Feature dependence in the automatic identification of musical woodwind instruments

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

Olivier Houix and Stephen McAdams
Institut de Recherche et de Coordination Acoustique/Musique (Ircam-CNRS), 1 place Igor Stravinsky, F-75004 Paris, France

(Received 18 May 1999; revised 16 November 2000; accepted 22 November 2000)

The automatic identification of musical instruments is a relatively unexplored and potentially very important field for its promise to free humans from time-consuming searches on the Internet and indexing of audio material. Speaker identification techniques have been used in this paper to determine the properties (features) which are most effective in identifying a statistically significant number of sounds representing four classes of musical instruments (oboe, sax, clarinet, flute) excerpted from actual performances. Features examined include cepstral coefficients, constant-Q coefficients, spectral centroid, autocorrelation coefficients, and moments of the time wave. The number of these coefficients was varied, and in the case of cepstral coefficients, ten coefficients were sufficient for identification. Correct identifications of 79%–84% were obtained with cepstral coefficients, bin-to-bin differences of the constant-Q coefficients, and autocorrelation coefficients; the latter have not been used previously in either speaker or instrument identification work. These results depended on the training sounds chosen and the number of clusters used in the calculation. Comparison to a human perception experiment with sounds produced by the same instruments indicates that, under these conditions, computers do as well as humans in identifying woodwind instruments. © *2001 Acoustical Society of America.* [DOI: 10.1121/1.1342075]

PACS numbers: 43.60.Gk, 43.75.Cd, 43.75.Ef [JCB]

I. INTRODUCTION AND BACKGROUND

Despite the massive research which has been carried out on automatic speaker identification, there has been little work done on the identification of musical instruments by computer. See Brown (1999) for a summary. Applications of automatic instrument identification include audio indexing (Wilcox *et al.*, 1994), automatic transcription (Moorer, 1975), and Internet search and classification of musical material.

One technique used widely in speaker identification studies is pattern recognition. Here, the most important step is the choice of a set of features which will successfully differentiate members of a database. Brown (1997, 1998a, 1999) applied this technique to the identification of the oboe and the saxophone using a Gaussian mixture model with cepstral coefficients as features. Included in this reference is an introduction to pattern recognition and to the method of clusters. Definitions which will be useful for this paper can be found in the Appendix.

Two later reports on computer identification of musical instruments also use cepstral coefficients as features for pattern recognition. Dubnov and Rodet (1998) used a vector quantizer as a front end and trained on 18 short excerpts from 18 instruments, but reported no quantitative classification results. Marques (1999) examined eight instruments trained on excerpts from one CD with the test set excerpted from other CDs (one per instrument class) and reported a 67% success rate. In a study which will be examined further in this paper, Dubnov *et al.* (1997) explored the effectiveness of higher-order statistics using the calculation of moments for musical instrument identification. They concluded that these features were effective in distinguishing families of musical instruments, but not the instruments within families. As with earlier work, none of these studies includes enough samples for statistically valid conclusions.

In marked contrast to the relatively few articles on automatic recognition of musical instruments, there has been a great deal of interest in human timbre perception. For comparison with this study, we focus on experiments involving the woodwind family. These instruments are difficult to distinguish from each other since they have similar attacks and decays, overlapping frequency ranges, and similar modes of excitation. The literature on these experiments is summarized in Table I. For a short, general summary of human perception experiments, see Brown (1999). For more complete reviews, see McAdams (1993), Handel (1995), and Hajda *et al.* (1997).

Although the vast majority of the experiments of Table I has been on single notes or note segments, Saldanha and Corso (1964) pointed out that the transitional effects from note to note could provide one of the major determiners of musical quality. In the earliest study including note-to-note transitions, Campbell and Heller (1978) found more accurate identifications using transitions than with isolated tones. They called the transition region the legato transient. In an-

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

TABLE I. Summary of percent correct for previous human perception experiments on wind instruments. Results for the oboe, sax, clarinet, and flute are given when possible. The final column is the total number of instruments included in the experiment.

	Date	Oboe	Sax	Clar	Flute	Overall	Number of instruments
Eagleson/Eagleson	1947		59	45	20	56	9
Saldanha/Corso	1964	75		84	61	41	10
Berger	1964					59	10
Clark/Milner	1964					90	3 (flute, clar, oboe)
Strong/Clark	1967a					85	8
Campbell/Heller	1978					72	6 (2-note legato)
Kendall	1986					84	3 (trumpet, clar, violin)
Brown	1999	85	92			89	2 (oboe, sax)
Martin	1999					46	27 (isolated tone)
						67	27 (10-s excerpt)
Houix/McAdams/Brown		87	87	71	93	85	4 (oboe, sax, clar, flute)

other study using musical phrases, Kendall (1986) emphasized the importance of context and demonstrated that results on musical phrases were significantly higher than on single notes.

More recently, Brown (1997, 1998a, 1998b, 1999) has found excellent results using multinote segments from actual musical performances. Martin (1999) has explored both types of experiments and found more accurate results with multinote segments than with isolated single notes. The results of Houix, McAdams, and Brown (unpublished) on multinote human perception will be compared to our calculations in a later section.

In this paper we have used a large database of sounds exerpted from actual performances with the oboe, saxophone, clarinet, and flute. We present calculations to show:

(i) The accuracy with which computers can be used to identify these very similar instruments;
(ii) The best signal processing features for this task; and
(iii) The accuracy compared with experiments on human perception.

II. SOUND DATABASE

A. Source and processing

Sounds were excerpted as short segments of solo passages from compact disks, audio cassettes, and records from the Wellesley College Music Library. This method of sample collection ensured a selection of typical sounds produced by each instrument, such as might be encountered on Internet sites or stored audio tapes. At least 25 sounds for each instrument were used to provide statistical reliability for the results. Features were calculated for 32-ms frames overlapping by 50% and having rms averages greater than 425 (for 16-bit samples).

B. Training and test sets

Sounds of longer duration (1 min or more) representing each instrument were chosen as training sounds and are given in Table II. These training sounds were varied in the calculations with one sound representing each instrument in all possible combinations to determine the optimum combination for identification. From Table II, with two, four, three, and four sounds for each of the four instruments, there were 96 combinations.

The constant-Q transforms of the most effective training sounds are shown in Fig. 1. Both the oboe and flute examples have strong peaks at a little over 1000 Hz. The oboe has an additional bump at 1200 Hz, giving rise to its nasal quality. The saxophone has a low-frequency spectral-energy distribution with a peak around 400 Hz, while the clarinet has less prominent peaks at around 400 and 900 Hz.

Properties of the test set are given in Table III. The training sounds were included in the identification calculations but were not included in the calculation of the average durations reported here. Two longer flute sounds with durations on the order of 40 s were also omitted as their durations were not representative of the flute data as a whole and skewed the average.

TABLE II. Training sounds identified by performer and piece of music performed. The third column is the length of the sound in seconds which was exerpted for the calculation.

Performer	Music	Length (s)
Peter Christ	Persichetti's Parable for Solo Oboe	60.7
Joseph Robinson	Rochberg's Concerto for Oboe and Orchestra	82.2
Frederick Tillis	''Motherless Child''	77.7
Johnny Griffin	''Light Blue''	99.3
Coleman Hawkins	''Picasso''	63.0
Sonny Rollins	''Body and Soul''	88.8
Benny Goodman	Copland's Concerto for Clarinet and String Orch	74.05
Heinrich Matzener	Eisler's Moment Musical pour clarinette Solo	70.1
David Shifrin	Copland's Concerto for Clarinet and String Orch	63.2
Samuel Baron	Martino's Quodlibets for Flute	74.1
Sue Ann Kahn	Luening's Third Short Sonata for Flute and Piano	69.0
Susan Milan	Martinu's Sonata for Flute and Piano	54.3
Fenwick Smith	Koechlin's Sonata for 2 Flutes Op 75	106.0

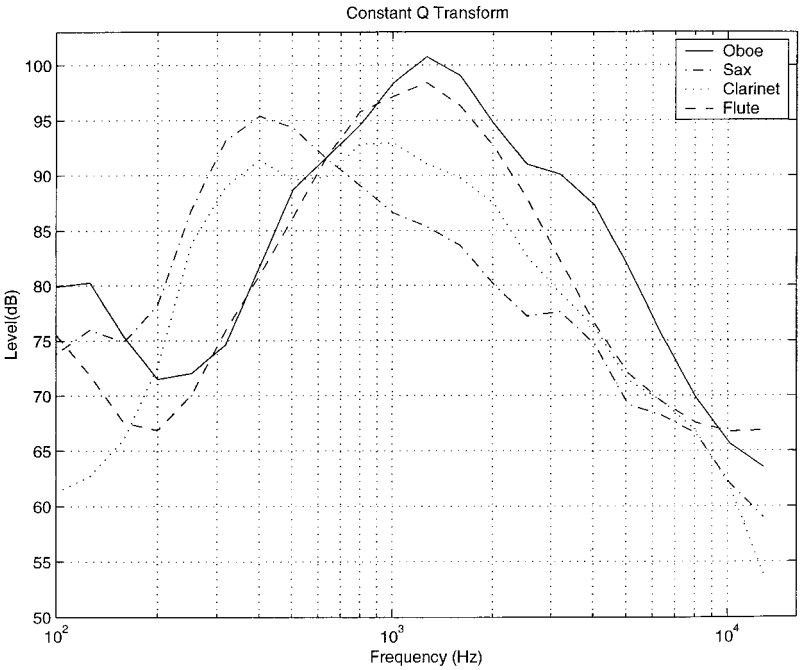

FIG. 1. Comparison of the constant-Q spectra for examples of successful training sounds for each of the four instrument classes. These were the sounds performed by Christ, Griffin, Matzener, and Baron. See Table II for details.

III. CALCULATIONS

A. Probability calculation

The details of the calculations described in Brown (1999) will be summarized here. For each training sound, cepstral coefficients or other features were calculated for each frame; and from these values, a *k*-means algorithm was used to calculate clusters. A Gaussian mixture model (Reynolds and Rose, 1995), i.e., a sum of weighted Gaussians, was then calculated based on the mean $\boldsymbol{\mu}_k$, standard deviation $\boldsymbol{\sigma}_k$, and population given by the cluster calculation for this sound; this model was used to give the probability density function representing the data calculated for the training sounds. For a single cluster k belonging to class Ω, the probability density of measuring the feature vector $\mathbf{x^i}$ is

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

Summing over all K clusters, the total probability density that feature vector $\mathbf{x^i}$ is measured if unknown sound $\mathbf{U}$ belongs to class Ω is

$$p(\mathbf{x^i}|\Omega)=\sum_{k=1}^{K} p_k p(\mathbf{x^i}|\Omega_k), \quad (2)$$

TABLE III. Data on sounds in the test set by instrument class. The number of sounds is given in column two with the average length and standard deviation in the last two columns.

Instrument	Number of sounds	Average length (s)	Standard deviation (s)
Oboe	28	2.5	2.1
Sax (Gp I)	31	2.0	0.8
Sax (Gp II)	21	7.8	2.4
Clarinet	33	6.1	2.1
Flute	31	7.8	4.1

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. If we define $\mathbf{X}=\{\mathbf{x^1},\ldots,\mathbf{x^N}\}$ as the set of all feature vectors measured for $\mathbf{U}$, then 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

$$p(\mathbf{X}|\Omega)=p(\mathbf{x^1},\ldots,\mathbf{x^N}|\Omega)=\prod_{i=1}^{N} p(\mathbf{x^i}|\Omega). \quad (3)$$

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). As the sounds used in the study had many rapid note changes, it proves a better assumption here than for speech. Equation (3) is the probability density of measuring the set of feature vectors $\mathbf{X}$ for unknown U if U belongs to class Ω, whereas the quantity of interest for a Bayes decision rule is the *a posteriori* probability

$$\hat{\Omega}=\arg\max \Pr(\Omega^{(m)}|\mathbf{X}) \quad (4)$$

that a measurement of $\mathbf{X}$ means it is more probable that U is a member of a particular class $\Omega^{(m)}$ than another class. Here, $\Omega^{(m)}$ represents the mth class, $\hat{\Omega}$ is the class which maximizes this probability, and $m=1,2,\ldots,M$.

Using the argument that the four classes are equally probable and dropping terms which do not vary with class, it can be shown for the present case (Brown, 1999) that $\hat{\Omega}$ in Eq. (4) above can be expressed as

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

This equation states the results in terms of the probability density of Eq. (3), which is the quantity calculated in our

experiment. Here, $m=1,2,3,4$, and each sound in the test set is assigned to the class which maximizes the probability in this equation.

The values for the features from each frame of a particular sound from the test set were used to calculate the probability density of Eq. (3) for each of the four instrument classes. That sound was then assigned to the class for which this function was a maximum. After this was done for each of the sounds, a four-by-four confusion matrix was computed showing what percent of each of the test sounds in each of the classes was assigned to each of the four possibilities. An overall percent correct (equal to the total number of correct decisions divided by the total number of members of the test set) for this particular set of training sounds was also computed.

The training sounds (listed in Table II) and total number of clusters were then varied. Pairwise comparisons were also made with calculations identical to those described in Brown (1997, 1998a, 1998b, 1999).

B. Features

Features from both the frequency and time domains were examined; in some cases approximations to the frequency and time derivatives were calculated as well.

1. Frequency domain

Cepstral coefficients provide information about formants for speech/speaker identification in humans which translates into resonance information about musical instruments. They were calculated (O'Shaughnessy, 1987) from 22 constant-Q coefficients with frequency ratio 1.26 and frequencies ranging from 100–12 796 Hz. Channel effects were explored, where the long-term average is subtracted from each coefficient to eliminate the effects of different recording environments (Reynolds and Rose, 1995). Cepstral time derivatives (approximated by subtracting coefficients separated by four time frames) were calculated, again to eliminate effects of the recording environment. Other features derived from the spectrum were the constant-Q coefficients and their bin-to-bin differences as a measure of spectral smoothness (McAdams, Beauchamp, and Meneguzzi, 1999). Spectral centroid (the Fourier amplitude-weighted frequency average) and average energy (Beauchamp, 1982) were calculated from the Fourier transform.

2. Time domain

In addition to autocorrelation coefficients, the Dubnov *et al.* (1997) method of calculating moments of the residual of the LPC (linear prediction coefficients) filtered signal was examined along with the straightforward calculation of the third (skew), fourth (kurtosis), and fifth moments of the raw signal. Finally, the second through fifth moments of the envelope of the signal were examined by taking the Hilbert transform (Hartmann, 1998) of the signal and low-pass filtering its magnitude.

IV. RESULTS AND DISCUSSION

A. Four instruments

1. Feature dependence

Results with different sets of features are summarized in Fig. 2. The optimum choice of training sounds and clusters is indicated by ''Opt.'' The mean is the average over all training sounds and numbers of clusters, and is the accuracy obtainable with an arbitrary set of training sounds. The standard deviation is a measure of the confidence interval of the results. Note that all features except moments of the time wave gave much better identification than chance.

Feature sets and number of coefficients are indicated on the graph. The most successful feature set was the frequency derivative of the constant-Q coefficients measuring spectral smoothness (also called spectral irregularity in the human perception literature) with 84% correct. Next most successful were bin-to-bin differences (quefrency derivative) of the cepstral coefficients with 80%, even though, considering the roughly 7% standard deviation, this does not mark a significant difference from cepstral coefficients. An explanation for this slight advantage is that taking differences removes the effect of frequency-independent interference, and this gives a constant additive term for all cepstral coefficients.

Other successful features were cepstral coefficients and autocorrelation coefficients with over 75% correct. From the point of view of computational efficiency, the best choice is cepstral coefficients, since only ten were required. The cepstral transform acts as an information compaction transform with most of the variance (and hence information) in the lower coefficients.

Spectral centroid alone, i.e., a one-dimensional feature or single number per frame, was sufficient to classify the sounds with close to 50% accuracy. There is an optimum range for the number of features (10–22 for cepstra and 25–49 for autocorrelation) as has been discussed for pattern recognition calculations (Schmid, 1977; Kanal, 1974).

Unlike improvements obtained in calculations for speaker identification with the inclusion of channel effects and frame-to-frame differences in cepstral coefficients, we found no such improvement in our results. This indicates that for music, in contrast to speech, significant information is contained in the long-term average value.

That autocorrelation coefficients were successful as features is surprising since they have not been used for speaker or vowel identification, and there is no *a priori* reason to anticipate this success. Also of note is the fact that changing the sample rate from 11 to 32 kHz has little effect on the autocorrelation results, since the time range examined varies by a factor of about 3. This indicates the importance of high-frequency or formant information present in both representations.

Cepstral coefficients were combined with spectral centroid to determine whether combining features would lead to better identifications. The result was slightly poorer than that with cepstral coefficients alone, although not outside the standard deviation.

Finally, consistent with the findings of Dubnov *et al.* (1997), the average moment calculations gave results no bet-

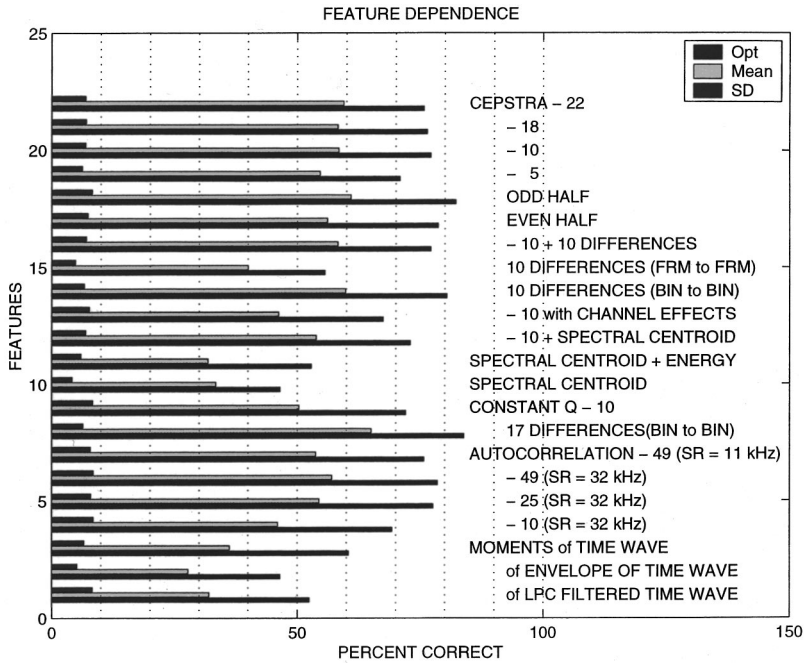

FIG. 2. Accuracy as a function of features. ''Opt'' gives the percentage correct with the optimum choice of training sounds and number of clusters for the four instruments. The mean and standard deviation were obtained by varying the training sounds and clusters.

ter than random, indicating that instruments cannot be distinguished within an instrumental family with these features.

The most successful feature sets (cepstra, constant-*Q* differences, and autocorrelation coefficients) can all be derived from the Fourier transform and in that sense can be considered as transformations of spectral information. The advantage of taking the transforms is that they decorrelate the components of the feature vector, as tacitly assumed in Eq. (1). In contrast, components of the Fourier transform are highly correlated since they are proportional to the amplitude of the original sound wave. Decorrelation occurs in taking the log for the transformation to cepstral coefficients; the amplitude information is all contained in the dc component, which is usually dropped. Similarly, with the constant-*Q* differences, the overall amplitude term is a constant additive term for each coefficient (expressed in dB) and drops out when taking the differences (Macho *et al.*, 1999).

2. Number of clusters

The maximum number of clusters was varied, with the results given in Fig. 3. They show no significant change in going from seven to ten clusters, and only 4 percent from two to ten clusters, so calculations can be carried out using seven clusters with confidence that there will be no loss of accuracy.

3. Training sounds

The results shown in Fig. 2 indicate that the choice of training sound combinations is significant in obtaining optimum results. Information on the best training sounds and corresponding number of clusters for the most successful features is collected in Table IV. The features are identified in column one, followed by the number of combinations of training sounds which gave identical results. Column three indicates the number of combinations from column two in which only the number of clusters varied, i.e., the sounds were identical. Finally, in columns four to seven, the training sounds referred to in column three are identified along with the range of cluster values of each in parentheses.

The sounds by Christ (oboe), Griffin (sax), Matzener (clarinet), and Baron (flute) were the most effective for the majority of these feature sets, indicating that a single set of training sounds is optimum for different feature sets. Analysis of these sounds shows that it is important to have many notes (rapid passages) over a wide frequency range with a reasonably smooth spectrum.

As a further test of generality of training sounds, the sounds in the test set were split arbitrarily (odd and even sample numbers) into two halves and run independently. As shown in Fig. 2, the results were similar (82% vs 79%), indicating no disparity in the two sets of data. The calculation was then carried out using the optimum training sounds for the second half on the first half and vice versa. The results on the first half changed from 82% correct with its optimum training sounds to 73% correct with the sounds from Table IV optimized for the second half. The corresponding change for the second half of the sounds was from 79% to 67%. The effect is greater than the 7% significance level, but the results are still quite good and indicate that this method is generalizable.

4. Confusion matrices

Confusion matrices were calculated for each of the feature sets and can be obtained from the author. Figure 4 is a summary of the diagonal elements (percent correct for each instrument) of the confusion matrices for the best feature sets.

For ten cepstral coefficients, the clarinet identification is poor with only 50% correct. It was confused with the sax 27% of the time, with all other confusions 12% or less. With 18 cepstral coefficients, the results on the clarinet are much better than with ten coefficients, although more confusions of

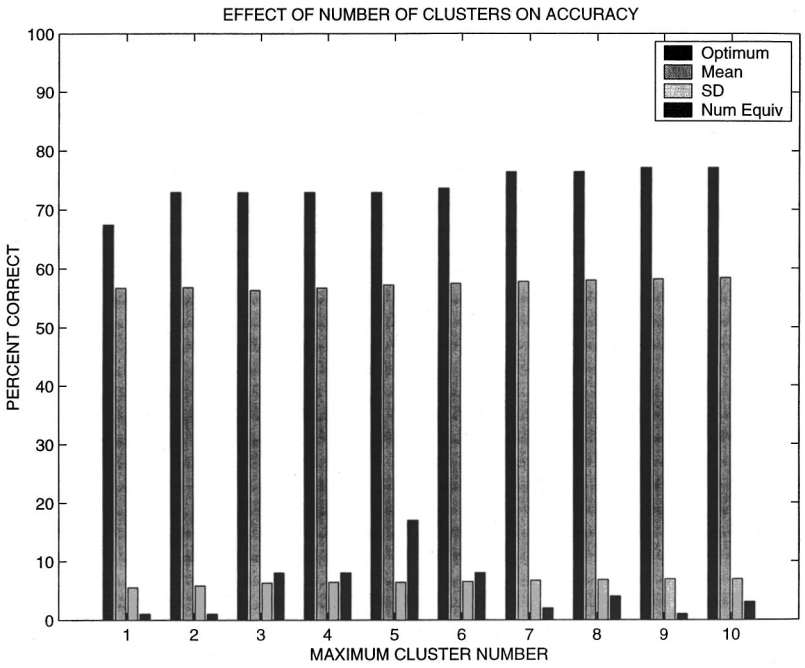

FIG. 3. Effect of varying the maximum number of clusters with ten cepstral coefficients as features. ''Optimum'' gives the percent correct for the optimum choice of training sounds and number of clusters. The mean and standard deviation are taken over all combinations of training sounds and cluster numbers up to the maximum. ''Num equiv'' is the number of combinations which gave identical optimum results.

other instruments identified as clarinet occur. Results on the oboe and flute are somewhat poorer. For better overall identifications, 18 coefficients would be preferable to ten. The largest confusions were of the flute as clarinet (26%) and the oboe as clarinet (19%). Strong and Clark (1967b) also found oboe–clarinet confusions.

Results with 25 autocorrelation coefficients were quite good overall with all identifications of instruments 70% or above. The major confusions were sax–clarinet confusions of 19% and 24%. Better overall correct identifications were found for 49 autocorrelation coefficients as seen in Fig. 4. Here, all diagonal elements are over 75%. Confusions in the range 10%–16% were found for sax as oboe, clarinet as sax, clarinet as flute, and flute as clarinet.

The results for the bin-to-bin frequency differences were of particular interest since they are directly related to the spectral smoothness studied by McAdams, Beauchamp, and Meneguzzi (1999). These are the best overall results, and unlike the others, clarinet identifications are the best. This is due to the missing even harmonics at the lower end of the spectrum, which make bin-to-bin differences distinctive, and is consistent with the results of Saldanha and Corso (1964). The oboe was identified as a flute almost 30% of the time. Other confusions were all less than 10%.

For all other feature sets, oboe and sax identifications are best overall.

B. Pairs of instruments

The sounds from the four instruments were also compared in pairs, as was done for the oboe and sax in Brown (1999). Results are given in Fig. 5, which plots percent error for each of the six pairs along with an overall percent error. As with the four-way calculations, the poorest results were obtained with spectral centroid, a single number. Again, the best results occurred with bin-to-bin differences of constant-Q coefficients as features. There, the error was only 7% overall. Confusions of the flute with each of the three

TABLE IV. Optimum choice of training sounds for different features for four instrument identification. Column one indicates the features. Column two (NW=number of winners) gives the number of combinations of training sounds and clusters which gave optimum results. Column three gives the number of identical (NI) sounds from column two in which only the number of clusters is different. The last four columns give the optimum training sound for each instrument with the range of cluster values in parentheses or simply the number if there was a single cluster value.

Features	NW	NI	Oboe	Sax	Clarinet	Flute
10 Cepstral coefficients	3	3	Christ2	Griffin(2–3)	Matzener(9–10)	Baron2
18 Cepstral coefficients	24	24	Christ(6–10)	Griffin(9–10)	Goodman10	Baron(7–9)
22 Cepstral coefficients	8	8	Christ(8–10)	Griffin(9–10)	Goodman10	Baron(5–6)
10 Cepstra—half of sounds	12	12	Christ2	Griffin(2–3)	Matzener(9–10)	Baron(2–6)
10 Cepstra—other half of sounds	4	4	Christ4	Griffin(6–7)	Goodman9	Baron(4–6)
17 Cepstral diffs (bin-to-bin)	6	6	Robinson6	Griffin(6–9)	Matzener10	Baron(4–10)
17 Constant-Q diffs (bin-to-bin)	1	1	Christ(9)	Griffin(5)	Matzener(9)	Luening(10)
49 Autoc coeffs (SR=11 kHz)	14	12	Christ(7–10)	Griffin(5,10)	Matzener(4–6)	Luening(7–10)
49 Autoc coeffs (SR=32 kHz)	2	2	Christ9	Griffin10	Matzener9	Baron9
25 Autoc coeffs	12	12	Christ(9–10)	Griffin(8–9)	Matzener(9–10)	Baron(7–8)
10 Autoc coeffs	6	3	Christ(9–10)	Griffin(7–8)	Matzener(7,10)	Baron(7–8)

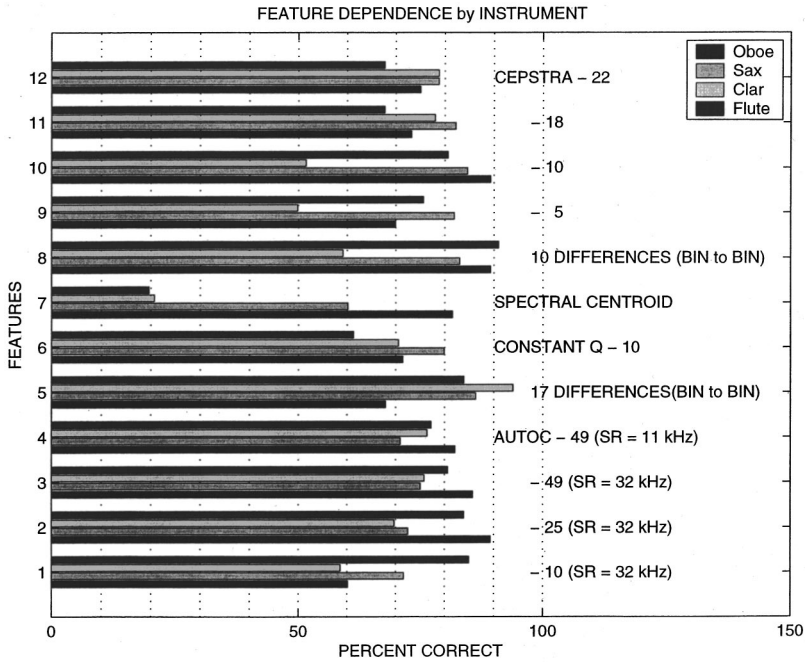

FIG. 4. Summary of correct identifications of each instrument class taken from diagonal elements of confusion matrices for feature sets indicated. Note that data are in inverse order from captions.

other instruments were highest, consistent with Berger's finding of maximum confusions for the flute as oboe and flute as sax. The clarinet was most easily identified, in agreement with Saldanha and Corso's (1964) finding.

C. Human perception experiment

None of the published human perception studies was carried out with exactly the same instruments as were used in these calculations; for the most part, they were carried out on single notes. For purposes of comparison, therefore, we conducted a free classification experiment on short solo segments of music played by the oboe, sax, clarinet, and flute. In many cases these were the same segments used for the calculations.

Fifteen musicians were asked to classify 60 sound samples into as many categories as they wished, but to make no distinction regarding the register of instrument, e.g., soprano or alto. They organized the sounds into five major groups. If four of these groups are named for the instrument with the most sounds present (one group was a mixture of several instruments), then the percent correct is given in the last row of Table I. More details on this experiment will be given in a subsequent paper (Houix, McAdams and Brown, unpublished).

Confusions were on average small, with no overall pattern. The overall percent correct for all classifications is 85%, which is close to the results for the computer calculations.

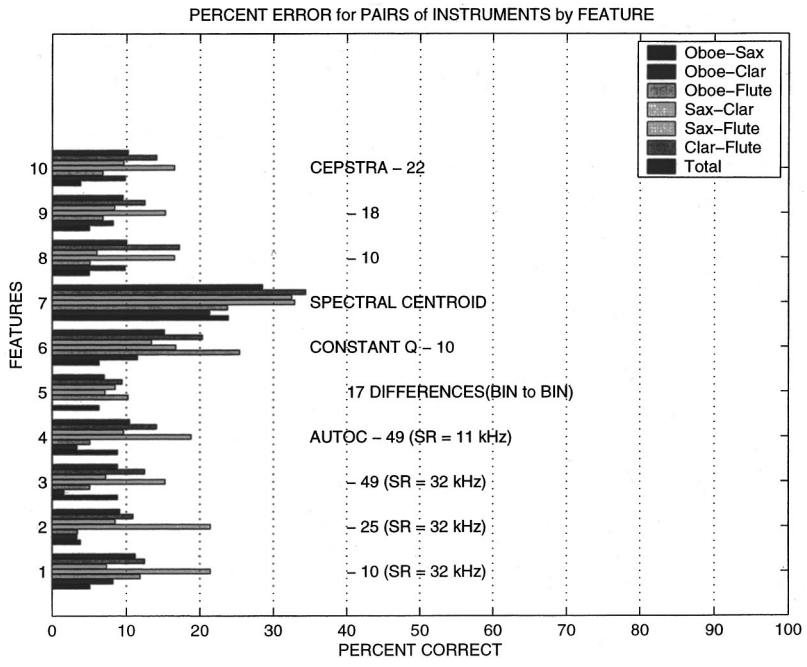

FIG. 5. Errors in identification of pairs of instruments. Instruments are given in the legend. The total represents the total number of errors divided by the total number of decisions for all pairs for a given feature set. Note that data are in inverse order from captions.

V. CONCLUSIONS

The success of cepstral coefficients (77% correct) for identification indicates that these woodwind instruments have distinct formant structures and can be categorized with the same techniques used for speaker/speech studies. Spectral smoothness (bin-to-bin differences of the constant-Q spectrum) was also effective (over 80% correct) and indicates a characteristic shape of the spectrum for sounds produced by these instruments. The success of these features is due to the property that individual components of their feature vectors are uncorrelated.

The actual numerical percentage correct for these sounds is dependent on the particular training set and number of clusters chosen. The choice of training sounds is generalizable for a randomly chosen set of test sounds with about a 10% drop in accuracy.

Most important, several sets of features can be used for computer identification of the oboe, sax, clarinet, and flute with 75%–85% accuracy. Because a much larger test set was used than in previous studies, the feature sets and methods used are applicable to arbitrary examples of these instruments. These results are as good or better than results on human perception and indicate that the computer can do as well as humans on woodwind instrument identification under the present conditions.

ACKNOWLEDGMENTS

J.C.B. is very grateful to the Marilyn Brachman Hoffman Committee of Wellesley College for a fellowship supporting this study. Part of this work was carried out during a sabbatical leave by J.C.B. tenured in the Music Perception and Cognition group at IRCAM and was made possible by Wellesley College's generous sabbatical leave policy. Finally, thanks go to Peter Cariani for suggesting the use of autocorrelation coefficients as features, and to Dan Ellis and Douglas Reynolds for valuable e-mail discussions.

APPENDIX: TERMS USED IN PATTERN RECOGNITION AND THE METHOD OF CLUSTERS

Pattern recognition—A method in which a set of unknown patterns called the *test set* is grouped into two or more *classes* by comparison to a *training set* consisting of patterns known to belong to each class.

Features—also called *feature vectors*—Properties (the patterns) calculated for the test set which are compared to the same properties of the training set for classification. In general, a feature has N associated values and can be considered an N-dimensional vector, e.g., for autocorrelation coefficients, each lag time gives one component of the vector.

Clustering—a means of summarizing the calculations on members of the training set to simplify comparison to the test set. In the calculation described in this paper, a feature vector is calculated every 16 ms for each training sound, each time contributing a point in an N-dimensional feature space. These data are summarized by grouping nearby points into *clusters* each with a mean μ, standard deviation σ, and probability p given by the number of points in that cluster divided by the total number of points for the sound.

Gaussian mixture model—A probability density function is formed as a sum of Gaussian functions obtained from the means, standard deviations, and probabilities for each cluster of a given member of the training set. This is described in more mathematical detail in Sec. III.

Beauchamp, J. W. (**1982**). ''Synthesis by spectral amplitude and brightness matching of analyzed musical instrument tones,'' J. Audio Eng. Soc. **30**, 396–406.

Berger, K. W. (**1964**). ''Some factors in the recognition of timbre,'' J. Acoust. Soc. Am. **36**, 1888–1891.

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

Brown, J. C. (**1998a**). ''Computer identification of wind instruments using cepstral coefficients,'' J. Acoust. Soc. Am. **103**, 1889–1890(A).

Brown, J. C. (**1998b**). ''Musical instrument identification using autocorrelation coefficients,'' Proceedings of the International Symposium on Musical Acoustics 1998, Leavenworth, Washington, pp. 291–295.

Brown, J. C. (**1999**). ''Computer identification of musical instruments using pattern recognition with cepstral coefficients as features,'' J. Acoust. Soc. Am. **105**, 1933–1941.

Campbell, W. C., and Heller, J. J. (**1978**). ''The contribution of the legato transient to instrument identification,'' in Proceedings of the Research Symposium on the Psychology and Acoustics of Music, edited by E. P. Asmus, Jr. (University of Kansas, Lawrence, KS), pp. 30–44.

Clark, M., and Milner, P. (**1964**). ''Dependence of timbre on the tonal loudness produced by musical instruments,'' J. Audio Eng. Soc. **12**, 28–31.

Dubnov, S., and Rodet, X. (**1998**). ''Timbre recognition with combined stationary and temporal features,'' Proceedings of the International Computer Music Conference, Los Angeles.

Dubnov, S., Tishby, N., and Cohen, D. (**1997**). ''Polyspectra as measures of sound texture and timbre,'' J. New Music Res. **26**, 277–314.

Eagleson, H. V., and Eagleson, O. W. (**1947**). ''Identification of musical instruments when heard directly and over a public-address system,'' J. Acoust. Soc. Am. **19**, 338–342.

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 Irene Deliege and John Sloboda (Psychology, East Essex, UK), pp. 253–307.

Handel, S. (**1995**). ''Timbre perception and auditory object identification,'' in *Hearing*, edited by B. C. J. Moore (Academic, New York).

Hartmann, W. M. (**1998**). *Signals, Sound, and Sensations* (Springer, New York, Secaucus, NJ).

Houix, O., McAdams, S., and Brown, J. C. (unpublished).

Kanal, L. (**1974**). ''Patterns in pattern recognition 1968–1974,'' IEEE Trans. Inf. Theory **IT-206**, 697–722.

Kendall, R. A. (**1986**). ''The role of acoustic signal partitions in listener categorization of musical phrases,'' Music Percept. **4**, 185–214.

Macho, D., Nadeu, C., Janovic, P., Rozinaj, G., and Hernando, J. (**1999**). ''Comparison on time and frequency filtering and cepstral-time matrix approaches in ASR,'' Proceedings of Eurospeech '99, Vol. 1, pp. 77–80.

Marques, J. (**1999**). ''An automatic annotation system for audio data containing music,'' Master's thesis, MIT, Cambridge, MA.

Martin, K. D. (**1999**). ''Sound-source recognition: A theory and computational model,'' Ph.D. thesis, Massachussetts Institute of Technology, Cambridge, MA.

McAdams, S. (**1993**). ''Recognition of Auditory Sound Sources and Events,'' in *Thinking in Sound: The Cognitive Psychology of Human Audition*, edited by S. McAdams and E. Bigand (Oxford University Press, Oxford).

McAdams, S., Beauchamp, J. W., and Meneguzzi, S. (**1999**). ''Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters,'' J. Acoust. Soc. Am. **105**, 882–897.

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).

Rabiner, L. R., and Huang, B.-H. (**1993**). *Fundamentals of Speech Recognition* (Prentice Hall, Englewood Cliffs, NJ).

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.
Saldanha, E. L., and Corso, J. F. (**1964**). ''Timbre cues and the identification of musical instruments,'' J. Acoust. Soc. Am. **36**, 2021–2026.
Schmid, C. E. (**1977**). ''Acoustic Pattern Recognition of Musical Instruments,'' Ph.D. thesis, University of Washington.
Strong, W., and Clark, M. (**1967a**). ''Perturbations of synthetic orchestral wind-instrument tones,'' J. Acoust. Soc. Am. **41**, 277–285.
Strong, W., and Clark, M. (**1967b**). ''Synthesis of wind-instrument tones,'' J. Acoust. Soc. Am. **41**, 39–52.
Wilcox, L., Kimber, D., and Chen, F. (**1994**). ''Audio indexing using speaker identification,'' ISTL Technical Report No. ISTL-QCA-1994-05-04.