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. @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 cp.uantitati8iontern coeexcer7n9.3(no)-3ltnMarqueiuou .466.2(excer7n9.eures

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

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.

III. CALCULATIONS

A. Probability calculation

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

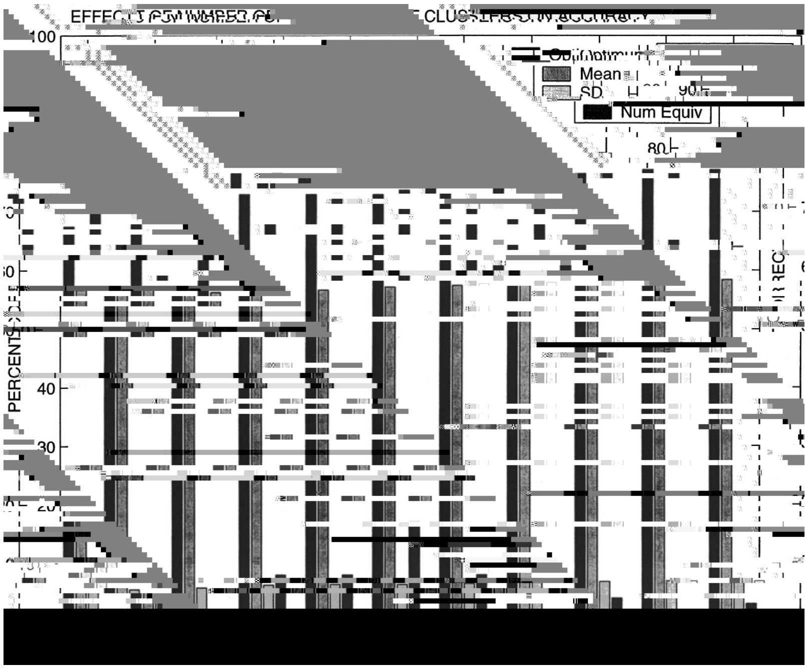

FIG. 3. Effect of varying the maximum number of clusters with ten cepstral coefficients as features. ''Optimum'' gives the percent correct for the optimum choice of training sounds and number of clusters. The mean and standard deviation are taken over all combinations of training sounds and cluster numbers 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!	Goodman9different features for fo/F13 1 Tf 1.6209 0 TD [(Goodman9)-2	

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 m, standard deviation s, and probability p

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.