Computer Identification of Wind Instruments using Cepstral Coefficients
Judith C. Brown

Physics Department, Wellesley College, Wellesley, Massachusetts, 01281 and Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139

ABSTRACT Cepstral coefficients based on a constant Q transform have been calculated for 28 short segments of oboe sounds and 52 saxophone sounds. These were used as features in a pattern analysis to determine for each of these sounds comprising the test set whether it belongs to the oboe or to the sax class. The training set consisted of longer sound segments of approximately 1 minute for each of the instruments. A k-means algorithm was used to calculate clusters for the training data, and Gaussian probability density functions were formed from the mean and variance of each of the clusters. Each member of the test set was then analyzed to determine the probability that it belonged to each of the two classes; and a Bayes decision rule was invoked to assign it to one of the classes. Results have been extremely promising.

INTRODUCTION

One of the most successful methods of automatic speaker identification (Reynolds and Rose, 1995 and references therein) has involved the use of a Gaussian mixture model with cepstral coefficients as features. Since a musical instrument can be viewed as a source plus resonator similar to a human speaker, it is of interest to determine whether this method works equally well for musical instrument identification. Thus a similar procedure has been followed using pattern recognition and other techniques developed by the speech community to distinguish between sounds made by oboes and saxophones, as examples of wind instruments.

CALCULATIONS USING PATTERN RECOGNITION

In any pattern recognition problem the most important step is the choice of features. Here we model a musical instrument as a resonator with a periodic excitation, and then analyze the sound produced in the same manner followed in the speaker identification work previously mentioned. For speech the glottal impulses are treated as a periodic excitation followed by a filter, which is the vocal tract or resonator. The musical analog for wind instruments is the pressure controlled opening and closing of the reed(s) delivering puffs of air into the cylindrical or conical bore resonator. One set of features which has been particularly successful in characterizing the vocal tract resonances which identify individual speakers are the cepstral coefficients.

In this implementation a constant Q transform was calculated using a method described previously using Hamming windowed kernels. (Brown and Puckette, 1992) The transformation from constant Q to cepstral coefficients was carried out using Equation 10.1 from O'Shaughnessy (1987).

\[ c[n] = \sum_{k=1}^{M} \log(X^{\text{cq}}[k_{cq}]) \cos \left( n\left(k - \frac{1}{2}\right) \frac{\pi}{M} \right) \]  

(1)

for \( n = 1, 2, ..., M \). Here \( X^{\text{cq}}[k_{cq}] \) is the \( k_{cq} \)th constant Q coefficient. In this implementation the constant Q coefficients are roughly equivalent to a third octave filterbank with 18 coefficients from 100 Hz to 5439 Hz.

Rather than comparing all points in an unknown to all points in the training set, Popat and Picard (1993) have used the method of clustering (Thierrien, 1989) to summarize the training data for each class. The method of clustering involves grouping the points from the known (or “training sound”) into so-called clusters. Thus each class can be characterized by a number of clusters and having determined the parameters for the clusters, one can model an arbitrary probability density for each class as a sum of these weighted Gaussians and calculate the probability that each point of the unknown sound belongs to that class.
All of the sound samples in this study were excerpted from the Wellesley College Music Library collection of compact disks, audio cassettes, and records or from the personal collection of a member of the Music Department faculty.

A K-means algorithm written by Kris Popat was used in all calculations to determine the cluster means and variances. The software allowed the number of clusters as an input, and the results could then be checked for the number of clusters which gave optimum performance.

In order to classify the unknown sounds into two classes A and B (oboe and sax in our case), calculations were made of the probability densities that the points defined by the feature vectors of an unknown U belonged to each of the classes. These were then compared to find the greater probability density.

It can be shown that this is equivalent to a log likelihood ratio:

$$\log(p(X|\Omega^1)) - \log(p(X|\Omega^2)) > 0$$

Here $X = \{x^1...x^N\}$ is defined as the set of all feature vectors measured for unknown sound U and $\Omega^1$ and $\Omega^2$ represent the two classes.

RESULTS

Results are shown in Table 1 where the machine calculations are compared with results on human perception carried out on a subset of the same sounds. The computer has a lower error rate than the humans for the oboe samples and one roughly the same for the sax samples. Since humans are considered the ultimate receivers, the success of the computer identification is impressive and hold promise for a number of areas such as audio indexing, data reduction, and automatic transcription.

TABLE 1 Summary of results on human and computer instrument identification. Each column gives the average fractional error for each experiment (average person errors divided by the number of sounds presented).

<table>
<thead>
<tr>
<th></th>
<th>Controlled Env</th>
<th>Auditorium</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oboe</td>
<td>4/27 = .15</td>
<td>2.7/16 = .17</td>
<td>0/28 = .00</td>
</tr>
<tr>
<td>Sax</td>
<td>2.5/31 = .08</td>
<td>2/17 = .12</td>
<td>5/52 = .10</td>
</tr>
</tbody>
</table>

ACKNOWLEDGEMENTS

I am very grateful to Kris Popat and Nicolas Saint-Arnaud of the MIT Media Lab for the many hours of discussions which made this work possible. Kris Popat was also very generous with his clustering software and additional time explaining how to use it.

REFERENCES


