

# MUSICAL INSTRUMENT FAMILY CLASSIFICATION

Ricardo A. Garcia

Media Lab, Massachusetts Institute of Technology  
20 Ames Street Room E15-401, Cambridge, MA 02139 USA  
PH: 617-253-0112 FAX: 617-258-6264  
e-mail: rago @ media . mit . edu

*A method to classify sounds of musical instruments (single monophonic notes) is introduced. The classification is done using in parallel two sets of perceptual features extracted from the sounds. The models used are a mixture of gaussians, whose parameters were found by training over a database of target sound families. The feature extraction procedure, model training and model usage for classification are explained. An implementation and results of the method are shown and discussed.*

## 1 Introduction

Humans are very good at classifying types of musical instruments, even under the most adverse conditions (i.e. noise, polyphonic sound, environmental perturbations). But training a computer to recognize between two different instruments of the same “family” is not an easy task. The problem is that the concept of a family of sounds is difficult to explicitly teach to a computer. A method that uses a set of perceptually derived features to train models of families of sounds is introduced. The sound is analyzed and mapped into two perceptual feature spaces: the Spectral Contours and the Cepstral Coefficients.

## 2 Approach

The proposed method uses a parallel feature-space modeled with gaussian-mixtures to approximate the families of musical instruments to be classified. The models are trained using methods of estimation-maximization. Two feature spaces are used: spectral contour and cepstral coefficients. These are calculated in a frame-by-frame basis for each musical note in the training set. Normalization and dimensionality reduction are applied to make the final training set for each musical instrument type. An independent gaussian mixture model is trained for each feature space for each instrument family. These models are used in parallel to compute the probability of each unknown note of being classified as belonging to a particular instrument type in every feature-space.

### 2.1 Feature-space

Digital audio signals are associated with high sampling rates. The amount of data that is produced per second is surprisingly high. Luckily, the characteristics of the sounds of musical instruments vary slowly in time, allowing a meaningful (and very reduced) set of data to be extracted from the original audio signal. In addition, it is important to take into account in some degree the human perceptual element when classifying musical instruments [ 4 ]. Features computed in a frame-by-frame basis of about 20 ms per frame and 30 percent overlap have shown to be good for analysis of musical sounds [ 5 ].

#### 2.1.1 Spectral contour

Each frame is transformed using a DFT to the discrete frequency domain. The basic spectral contour is defined as the smoothed energy spectrum. In this project, a modified spectral contour is applied. This modification is performed using a non-linear mapping of the frequency axis into the bark scale [ 2 ]. The

bark scale is a perceptually derived scale composed by non-regular divisions in the frequency domain, each division is equivalent to one “critical band”. This mapping is done using [ 2 ].

$$z = 13 \tan^{-1}\left(\frac{0.76 * f}{1000}\right) + 3.5 \tan^{-1}\left(\left(\frac{f}{7500}\right)^2\right) \quad (1)$$

The spectral contour is calculated for every frame, in all critical bands. In Figure 1 is possible to see the evolution in time vs. the critical bands of different musical instruments. (note that it is plotted the log of the magnitude to have a better understanding).

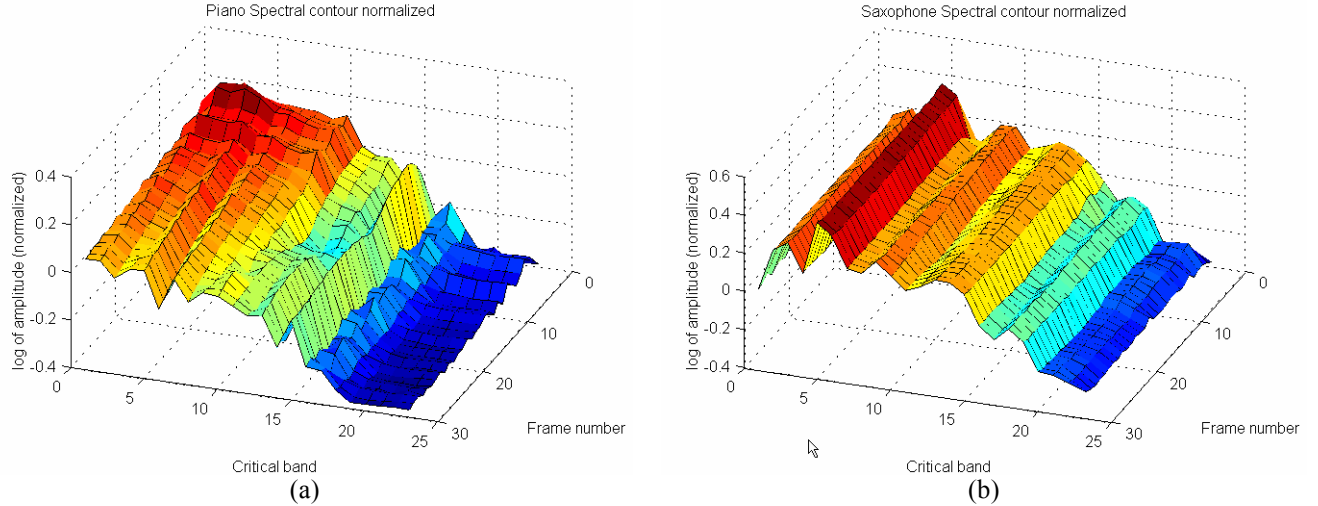


Figure 1 Spectral contour of (a) piano and (b) saxophone

### 2.1.2 Cepstral coefficients

A technique originally developed to analyze speech can be applied to extract some information about the structure of the sound [ 7 ], [ 5 ]. The “real cepstrum” (or simply cepstrum) is computed as the inverse Discrete Fourier Transform of the logarithm of the magnitude of the energy spectrum of the frame. This is expressed as:

$$cepstrum = IFFT(\log(|FFT(frame(t))|)) \quad (2)$$

For each frame, a set of the first cepstral coefficients is selected. These first coefficients are related to the distribution of the broadband energy in the frequency domain. If a sound is regarded as a time-invariant linear system that is excited by a quasi-periodic source, the first coefficients of the cepstrum will be related to the structure of the system, regardless of the frequency of the excitation [ 7 ]. This is very convenient for the task at hand, because the goal is to classify sounds regardless of their pitch. Also, the underlying assumption is that each family of instruments “preserves” some of the characteristics (cepstral coefficients) across the pitch range. Figure 2 shows the cepstral coefficients evolution for the same instruments in the previous figure.

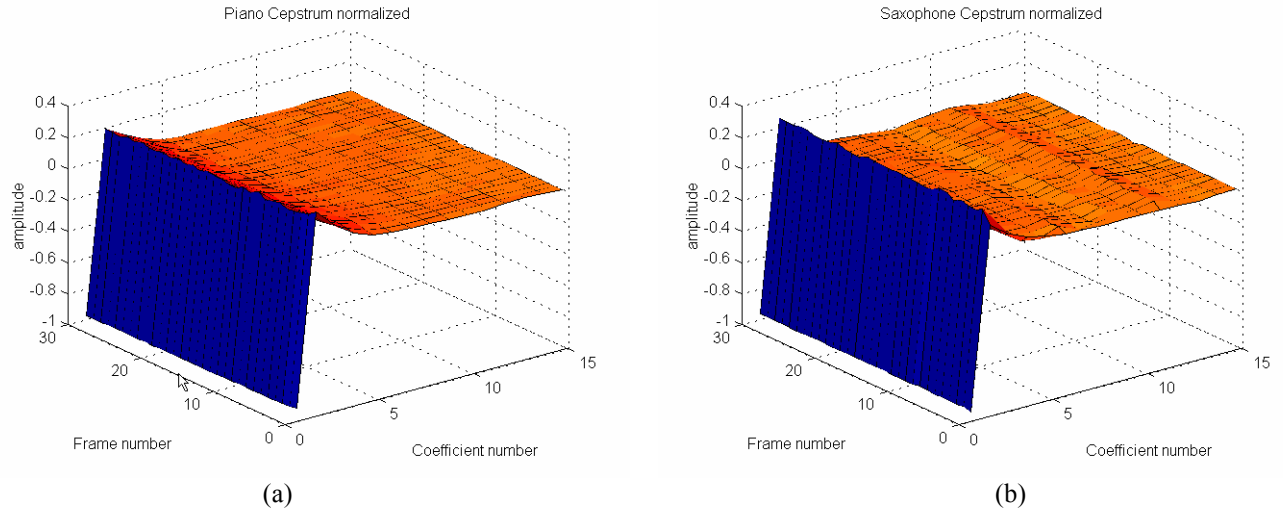


Figure 2 Cepstrum coefficients for (a) piano and (b) saxophone

## 2.2 Feature data manipulation

Because of the high number of dimensions that result typically in the feature extraction, it is necessary to realize a dimensionality reduction to be able to manage the data in a more efficient way. Normalization helps to isolate the effects of different energy levels in different frames and sound samples.

### 2.2.1 Normalization

Many of the feature values are related to the energy of the spectrum (or the log of the energy), and at the same time it depends on the energy of the original sound (or sound level). This can drive similar sounds with different sound level to be classified as from different families. The approach taken is to make the magnitude of each frame to be equal to unity.

$$\overline{frame} = \frac{frame}{\|frame\|} \quad (3)$$

Note that this will make some frames that fade out in time to loose this time dependant characteristic.

### 2.2.2 Dimensionality reduction

It is convenient to reduce the number of components to manipulate in each frame, because this reduces the time of computation of the algorithms, as well as it improves the performance of the selection when removing redundant (or noisy) information from within the data sets. The method used is Principal Component Analysis (PCA) [ 3 ].

$$y = \mathbf{W}x \quad (4)$$

## 2.3 Model of a musical instrument family: gaussian mixture

Each musical instrument family is modeled by two independent gaussian mixtures [ 1 ],[ 3 ],[ 6 ], one for each feature space used. Therefore, the representation of a single instrument is given by the linear combination of these gaussian mixtures:

$$p(x|type) = A \sum_{j=1}^{M_{contour}} P_{j_{contour}}(j) p_{contour}(x|j) + B \sum_{j=1}^{M_{cepstrum}} P_{j_{cepstrum}}(j) p_{cepstrum}(x|j) \quad (5)$$

Where each  $p_{contour/cepstrum}(x|j)$  is a multivariate, full covariance gaussian distribution, with dimension equal to the dimension of each feature space.

The weights A and B reflect some knowledge about the confidence in the measure of each feature space. ( $A+B=1$ ,  $A, B \geq 0$ );

### 2.3.1 Training: Expectation – maximization

An iterative procedure to estimate an optimal set of parameters for the models is used, namely Expectation – Maximization EM [1],[3], The algorithm selects a random initial guess for the parameters to optimize  $(p_j(j), u_j, \sigma_j^2)$  and computes the expected probability of each training point with relation to the mixture. Then, a new set of parameters is estimated and the procedure repeated until a good solution is found, or a limit in the number of iterations is reached.

### 2.3.2 Classification procedure

The classification of a new sound sample has the following steps:

- Feature extraction: Extract the frames, and the Spectral Envelope and Cepstrum coefficients.
- Normalization: each frame is normalized to unity magnitude.
- Dimensionality reduction: for each feature set and each type of musical instrument, a different transform matrix is used to reduce the number of dimensions.
- Calculate  $p(x|type)$  for all the types of instruments, all the frames of the sound, all the feature spaces.
- Select the highest  $p(x|type)$  for each frame.
- Classify the sound as belonging to the type that appears the most in all the classified frames/feature spaces.

## 3 Implementation

The suggested algorithm was implemented and tested. The used resources and the results are shown.

### 3.1 Program and required resources

Sound Database:

McGill University Master Samples (MUMS library).

CD 1: Solo Strings: solo violin, viola and cello. 321 notes

CD 2: Woodwind and brass: Flutes, clarinets, bassoons, trumpets, and trombones. 454 notes

CD 3: Piano: Grand pianos. 264 notes

Each set was randomly divided in 80% training and 20% testing data.

The programs were written in Matlab 6.0 and run in a Pentium III dual proc. 750 Mhz, 256 MB ram.

They took in average 800 seconds to find the parameters for a model using 10 dimensions and 7 gaussians, using all the training data set.

Because of the dependence in a good selection of initial conditions for a good estimate of the parameters using EM, each model was required to be estimated an average of 3 times, to avoid some problems with numeric precision (when a matrix of covariance “shrinks” to be zero).

## 4 Discussion

During the realization of this project, some interesting functional aspects of EM and gaussian mixtures were noted:

When having many dimensions (more than 8), and many gaussians (more than 6), it was common that the initial conditions of the parameter estimation seemed to make the system more sensitive to them. Many of the runs where stopped because of “numerical” problems with the evaluation after a bad choice of initial conditions.

A couple of restrictions have to be enforced, to not allow a gaussian to “shrink” in any given dimension too much. This will make it “flat” in this dimension and therefore worthless.

The number of dimensions needed to express both feature spaces was around 8 for the Spectral Contour, and about 5 for the Cepstrum Coefficients. The number of gaussians was about 6 and 3 respectively.

## 5 References

- [ 1 ] Duda, R. O., Hart, P. E., *Pattern Classification and Scene Analysis*, John Wiley & sons, New York, 1973
- [ 2 ] Garcia, R. A., *Digital Watermarking of Audio Signals using a Psychoacoustic Auditory Model and Spread Spectrum Theory*, University of Miami, Master thesis, 1999
- [ 3 ] Gershenfeld, N., *The Nature of Mathematical Modeling*, Cambridge University Press, New York, 1999
- [ 4 ] Martin, K. D., *Sound Source Recognition: A Theory and Computational Model*, MIT Ph.D Dissertation, 1999
- [ 5 ] Oppenheim, A. V., Schafer, R. W., Buck, J. R., *Discrete-Time Signal Processing*, Prentice hall, New Jersey, Second Edition, 1999
- [ 6 ] Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, Mc Graw Hill, Third Ed., 1991
- [ 7 ] Rabiner, L. R., Schafer, R. W., *Digital Processing of Speech Signals*, Prentice Hall, New Jersey, 1978