Clustering of Musical Sounds using Polyspectral Distance Measures.

Shlomo Dubnov and Naftali Tishby
Institute for Computer Science and
Center for Neural Computation
Hebrew University, Jerusalem 91904, Israel

ABSTRACT: This paper describes a hierarchical clustering of musical signals based on information derived from spectral and bispectral acoustic distortion measures. This clustering reveals the ultra metric structure that exists in the set of sounds, with a clear interpretation of the distances between the sounds as the statistical divergence between the sound models. Spectral, bispectral and combined clustering results are presented.

1 Introduction

The ability to assess similarity between musical timbres is basic both for the understanding of our musical perception and for compositional practice. Originally addressed by psychoacousticians, perceptually relevant timbre spaces were devised based on human listening experiments. These results, (Gray) reported in a series of classic papers, had become the standard representation for the musical instruments timbre space. We propose to probe into this space by mathematical tools that could measure this similarity directly from the musical signal.

Acoustical distortions measure, although well known in the speech literature (Gray Markel) received little attention in music applications. One reason for this is that the underlying linear model used for these functions cannot capture the wealth of sound found in musical signals. Here we discuss distance measures for acoustic signals based on spectral and higher order statistical information (Nikias Raghavachar) drawn from the steady state portion of the signal. Special attention is given to musical signals whose timbral properties are difficult to model.

The following issues will be addressed in this paper: (1) the notion of higher order spectra, (2) acoustic distance measure and its generalization to bispectral information, (3) a method for clustering stochastic models and finally (4) application to classification of musical instruments.

2 Higher Order Statistics

In order to understand our motivation for studying higher order correlations, we will recapitulate briefly some of the reasons for using the ordinary double correlation. The common assumption is that out ears perform spectral analysis of the incoming signal. Naturally, not all of the signal information is retained in our ears and the simplest assumption is that the phase is ignored. It is well known that the amplitude of the Fourier spectrum is the Fourier transform of the signal's autocorrelation. This autocorrelation in the time domain is the basic information extracted from the acoustic signal by our auditory system. This information represent the signal's spectral envelope in the frequency domain. Now we intend to extend the scope of the acoustic analysis by including triple and may be
higher correlations of the signal, also known as polyspectra in the frequency domain. The 4th-order correlation, \( h_4(t_1, ..., t_4) \), of a signal \( \{h(t)\}_{t=0}^{N} \) is defined as,

\[
h_4(t_1, ..., t_4) = \sum_{\tau = 0}^{N} h(t+\tau) h(t+\tau+1) h(t+\tau+2) h(t+\tau+3) \tag{1}
\]

In frequency domain it corresponds to the 4th-order spectrum,

\[
B_4(\omega_1, ..., \omega_4) = \sum_{i_1, i_2, i_3, i_4 = -N}^{N} h_4(i_1, ..., i_4) e^{-j2\pi(i_1\omega_1 + i_2\omega_2 + i_3\omega_3 + i_4\omega_4)} = H(\omega_1) \cdot H(\omega_2) \cdot H(\omega_3) \cdot H(\omega_4) \tag{2}
\]

3 Acoustic Distortion Measures

Acoustic distortion measures are widely used in speech literature as means of signal comparison, and are closely related to feature representation of the signal. One of the most common representations of speech signals is the linear predictive coding (LPC), which provides reasonable representations for speech analysis and synthesis. We believe, however, that this model is insufficient for the acoustic representation of musical signals and suggest a simple extension to it, which still avoids the numerically expensive schemes encountered in other common musical signal representations (such as phase vocoding for instance).

The LPC representation can be considered as an acoustical model that treats the signal as a random Gaussian excitation passing through a linear resonant chamber. A Gaussian random excitation is completely characterized by its first two moments and thus contains no additional information in its higher order spectra. In a previous work (Dubnov et al.) we have suggested an acoustic model that treats the signal as a random non-Gaussian excitation passing through a linear resonant chamber. Using this model, an acoustic distortion measure based on order statistics of the signal has been derived. It was shown also that this measure is a generalization of the Itakura-Saito distortion measure to include bispectral information. Thus, we propose a model of a linear filter driven by a non-Gaussian white noise, which may better capture aspects of musical sound related to timbre or "sound color" of the musical instrument.

By having a well defined statistical model of the signal the problem of measuring distances between two signals is turned into the question of evaluating the statistical similarity between the model distributions. The statistical similarity is measured by the Kullback-Liebler (KL) divergence between the distributions, and provides us with a distortion function

\[
D(P||P') = -N(\lambda_0 - \lambda_0') + (\lambda_1 \mu - \lambda_2 \mu'_{\omega_2}) \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \ln \left( \frac{S_0(\omega)}{S_0'(\omega)} \right) + (\lambda_3 \mu - \lambda_4 \mu'_{\omega_4}) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{d\omega_2 d\omega_4}{(2\pi)^2} B_4(\omega_1, \omega_2) B_4(\omega_3, \omega_4) \tag{3}
\]

where \( S_0(\omega) \) and \( B_4(\omega_1, \omega_2) \) stand for the spectrum and bispectrum of the signal, and \( (\mu, \lambda) \)'s are derived from the moments of the non-Gaussian input noise.

4 The clustering method

An interesting application of our new distortion measure is an acoustic taxonomy of various musical instrument sounds. Musical sounds are hard to cluster since it is difficult to model sounds

\[\text{The Itakura-Saito distortion is a statistical, Kullback-Liebler (KL) divergence, between signals represented by their LPC models.}\]
as points in a some low dimensional space. Though we limit ourselves to the stationary portion of the signal, a grouping of musical instruments, by the above similarity measure, appears in a rather interesting manner, as shown below.

In our approach, acoustic signals are treated as samples drawn from a stochastic source, namely, music signals are represented by their corresponding statistical models. Each model is chosen to be the most likely to have produced the observed waveform, in its parameter class. It is important to notice that our relevant observable is the spectrum and the bispectrum of the signal, but in order to evaluate all the parameters of the distortion function, the \( \{ \mu, \lambda \} \) must be estimated as well. Although this can be done in principle and an iterative solution procedure for that was described in another work (Rubnov, Tishby) here we prefer to avoid the complete estimation of these parameters by using the following arguments.

The distortion function is characterized by two integrals over ratios of the spectra and bispectra of the two signals involved. These are actually the only two quantities that are directly derived from the observed signal. We expect these two observables to contain all the relevant acoustic information and thus we substitute the original distortion function, Eq. 4, by the pair \((SR_{ab}, BR_{ab})\) where

\[
SR_{ab} = \int \frac{d\mu}{\mu} \frac{d\lambda}{\lambda} \int \frac{d\gamma}{\gamma} \frac{d\delta}{\delta}; \quad BR_{ab} = \int \frac{d\nu}{\nu} \frac{d\zeta}{\zeta} \int \frac{d\gamma}{\gamma} \frac{d\delta}{\delta}
\]

are the spectral and bispectral integral ratio expressions, respectively.

The idea is to represent a sound \( y \) by a collection of all pairs \((SR_{ab}, BR_{ab})\), calculated with respect to all signals \( y' \) in our data set. This “trace” vector provides a signature of the particular signal \( y \), characterizing it by the values of spectral and bispectral integral ratios, with respect to all the other signals. These signatures are expected to be similar for sounds with similar characteristics and serve as the basis for an acoustic classification, which in turn determines the underlying structure in the musical domain. By this transformation we have turned our data into a collection of points in a high dimensional space, the dimensionality being equal to the length of our signature vectors.

Using this vector-signature representation we explore our signal collection using a deterministic annealing hierarchical clustering approach (Rose et al.). As in other fuzzy clustering methods, each point (a signature vector in our case) is associated with probability with all the clusters. The “annealed” system consists of a set of probability distributions for associating points with clusters. The associations probability is controlled by a inverse temperature-like parameter, \( \beta \). As \( \beta \) gets larger, the temperature is lowered and the associations becomes less fuzzy. At the starting point \( \beta = 0 \) and each point is equally associated with all clusters. As \( \beta \) is raised, the cluster centroids bifurcate through a sequence of phase transitions and we obtain an ultrametric structure of associations, controlled by the temperature parameter.

5 Clustering Results

As described above, an hierarchical clustering was performed iteratively by splitting the models (cluster centers) while raising \( \beta \). This creates a sequence of partitions that give refined models at each stage, with clusters containing a smaller group of signature vectors, which are the representatives for the signals themselves.

In order to test in detail the spectral and bispectral components of our data, we performed three separate clusterings: (1) a clustering with the spectral integral ratios vector only, (2) a bispectral integral ratios vector only, and (3) a combined clustering with vectors consisting of the complete \((SR_{ab}, BR_{ab})\) set of pairs.

Our data set was taken from a collection of 31 sampled instruments sounds taken from the Portland sample player module. The results of spectral, bispectral and combined clustering are

---

1 The signature vector could be calculated with respect to a small representative subset of our signal set. We choose to evaluate the complete distances matrix, i.e. each signal is represented by the distances from all other signals, and the space dimension is equal thus to the size of the data set.
shown in figures 1, 2, 3 respectively. Regarding the spectral and bispectral clusterings, we first note that many similar pairs of sounds, such as Flutes, Clarinets, Trumpets are grouped together in both clustering methods. It is interesting to notice also that the Violin and Viola are related (to different extents) in both methods, while the Cello is distant from both of them in the two clustering trees. Notice also that the Bassoon is close to Contra Bassoon in the spectral tree, while being very distant in the bispectral case. Similarly, but in the other direction, the Alto Oboe, English Horn pair are close in bispectral tree and distant in spectral tree.

Qualitative analysis of these results suggests some possible interpretations to the behavior of spectral and bispectral clusterings. A rather "trivial" observation is that the spectral tree "grows" in direction of spectral richness - the Flutes are close to the root while the deeper branches belong to spectrally richer string and wind instruments. The bispectral tree exhibits clustering that puts at the deep nodes instruments which are normally identified as "solo" instruments, while the sounds closer to the root belong to a lower register instruments of a more "supportive" character. It may be said that the hierarchical clustering splitting corresponds to some "penetrating" quality of the sounds, with Flutes and Violin at the extreme.

The combined spectral/bispectral results are difficult to interpret, although traces of the spectral and bispectral trees are apparent there.

Acknowledgments

We would like to thank Prof. Dalia Cohen for fruitful musical discussions, and Mr. Yona Golan for discussions of clustering and for providing us with the routines.

Bibliography


Figure 1: Spectral clustering tree. The numbers on the nodes are the splitting values of $\beta$.
Figure 2: Biepectral clustering tree.
Figure 3: A combined spectral and bispectral clustering tree.