude surfaces over a time-frequency space) of the mono sum from fixed-duration samples of stereo CD-quality digital-audio music files and then relatively-coarsely quantizing the time and frequency axes into discrete bins to give data-compression in excess of 90%. The quantized spectrogram is decorated with one amplitude "point" per time-frequency bin, and the collections of points for a particular music sample are treated as coordinates for a single point in a high-dimensional space (the space of possible quantized spectrograms). For example, working with 30-second audio samples and dividing the spectrogram frequency axis into 1000 bins and the time axis into 50 bins per second gives 1000*50*30=1.5m points, so each processed sample is a single point in a 1.5m-dimensional space of possible inputs.

We compute a collection of such points, one per sample, for samples from a variety of music files, to give a cloud of points in input-space. We then apply linear principal components analysis (PCA) on deviations from the mean of this cloud to identify principal modes of variation, and then explore whether these principal modes correspond to intuitive notions of similarity. We also explore whether the early principal components can be interpreted as a set of basis vectors for a a subspace in which simple distance metrics correspond to measures of musical similarity.

The dimensionality of the input space is so large that neural-network approximators to PCA (Sanger, 1989) are used rather than analytic matrix-manipulation code. So far, our results (from samples of commercial recordings) are inconclusive. We are attempting to better understand the nature of our approach by working with music generated from MIDI files where we have full experimental control over various aspects of the music such as tempo, number of voices, instrument type, and so on. We are also looking to use nonlinear PCA neural networks (e.g. Fotheringhame & Baddeley, 1997) to test the possibility that our current use of linear methods represents an over-simplification.

Author Information

Dave Cliff Digital Media Systems Department Hewlett-Packard Labs Bristol BS34 8QZ England U.K. dave_cliff@hp.com Phone +44 117 312 8189	Heppie Freeburn Digital Media Systems Department Hewlett-Packard Labs Bristol BS34 8QZ England U.K. cf@hplb.hpl.hp.com Phone +44 117 312 8718

Suggested Readings

Fotheringhame, D. and Baddeley, R. 1997. "Nonlinear Principal components analysis of neuronal spike train data". *Biological Cybernetics* 77: 282-288.

Leeds, 1998. University of Leeds, School of Computer Studies, Computer Vision Group. http://www.scs.leeds.ac.uk/imv/

Sanger, T.D. 1989. "Optimal unsupervised learning in a single-layer linear feedforward neural network". *Neural Networks*, 2:459-473.

Wolfson 1999. Wolfson Medical Imaging Unit, University of Manchester. http://www.wiau.man.ac.uk/research/Flexible_Models/index.html