Computer identification of musical instruments using pattern recognition with cepstral coefficients as features

Judith C. Brown[a)]
Physics Department, Wellesley College, Wellesley, Massachusetts 02181 and Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 4 August 1997; revised 19 October 1998; accepted 12 November 1998)

Cepstral coefficients based on a constant *Q* transform have been calculated for 28 short (1–2 s) oboe sounds and 52 short 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 sounds of 1 min or more 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 good and are compared to a human perception experiment identifying a subset of these same sounds.
© *1999 Acoustical Society of America.* [S0001-4966(99)00703-1]

PACS numbers: 43.75.Cd [WJS]

INTRODUCTION

The perception of timbre by humans has been widely studied over the past four decades with the elusive goal of correlating the results of perceptual experiments with a small number of acoustical properties of the sounds studied. Many of the studies have used perceived proximity ratings as a measure of similarity of sounds followed by analysis using multidimensional scaling (MDS) techniques, while others have modified the acoustic signal and then done perception experiments to determine whether the altered signal could be distinguished from the original. There have been a wide range of results identifying various spectral and temporal properties and assigning weights to their saliency.

As is often the case in research in musical acoustics, the first work was reported in Helmholtz' (1885/1954) monumental work. He postulated that timbre perception depended on the spectral shape resulting from real-time frequency analysis on the basilar membrane. This is called his ''harmonic structure'' theory.

Saldano and Corso (1964) did experiments to distinguish between Helmholtz' theory and the formant theory of musical quality according to which there is a strengthening of partials in certain regions due to resonances of the musical instrument. Their spectral results were equivocal, but they demonstrated the importance of onset patterns for timbre recognition. Experiments by Risset and Mathews (1969) pointed out the importance of the time-varying properties of the onset in trumpet sounds.

Clark *et al.* (1964) found timbre to be determined by the attack transient, by modulation during the steady state, and to some degree by one or more formants. In an interesting study to determine the relative effects of spectral and temporal properties on timbre, Strong and Clark (1967a, b) interchanged spectral and temporal envelopes on sounds by a number of wind instruments and found that the results depended on the instrument.

In an early MDS study supporting the importance of spectral properties, Plomp (1970) found a 0.86 correlation between a three-dimensional perceptual space and a three-dimensional physical space derived from differences in SPL outputs of a $\frac{1}{3}$-octave band filter. Later experiments by Plomp (1976) supported the formant theory of musical instruments. Wedin and Goude (1972), in another MDS study, found three principal factors, all derived from the long term average spectrum.

In a series of widely cited papers, Grey (1977, 1978), Grey and Gordon (1978) and Grey and Moorer (1977), were able to use MDS to correlate experimental results with three acoustical properties, both spectral and temporal. Similar results were reported later by Krumhansl (1989) and McAdams and others (1995).

Charbonneau (1981) reduced the information in the acoustic signal and concluded that humans perceive the time variations of groups of partials rather than individual ones.

Kendall (1986) found that steady-state-alone conditions were matched as well as using the entire unaltered signal and concluded that ''the perceptual importance of transients in defining the characteristic sounds of instruments has been overstated.''

The preceding resume is admittedly sketchy and slanted toward information about the frequency domain as that is more pertinent to the present study. Two recent articles do an excellent job of reviewing the entire timbre perception literature (McAdams, 1993; Hajda *et al.*, 1997).

There has been little work attempting computer ''perception'' of timbre, or musical instrument identification, particularly when compared to the large volume of research by the speech community on automatic speaker/speech recognition. This emphasis is understandable in view of the enormous range of practical applications of the latter, for example, the identification of a speaker calling a bank to

a)Electronic mail: brown@media.mit.edu

conduct a transaction and the enormous field of communications applications.

The literature on automatic identification of musical instruments consists for the most part of reports in conference proceedings. Kaminskyj and others (1995, 1996) reported preliminary calculations using temporal features. De Poli and Tonella (1993) used a neural net calculation to classify sounds with a procedure similar to Grey's and based on his parameters. Cosi and others (1994) used features based on an auditory model followed by a neural net to classify instrument tones and reported successful classification.

Langmead (1995a,b) used auditory-based features and found that a temporal factor, which he called spectral onset asynchrony, was the most effective in creating timbre categories. Kostek and Wieczorkowska (1997) reported on an examination of various sound parameters, initially from the frequency domain but with temporal features added, with the as yet unachieved goal of separating musical instrument sounds.

With rare exceptions the timbre perception studies as well as the computer identification work has used a single example of each instrument. The current study, reported by Brown (1997, 1998a, b, c), does not compute a distance measure between single examples of a number of classes of instruments, but rather determines whether a statistically significant number of instruments from two different classes can be grouped correctly. It is more similar in method and goal to speaker/speech studies than to previous timbre studies.

This study is motivated by a long-time interest in the properties of musical signals which give rise to human percepts, and the objective measurement of these properties by computer. While a calculation of this type cannot answer the question of whether humans are using the same information in their decisions, it can, nevertheless, provide a solid scientific conclusion as to whether there is sufficient information present in the chosen properties of the waveform for a correct decision.

There are also a number of practical applications to this problem as formulated. One is in the field of audio indexing (Wilcox, 1994), the goal of which is the automatic identification of the segments in an audio stream. A successful solution to this problem would eliminate the necessity of sequential monitoring by a human. Another application is the long-studied automatic transcription problem (Moorer, 1975), which has as goal the conversion of an audio stream into a written score. Finally, and potentially most important, the explosion of internet sites has made automatic recognition methods of great importance in classifying and reducing the sheer volume of material to be searched on the internet.

The current study was inspired by one of the most successful methods of automatic speaker identification (Reynolds and Rose, 1995 and references therein), which involved the use of a Gaussian mixture model with cepstral coefficients as features. It can be argued that sound production by a musical instrument is analogous to quasi-steady vowel production by a singer in which the shape of the vocal tract determines the spectrum (Strong, 1998). This seems the appropriate description for the case of instruments, such as strings and natural horns, where the spectrum is largely determined by a resonator of fixed shape. The vowel analog may be less obvious in the case of instruments, such as the woodwinds of this study, where the dimensions of the resonator change to produce different notes.

Independent of analog chosen, a set of features used successfully for speech/vowel/speaker identification are the cepstral coefficients which have been used in this study. The method of analysis to be described does not take into account the temporal evolution of the spectral features as is done with hidden Markov models in speech identification, but rather treats all data frames independently, as done with speaker identification.

In choosing instruments for this study it was desirable to have similar (or at least overlapping) frequency ranges, a similar excitation mechanism, and similar resonators. The oboe and the saxophone met these criteria because they are both conical bore, reed instruments with overlapping frequency ranges, and in the case of the soprano sax, the frequency range is almost identical. They have similar attacks and decays, and a scientist/woodwind musician (Coleman, 1997) reported that the soprano sax and oboe can be difficult to distinguish when playing music of a similar style. Preliminary experiments with human subjects supported this observation. Since humans are thought to be the ultimate receivers, these seemed to be an extremely challenging pair of instruments to study for an initial test of this method as applied to musical instruments.

I. METHOD OF PATTERN RECOGNITION

A. Introduction

The goal of a pattern recognition calculation is to classify a group of patterns called the test set into two or more classes. This is done by calculating features for the test set and comparing them to the same features for known examples of the classes called the training set; i.e., the computer ''learns'' what the values of the features are for an example of each of the known classes. These are then used as a basis for comparison to the unknown sounds.

This process is analogous to that followed by human subjects in classifying sounds. Humans build up mental representations of sounds made by different sources. Then, when asked to classify a new, unknown sound, they compare it to each of these mental representations to make a decision. See, for example, Galotti (1994) or any other introductory text on cognitive psychology.

For computer identification, an unknown sound can be segmented into shorter ''frames'' and a feature vector calculated for each of these frames. If the feature vector is of dimension M, then each frame contributes a point in an M-dimensional space. The classifier must then adopt a distance measure for determining whether the given points are closer to those of class A or B (in the case of two classes) and, based on this, make a likelihood decision, i.e., make the decision that minimizes the probability of error.

B. Feature selection: Cepstral coefficients

In any pattern recognition problem the most crucial step is the choice of features since this determines whether the

classes can be differentiated. In this work the musical instrument was modeled as a resonator with a periodic excitation, with the sound produced analyzed in the same manner as in the speaker identification work previously mentioned. In speech analysis the glottal impulses are treated as a periodic excitation followed by a filter, which is the vocal tract or resonator (Rabiner and Schafer, 1978). The musical analog for reed woodwind instruments is the pressure-controlled opening and closing of the reed(s) delivering puffs of air into a cylindrical or conical bore resonator.

A set of features which has been particularly successful in characterizing the vocal tract resonances which identify individual speakers, speech, or vowels are the cepstral coefficients. See Rabiner and Schafer (1978) and Rabiner and Juang (1993) for a discussion of the use of cepstra for speech applications. The cepstrum is the Fourier transform of the log magnitude spectrum (Oppenheim and Schafer, 1975); it involves two transforms which makes it computationally more intensive than FFT-based calculations. These features will be applied to musical instruments to determine whether they offer instrument specific information sufficient to differentiate the instruments producing the sounds.

For musical signals the information in a constant Q transform is more useful than that of a linear FFT as the frequencies can be chosen to map directly on to the notes of the musical scale (Brown, 1991). The cochlea of the ear, except at the low-frequency end, is usually modeled by a third octave filter bank which has a ratio of adjacent center frequencies of $2^{1/3}$ or 1.26. In this implementation, the signal input was divided into 23-ms frames and a FFT was calculated. A constant Q transform equivalent to a third octave filter bank was calculated from these Fourier coefficients using the method described by Brown and Puckette (1992) with Hamming windowed kernels. Although the previous description was for $\frac{1}{12}$- or $\frac{1}{24}$-octave filters, it is equally applicable for any desired filter bandwidth.

The transformation from constant Q coefficients to cepstral coefficients was carried out using Eq. 10.1 from O'Shaughnessy (1987),

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

for $n=1,2,...,N$.

Here $X^{cq}[k]$ is the kth constant Q coefficient, $M=18$, and $N=18$ to give 18 cepstral coefficients.

In this implementation the constant Q coefficients were roughly equivalent to a third octave filter bank with 18 coefficients and center frequencies from 100 to 5439 Hz. They were transformed to 18 cepstral coefficients using Eq. (1) above. The lower limit of 100 Hz was chosen to match that of the lowest note of the tenor sax and a frequency ratio of 1.265 was chosen to give an upper frequency below the Nyquist for a sampling rate of 11 025 Hz.

C. Method of clusters

Rather than comparing all points in an unknown sound to all points in the training set, Popat and Picard (1997) have used the method of clustering (Therrien, 1989) to summarize the training data for each class. The method of clustering involves grouping points calculated from the 23-ms frames of the known, or ''training'' sound, into so-called clusters. Thus each class can be characterized by a number of clusters. After determining the parameters for the clusters, an arbitrary probability density for each class can be modeled as a sum of weighted Gaussians, called a Gaussian mixture model. The probability that each point of the unknown sound belongs to that class can then be calculated.

The optimum number of clusters can be determined either by software or by heuristics. Once the number of clusters is decided, the cluster ''centers'' for the training set are determined by choosing a distance measure between the training samples and the cluster centers. A common method, called the K-means algorithm, minimizes the Euclidian distances between the sample features and the cluster centers by an iterative procedure. It can be shown that the mean of the feature values of the samples assigned to a given cluster will minimize this distance. Thus the kth cluster can be characterized by a mean $\boldsymbol{\mu}_k$ and a standard deviation $\boldsymbol{\sigma}_k$. For the case of a single cluster, the training data are represented by the mean and standard deviation of the totality of feature values for this sound. The mathematics of this process is described in the section on calculations.

II. EXPERIMENTAL SETUP

A. Sound database and processing

In order for a study of this type to be statistically valid, it should include examples by many different instruments of the same class. This means doing in-house recordings of mostly amateurs with instruments which are probably not examples of the highest quality available, or of taking performances by professionals from recordings. There is also the choice of whether to take excerpts from real performances or to study isolated notes or sequences of notes. For example, the McGill set of CDs is available for studies of single notes usually by a single instrument.

All of these studies are of interest, but it was decided that the most general and challenging study, the one which was most similar to speaker/speech problems as well as having greater relevance to the other applications mentioned in the Introduction, would involve the study of excerpts from real performances by experts. This meant taking these excerpts from recordings and gave the widest variety of sounds from each of these instruments as could possibly be found.

All of the sound samples in this study were excerpted either from the Wellesley College Music Library collection of compact disks, audio cassettes, and records or from the personal collection of Jay Panetta, a member of the Music Department faculty. These were listened to and segments of solo oboe and sax were excerpted for the experiment. The excerpts were of solos by the instruments in order to insure that the features calculated belonged to the instrument class to be identified. Each of the excerpts involved a number of different notes, the intention being to capture a random selection on the recording as would be appropriate for an audio indexing problem.

It was particularly difficult to get examples of solo sax

music, and there remains a small amount of background (low intensity contribution from other instruments in the ensemble) music in a few of these sax sounds. There are soprano, alto, and tenor saxophone sounds included in the study. Ideally it would have been preferable to include only soprano sax sounds, but there would not have been a large enough number of sounds for good statistics. In practice this was of less concern since the subjects in the expert group reported below were unable to identify the type of sax reliably and the documentation on the recording often did not include information about the type of sax in the performance.

For the calculation of features, the sounds were downsampled from a 44 100-Hz sampling rate to 11 025 Hz since an upper range of 5500 Hz was thought to be sufficient to characterize the sounds, and this speeded up the calculations. The resampling software is part of Dan Ellis' dspB software (Ellis, 1992). It consists of a Hanning-windowed ideal sinc interpolator in a polyphase rational-intermediate-frequency implementation, where both the window length and the sampling-frequency ratio are chosen to allow arbitrary accuracy in aliasing rejection and output sampling rate. A 256-point Fourier transform was taken with a hop size of 128 points giving 23-ms segments with frames overlapping by 11.5 ms. Frames with average amplitude less than 600 (for 16-bit samples) were dropped.

B. Training set

As is customary in pattern recognition problems, the system must be trained on examples of each of the classes. An oboe sound and a sax sound of approximately 1-min duration representing each of the instrument classes were chosen for the training set. Alternatively, a number of shorter sounds by different instruments of each class might also have been used, but it was thought to be of interest to determine whether the system could be trained on a single instrument.

There were two such long oboe sounds from different instruments and four long sax sounds also from different instruments. The calculations were carried out with each of the two oboe sounds paired with each of the four sax sounds (eight combinations with a single representative of each class in any one calculation) to determine the effect of choice of training sounds. The best results were obtained with a 61-s (5254 frames) excerpt from a performance of Peter Christ playing Persichetti's ''Parable for Solo Oboe'' and a 53-s (4565 frames) tenor saxophone excerpt of Archie Shepp from the CD ''Yasmina/Poem for Malcolm.'' The average over the entire time interval of the 18 constant Q coefficients for these two sounds can be found in Figs. 1 and 2. These spectra are quite distinctive. The oboe has a formant at around 1300 Hz in agreement with the literature (Strong and Plitnik, 1977) which attributes oboe formants at around 1000 and 3000 Hz to the mechanical properties of the reeds (Fransson, 1967). Rossing (1990) summarizes these results. The energy in the first three coefficients in this figure is somewhat perplexing, as the corresponding frequencies are below the normal range of the oboe. The recording is taken from a record, so possibly there is turntable noise, or there might be subharmonics for some of the notes which occur. In any case the magnitudes of these coefficients are well below the peak energy of this spectrum.

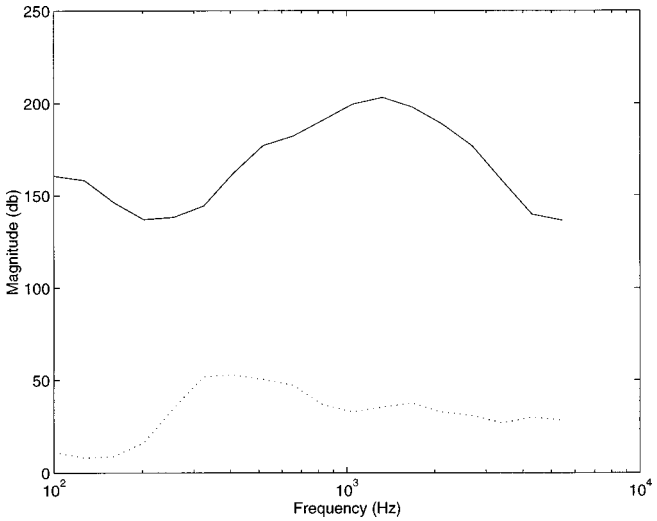

FIG. 1. Constant Q coefficients (−) and standard deviations (···) for Persichetti's ''Parable for Solo Oboe'' averaged over 61 s. Frequencies vary from 100 Hz to 5.43 kHz by third octaves.

The saxophone spectrum has a broad peak at around 500–600 Hz, well below this. Cepstral coefficients calculated from these constant Q coefficients using Eq. (1) are shown in Figs. 3 and 4. These are the mean values of these coefficients averaged over the entire sound. Notes from $A\sharp_2$ to G_5 were included in the sound from Figs. 2 and 4. The mean note was $G\sharp_3$ with a standard deviation of eight semitones. Notes from B_3 to G_6 were included for the oboe (Figs. 1 and 3) with a mean of $C\sharp_5$ and a standard deviation of eight semitones as well.

For comparison to the spectra of Figs. 1 and 2, recordings of soprano, alto, and tenor saxophones and an oboe playing the same chromatic scale from C_4 to B_4 were taken from the University of Iowa Electronic Music Studios Web Site (http://theremin.music.uiowa.edu/). This octave was chosen because it is in a range accessible to each of these

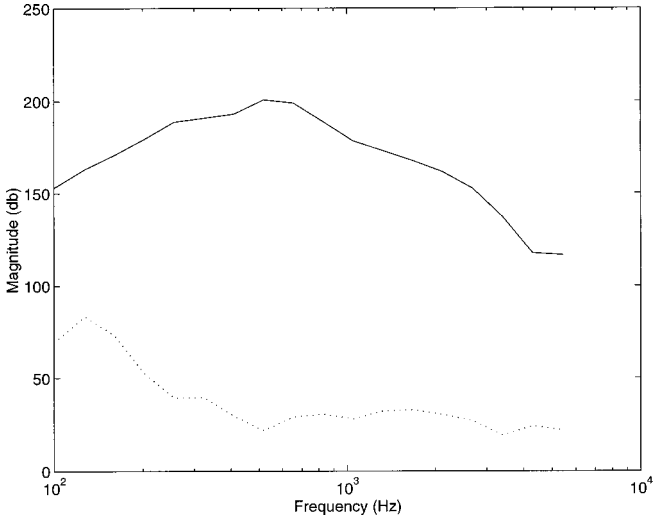

FIG. 2. Constant Q coefficients (−) and standard deviations (···) for ''Yasmina/Poem for Malcolm'' for saxophone averaged over 53 s. Frequencies vary from 100 Hz to 5.43 kHz by third octaves.

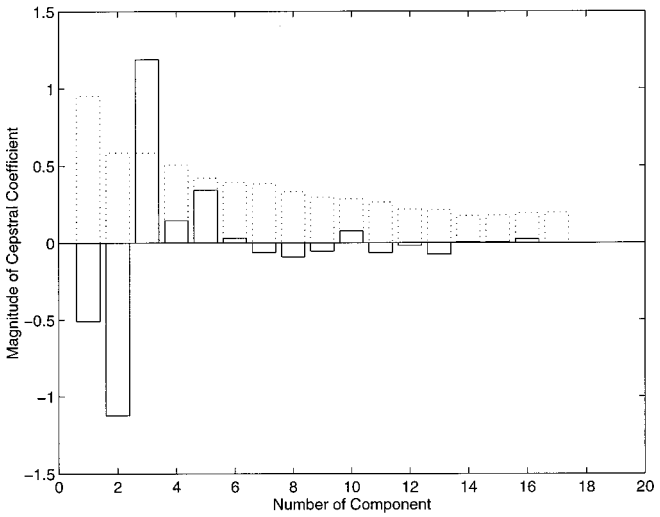

FIG. 3. Cepstral coefficients (−) and standard deviations (···) for Persichetti's ''Parable for Solo Oboe'' averaged over 61 s.

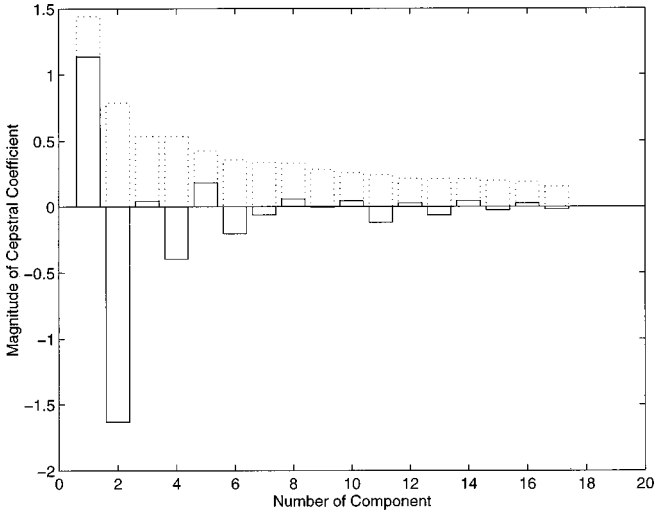

FIG. 4. Cepstral coefficients (−) and standard deviations (···) for ''Yasmina/Poem for Malcolm'' for saxophone averaged over 53 s.

instruments, and it is possible to compare spectra with the same notes being played. See Fig. 5 for their average constant Q spectra. The resolution of this transform was the same as that of the previous two figures, but a sampling rate of 32 000 Hz rather than 11 025 Hz was chosen to show the behavior of the higher frequency components. The oboe shows a resonance peak at 1400 Hz similar to that of Fig. 1; the soprano sax has a smaller peak roughly centered at 800 Hz and the alto and tenor spectra peak at around 400 Hz though the tenor has a second smaller peak at slightly over 1000 Hz. All of the sounds dropped off from their maximum values by around 50 dB or more at 5000 Hz showing that little information was lost by resampling at 11 025 for the calculations.

Finally, examples of the variation of the cepstral coefficients with note can be found in Figs. 6 and 7. The low-quefrency values are similar for the notes of a given instrument class, and it is these which give the most important information on the resonator.

C. Test set

Initially a set of 31 sax sounds was included in the study. Excluding two long sounds, these were of average length 2.0 s and standard deviation 0.8 s, and the perception experiment with the expert group was carried out on these sounds. Later, with the help of Jay Panetta, more solo sax sounds were added to the study since it was felt that some of the sounds in the initial group were less ''clean'' than was desirable. All of these sounds were included in the machine identification calculations. This latter group of 21 sounds was longer with a mean of 7.8 s and s.d. of 2.4 s.

There were 28 oboe sounds included in the study with average length 2.5 s and standard deviation 2.1 s. One of the

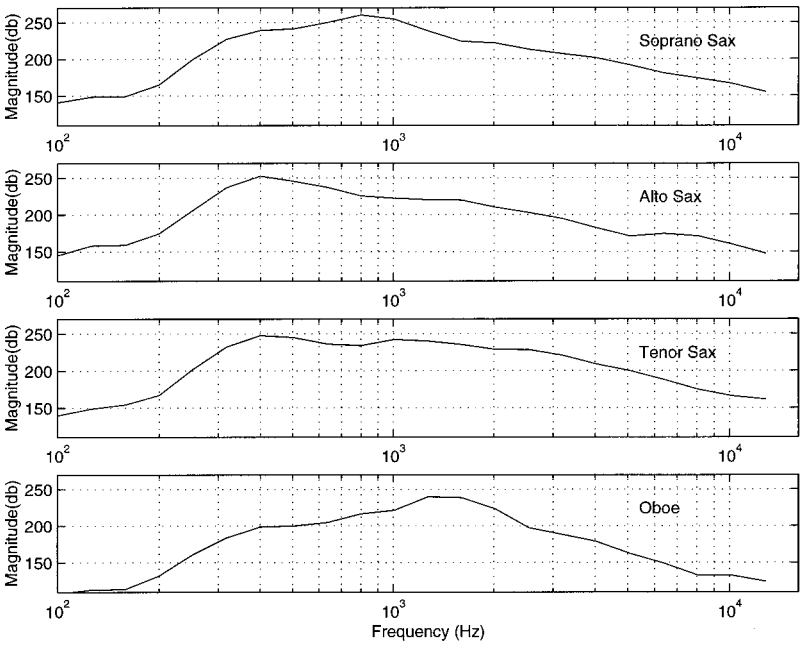

FIG. 5. Average constant Q spectra of soprano, alto, and tenor saxophones and an oboe playing a two octave chromatic scale from C_4 to B_4.

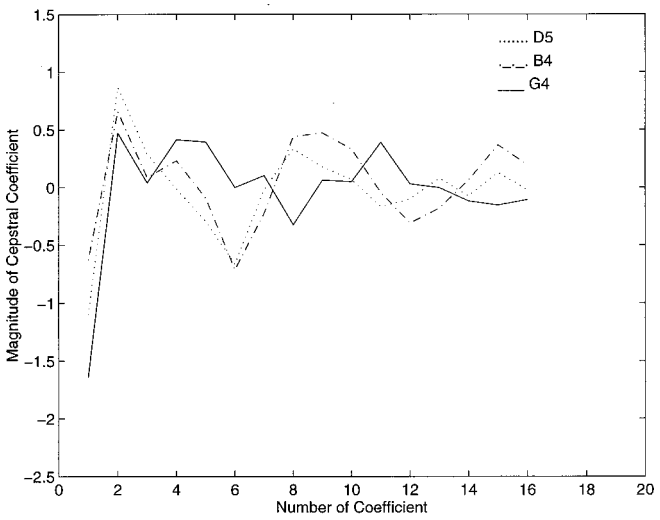

FIG. 6. Example of cepstral coefficients for three notes for oboe excerpted from Mozart's ''Concerto for Oboe and Orchestra in C'' K314 played by Niesemann.

sounds was inadvertantly omitted from the perception experiment by experts reported in Table I, but is included in the computer identification experiment.

III. CALCULATIONS

A. Background

A *K*-means algorithm written by Kris Popat was used in all calculations to determine the cluster means and variances. The number of clusters was an input parameter allowing the results to be checked for optimum performance.

In order to classify the unknown sounds into two classes A and B (oboe and sax in this 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. For example, given *N* 23-ms segments of an unknown sound *U*, then for each of these segments, a feature

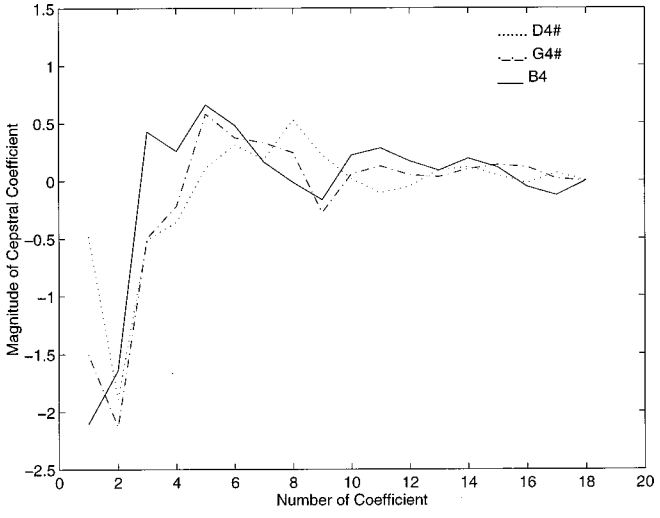

FIG. 7. Example of cepstral coefficients for three notes for soprano sax from a CD of demonstration sounds.

TABLE I. Summary of results on human and computer instrument identification. Each column gives the average fractional error for each experiment (average per person errors divided by the number of sounds presented). Listening conditions are described in the section on human perception results.

	Controlled env	Auditorium	Computer
Oboe	4/27=0.15	2.7/16=0.17	1/28=0.04
Sax	2.5/31=0.08	2/17=0.12	5/52=0.10

vector $\mathbf{x}^i$ with components $x_1^i, x_2^i, \ldots, x_{18}^i$ was calculated, where the components were the cepstral coefficients for the ith segment and $i=1,\ldots,N$.

Then the probability density of measuring feature vector $\mathbf{x^i}$ for cluster k of class Ω if U is a member of class Ω is:

$$p(\mathbf{x^i}|\Omega_k)=\frac{1}{\sqrt{2\pi\boldsymbol{\sigma}^2_{\boldsymbol{\Omega}_\mathbf{k}}}}\exp-(\mathbf{x^i}-\boldsymbol{\mu}_{\boldsymbol{\Omega}_\mathbf{k}})^2/2\boldsymbol{\sigma}^2_{\boldsymbol{\Omega}_\mathbf{k}}. \tag{2}$$

Using a Gaussian mixture model (Reynolds and Rose, 1995) to model the probability density function best describing the training data, and summing over all clusters, then the total probability density that feature vector $\mathbf{x^i}$ is measured if U belongs to class Ω is

$$p(\mathbf{x^i}|\Omega)=\sum_{k=1}^{K} p_k p(\mathbf{x^i}|\Omega_k), \tag{3}$$

where p_k is the probability of occurrence of the kth cluster. It is equal to the number of vectors in the training set assigned to this cluster divided by the total number of vectors in the training set.

The total probability density that all of the *N* feature vectors measured for unknown *U* belong to class Ω is given by the product of the individual probability densities:

$$p(\mathbf{X}|\Omega)=p(\mathbf{x^1}\cdots\mathbf{x^N}|\Omega)=\prod_{i=1}^{N} p(\mathbf{x^i}|\Omega), \tag{4}$$

where $\mathbf{X}=\{\mathbf{x^1}\cdots\mathbf{x^N}\}$ is defined as the set of all feature vectors measured for unknown sound *U*. This assumes statistical independence of the feature vectors. While this simplifying assumption is not strictly valid here, it is a widely accepted technique in the speech community and has been experimentally shown to be effective in calculations (Rabiner and Huang, 1993).

Equation (4) is the probability density of measuring the set of feature vectors $\mathbf{X}$ for unknown *U* if *U* belongs to class Ω. The quantity of interest for the Bayes decision rule is the *a posteria* probability that a measurement of $\mathbf{X}$ means it is more probable that *U* is a member of class Ω than another class. Letting $\Omega^{(m)}$ where $m=1,2,\ldots,M$ represent the *M* classes, then the desired *a posteria* probability is:

$$\hat{\Omega}=\arg\max \Pr(\Omega^{(m)}|\mathbf{X}),$$

where $1\leqslant m\leqslant M$ and $\hat{\Omega}$ is the class which maximizes this probability.

From Bayes' rule,

$$\hat{\Omega} = \arg\max \frac{p(\mathbf{X}|\Omega^{(m)})\Pr(\Omega^{(m)})}{\Pr(\mathbf{X})}. \quad (5)$$

The probability of measuring the set of features **X** is

$$\Pr(\mathbf{X}) = \sum_{i=1}^{M} p(\mathbf{X}|\Omega^{(m)})\Pr(\Omega^{(m)}) \quad (6)$$

and is the same for all classes. Similarly, if the classes are equally probable, $\Pr(\Omega^{(m)}) = 1/M$ independent of the class. This is the case with equal numbers of test sounds, but it is true as well if there is no *a priori* reason for one class's being more probable. Although in the present experiment the numbers of members of the classes in the test set was known, the goal of the experiment is to be able to distinguish between instruments in a situation where there is no information as to the composition of the test set. It would run counter to this purpose to make an assumption about a particular test set which would not be valid for the general case.

Dropping functions which do not vary with class,

$$\hat{\Omega} = \arg\max p(\mathbf{X}|\Omega^{(m)}).$$

For $1 \leqslant m \leqslant M$, and this is the probability density in Eq. (4).

With two classes $\Omega^{(1)}$ and $\Omega^{(2)}$, then if

$$p(\mathbf{X}|\Omega^{(1)}) > p(\mathbf{X}|\Omega^{(2)}),$$

the unknown U is assigned to class $\Omega^{(1)}$, and otherwise to class $\Omega^{(2)}$. This is called a likelihood ratio test. It is a decision rule which minimizes the probability of error and is thus an optimal decision rule. Equivalently, this can be stated as the log likelihood ratio:

$$\log(p(\mathbf{X}|\Omega^{(1)})) - \log(p(\mathbf{X}|\Omega^{(2)})) > 0, \quad (7)$$

which is the function graphed in the results section. This function was chosen for reasons of convention and ease of interpretation, since it is easy visually to pick out the positive cases.

B. Results

1. Human perception

a. Controlled listening experiment. The first experiment on human instrument identification used seven expert listeners as subjects. These were people who had had extensive listening and/or performing experience with these instruments. Twenty-seven oboe samples and 31 sax samples were presented, and the subjects were asked to classify them as either oboe or sax. If they were classified as sax, the subjects were also asked to identify the kind of sax. There was imperfect agreement among subjects in this secondary classification, and these experiments were not pursued.

The samples were presented in a controlled listening environment, with the subjects either listening over speakers in a soundproof recording studio or over headphones in a quiet room. The subjects could listen to a given sound as often as they wished before making a choice. The experiment was forced choice, so the subjects guessed if they were otherwise unable to make a decision.

The results are presented in Table I. The subjects made an average of 4 errors per person for an average error of 15% for the oboe and an average of 2.5 errors for the sax for an average error of 8%.

b. Auditorium experiment. Another test with quite different listening conditions was carried out in an auditorium using as subjects the 32-member audience at a talk on musical perception. These subjects had varied musical backgrounds ranging from members of the Wellesley College Music Department to people with no formal musical training, but all had an interest in music or perception since they had chosen to attend the lecture. The sound system in the room was excellent as it was designed for a musical acoustics course rather than for a typical auditorium. The subjects were presented with roughly half the sounds that had been heard by the previous group. Each sound was followed by 4 s of silence with a forced choice of oboe or sax. The total number of sounds was reduced compared to the expert listener group as it was felt that a 3-min test was sufficiently taxing for the patience of this captive group.

The results are shown in the third column of Table I. They essentially duplicated the 15% error rate for the previous group for the oboe, and had a slightly higher 12% error rate for the sax sounds.

2. Machine identification

Figure 8 summarizes the results on machine identification using the optimum choice of training sounds (indicated in Figs. 1–4) and number of clusters, which were three for the oboe and one for the sax. It is a plot of the log probability that each sound fits the Gaussian describing the correct class minus the log probability for the other class from Eq. (7). The upper curve is for the oboe sounds and the lower curve for the sax sounds. The x axis is numbered by sound sample. For both curves the positive values represent correct choices by the computer and the negative values are misses. There are five errors for the sax and one for the oboe, although it could be argued that one of the sax errors was sufficiently close to zero to be called no decision. The average number of errors varying all possible training sounds and varying the number of clusters from one to three was 15.9% total with a standard deviation of 6.8%.

IV. DISCUSSION AND CONCLUSIONS

Overall results are summarized in Table I comparing the average fractional error (average per person number of errors over the number of sounds evaluated) for each instrument with those of the computer. The computer has a lower error rate than the humans for the oboe samples and an error rate roughly the same for the sax samples. This success for the computer occurred despite the advantage of context which some of the human subjects commented upon as an important indicator for them, i.e., the humans assumed, usually correctly, that jazz sounds were played by a saxophone. The computer based its decision on feature values and was, of course, unable to distinguish jazz from classical styles.

The similarity of the success rates of human and machine identification is interesting for two reasons. First, it

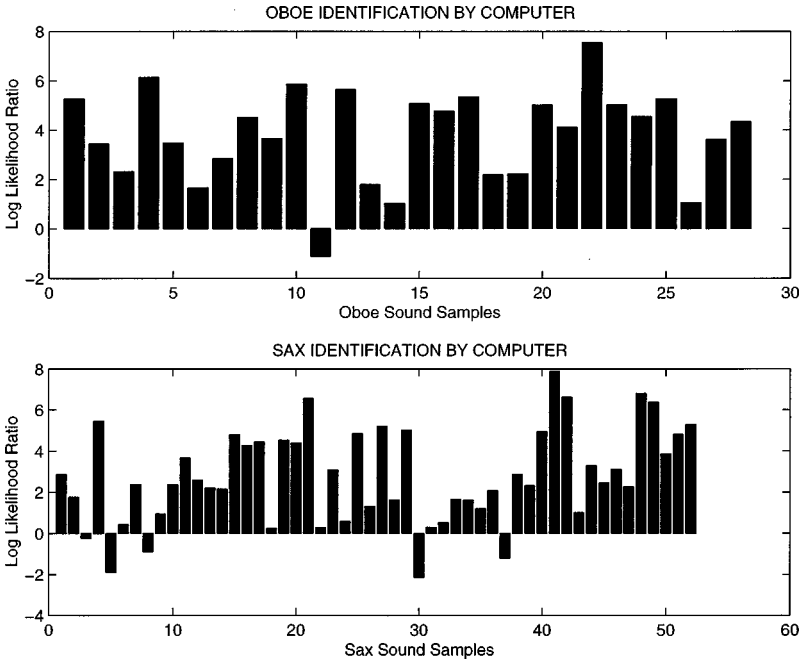

FIG. 8. Log likelihood ratio for oboe and sax sounds plotted against sound sample number. The calculations were done with three clusters for the oboe and one cluster for the sax.

should be recalled that humans have trained for many years on many different pieces of sax and oboe music; whereas the computer has only approximately a minute of one particular saxophone performance and one particular oboe performance to analyze and ''learn'' the oboe and sax features. Second, the success of the machine with this method is somewhat surprising since it is based totally on spectral information, whereas it is generally thought that human instrument identification is based on temporal as well as spectral cues.

The study of musical sounds using cepstral coefficients as features holds promise in a number of areas. As computations on properties of instruments by the study of the sounds they produce, they give new and complementary information, adding to knowledge from impedance and other passive methods. New knowledge of instrument properties can also lead to new information about human perception, as spectral cues are certainly of great importance to humans in their decision processes.

Finally, just as no number of experiments can fully validate a scientific theory, while a single counterexample can refute it, a computer identification experiment cannot prove what cues are actually used in human perception but rather demonstrate whether sufficient information for identification is present in the features studied. Thus, this study cannot assess the relative importance of spectral and temporal information used by humans for instrument identification, but it can indeed affirm that there is sufficient information in features derived from spectral properties to distinguish between these two instruments.

ACKNOWLEDGMENTS

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 much additional time explaining how to use it. I would also like to thank Deb Roy for his extremely helpful suggestions and Douglas Reynolds for his kindness in answering e-mail questions about his papers. Julie Pollack and Jay Panetta of Wellesley College were invaluable sources of information for the oboe and sax collections and gave generously of their time. I would also like to thank Jay Panetta, William Coleman, and the Machine Listening Group of the Media Lab for participating as subjects in the expert listener experiment. Finally, many thanks to Larry Fritts and Josh Nichols for their efforts in getting the sax scales up on the University of Iowa Electronic Music Studio web site.

Brown, J. C. (**1991**). ''Calculation of a constant *Q* spectral transform,'' J. Acoust. Soc. Am. **89**, 425–434.

Brown, J. C. (**1997**). ''Cluster-based probability model for musical instrument identification,'' J. Acoust. Soc. Am. **101**, 3167A.

Brown, J. C. (**1998a**). ''Computer Identification of Wind Instruments using Cepstral Coefficients,'' in *Proceedings of the 16th International Congress on Acoustics and 135th Meeting of the Acoustical Society of America*, Seattle, Washington, pp. 1889–1890.

Brown, J. C. (**1998b**). ''Computer identification of wind instruments using cepstral coefficients,'' J. Acoust. Soc. Am. **103**, 2967 (abstract).

Brown, J. C. (**1998c**). ''Musical Instrument Identification using Autocorrelation Coefficients,'' in *Proceedings International Symposium on Musical Acoustics 1998*, Leavenworth, Washington.

Brown, J. C., and Puckette, M. S. (**1992**). ''An efficient algorithm for the calculation of a constant *Q* transform,'' J. Acoust. Soc. Am. **92**, 2698–2701.

Charbonneau, G. R. (**1981**). ''Timbre and the perceptual effects of three types of data reduction,'' Comput. Music J. **5**, 10–19.

Clark, M., Robertson, P., and Luce, D. A. (**1964**). ''A preliminary experiment on the perceptual basis for musical instrument families,'' J. Audio Eng. Soc. **12**, 199–203.

Coleman, W. F. (**1997**). Private communication.

Cosi, P., De Poli, G., and Lauzzana, G. (**1994**). ''Timbre classification by NN and auditory modeling,'' in *Proceedings of the International Conference on Artificial Neural Networks*, pp. 925–928.

De Poli, G., and Tonella, P. (**1993**). ''Self-organizing neural networks and Grey's timbre space,'' ICMC Proceedings, pp. 441–444.

Ellis, D. P. W. (**1992**). ''A Perceptual Representation of Audio,'' MS thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Fransson, F. (**1967**). ''The Source Spectrum of Double-Reed Woodwind Instruments,'' Report STL-APSR 1, 25 Royal Institute of Technology, Stockholm.

Galotti, K. M. (**1994**). *Cognitive Psychology In and Out of the Laboratory* (Brooks/Cole, Belmont, CA).
Grey, J. M. (**1977**). ''Multidimensional perceptual scaling of musical timbres,'' J. Acoust. Soc. Am. **61**, 1270–1277.
Grey, J. M. (**1978**). ''Timbre discrimination in musical patterns,'' J. Acoust. Soc. Am. **64**, 467–472.
Grey, J. M., and Gordon, J. W. (**1978**). ''Perceptual effects of spectral modifications on musical timbres,'' J. Acoust. Soc. Am. **63**, 1493–1500.
Grey, J. M., and Moorer, J. A. (**1977**). ''Perceptual evaluations of synthesized musical instrument tones,'' J. Acoust. Soc. Am. **62**, 454–462.
Hajda, J. M., Kendall, R. A., Carterette, E. C., and Harshberger, M. L. (**1997**). ''Methodological issues in timbre research,'' in *Perception and Cognition of Music*, edited by I. Deliege and J. Sloboda (Psychology Press, East Essex, UK), pp. 253–307.
Helmholtz, H. L. F. (**1885/1954**). *On the Sensations of Tone* (Dover, New York).
Kaminskyj, I., and Materka, A. (**1995**). ''Automatic Source Identification of Monophonic Musical Instrument Sounds,'' IEEE Int. Conf. on Neural Networks **1**, 189–194.
Kaminskyj, I., and Voumard, P. (**1996**). ''Enhanced Automatic Source Identification of Monophonic Musical Instrument Sounds,'' in *Proceedings of the 1996 Australian New Zealand Conference on Intelligent Information Systems*, pp. 76–79.
Kendall, R. A. (**1986**). ''The role of acoustic signal partitions in listener categorization of musical phrases,'' Music Perception **4**, 185–214.
Kostek, B., and Wieczorkowska, A. (**1997**). ''Parametric Representation of Musical Sounds,'' Arch. Acoust. **22**, 3–26.
Krumhansl, C. L. (**1989**). ''Why is musical timbre so hard to understand?'' in *Structure and Perception of Electroacoustic Sound and Music*, edited by S. Nielzen and O. Olsson, Excerpta Medica 846 (Elsevier, Amsterdam), pp. 43–53.
Langmead, C. J. (**1995a**). ''Sound Analysis, Comparison and Modification Based on a Perceptual Model of Timbre,'' ICMC **1995**, 475–478.
Langmead, C. J. (**1995b**). ''A theoretical model of timbre perception based on morphological representations of time-varying spectra,'' Master's Thesis, Dartmouth College.
McAdams, S., and Bigand, E. (**1993**). ''Recognition of Auditory Sound Sources and Events,'' in *Thinking in Sound: The Cognitive Psychology of Human Audition* (Oxford U.P., Oxford).
McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (**1995**). ''Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes,'' Psychological Research **58**, 177–192.
Moorer, J. A. (**1975**). ''On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer,'' Ph.D. dissertation, Stanford Department of Music Report No. STAN-M3.
O'Shaughnessy, D. (**1987**). *Speech Communication: Human and Machine* (Addison–Wesley, Reading, MA).
Oppenheim, A. V., and Schafer, R. W. (**1975**). *Digital Signal Processing* (Prentice–Hall, Englewood Cliffs, NJ).
Plomp, R. (**1970**). ''Timbre as a multidimensional attribute of complex tones,'' in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (Sijthoff, Leiden), pp. 397–414.
Plomp, R. (**1976**). *Aspects of Tone Sensation* (Academic, London).
Popat, K., and Picard, R. W. (**1997**). ''Cluster-Based Probability Model and Its Application to Image and Texture Processing,'' IEEE Trans. Image Process. **6**, 268–284.
Rabiner, L. R., and Huang, B.-H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).
Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signals* (Prentice-Hall International, London).
Reynolds, D. A., and Rose, R. C. (**1995**). ''Robust text-independent speaker identification using Gaussian mixture speaker models,'' IEEE Trans. Speech Audio Process. **3**, 72–83.
Risset, J. C., and Mathews, M. V. (**1969**). ''Analysis of musical-instrument tones,'' Phys. Today **22**, 23–30.
Rossing, T. D. (**1990**). *The Science of Sound* (Addison-Wesley, Reading, MA).
Saldanha, E. L., and Corso, J. F. (**1964**). ''Timbre cues and the identification of musical instruments,'' J. Acoust. Soc. Am. **36**, 2021–2026.
Strong, W. (**1998**). Private communication.
Strong, W., and Clark, M. (**1967a**). ''Synthesis of wind-instrument tones,'' J. Acoust. Soc. Am. **41**, 39–52.
Strong, W., and Clark, M. (**1967b**). ''Perturbations of synthetic orchestral wind-instrument tones,'' J. Acoust. Soc. Am. **41**, 277–285.
Strong, W. J., and Plitnik, G. R. (**1977**). *Music, Speech, and High Fidelity* (Brigham Young Univ., Provo, UT).
Therrien, C. W. (**1989**). *Decision Estimation and Classification* (Wiley, New York).
Wedin, L., and Goude, G. (**1972**). ''Dimensional analysis of the perception of instrument timbres,'' Scandinavian Journal of Psychology **18**, 228–240.
Wilcox, L., Kimber, D., and Chen, F. (**1994**). ''Audio indexing using speaker identification,'' ISTL Technical Report No. ISTL-QCA-1994-05-04.