LINEAR PREDICTIVE MODELS FOR MUSICAL INSTRUMENT IDENTIFICATION

Nicolas Chétry and Mark Sandler

Centre for Digital Music
Queen Mary, University of London
Mile End Road
E1 4NS London

ABSTRACT

This paper deals with musical instrument identification. The proposed method consists of building the instrument models using a set of linear predictive coefficients, the Line Spectrum Frequencies. The models consist of characteristic short-term spectral envelopes calculated for each instrument in the database. The identification process involves the calculation of a similarity measure between two codebooks, one taken from the models database, one corresponding to the sample to identify. Next, the use of Support Vector Machines as classifier is investigated. The two systems are then applied to the identification of monophonic phrases extracted from commercial recordings. It is shown that good performance can be achieved for the classification of one unknown excerpt amongst 6 instruments.

1. INTRODUCTION

The representation and modelling of sound textures is an essential aspect in the design of automated musical instrument identification systems. Within the Musical Information Retrieval (MIR) framework, it is desirable to automatically evaluate the similarities or differences between two instruments or two families of instruments.

Timbre is an important property of sound. Thanks to the timbre, humans are able to recognise quasi-instantaneously familiar voices over the telephone or can pick out different instruments even if they are playing notes at the same pitch and loudness. Indeed, musical instrument identification systems attempt to model this faculty of differentiating between two sounds by extracting salient properties from the signals.

Research in the field concentrates on the following techniques. Firstly, there are *timbre modelling* approaches. These are descriptor-based algorithms which include a feature extraction module. This stage is concerned with the choice and calculation of relevant features used to model the timbre of a sound which are then fed into a machine learning algorithm to obtain a condensed and representative model of each instrument. Mono-feature and multi-feature systems are also used. The mono-feature approach originates from research in speech / speaker recognition and can be understood by the fact that sounds are produced by identical organs. On the other hand, the mechanisms of sound production by musical instruments are multiple and involve different physical principles. The resulting signals, although being and sounding musical (i.e. made of harmonics in our case) can exhibit peculiar properties that a learning machine should take into account. Examples of such approach can be found in [1] or [2]. The current trend in the research is clearly oriented towards multi-features systems including an automatic feature selection stage [3], [4].

Secondly, *instrument modelling* approaches put the emphasis on the learning of the mechanisms of sound production by musical instruments. Starting from a mathematical assumption about the signal content, the process consists of adapting this model to a training set, evaluating the relevant parameters during a training stage and using it to identify new excerpts. As an example, a log-power spectrum plus noise model in an independent subspace analysis framework has been used in [5].

Finally, *mixed models* combine the two approaches described above. On one hand, a prior is set on the mechanisms of sound production or on the signal composition, whereas on the other hand, features are extracted and used to build the instrument models. As an example, the use of synchronous and asynchronous deviations of the phase of the partials for instrument identification purposes is investigated in [6]. It is suggested such features may help to distinguish between instruments.

The technique presented here belongs to the last category. To a certain extent, the linear predictive framework is transposed here to model the mechanisms of sound production by musical instruments. More specifically, the deconvolution between the contributions of the excitation (source) and the instrument body (filter) is investigated through the consideration of a particular set of linear predictive coefficients, the Line Spectrum Frequencies (LSF). The use of two types of classifiers is proposed for building the instrument models. The first one consists of learning characteristic spectral shapes for each instrument using a K-means algorithm. The second one, using Support Vector Machines (SVM), consists of finding boundaries between the feature data distribution of each instrument and the others in the database. The emphasis in this paper is on the application of the technique for identifying monophonic excerpts extracted from commercial recordings.

In section 2, we describe the principle of the system that has been designed to identify isolated notes. Next, Support Vector Machines which are considered in building the instrument models are introduced in section 3. Experiments involving databases of isolated notes and solo phrases have been conducted to evaluate the systems performance. The corresponding results are summarised in section 4. Finally, a discussion about the pitch dependence in musical instrument identification systems closes the paper.

2. BASE SYSTEM DESCRIPTION

In this section, details about the system designed to classify isolated note recordings that has been described in [7] are recalled.

N. Chétry is supported by the Semantic Interaction with Music Audio Contents (SIMAC) project (EU-FP6-IST-507142) and by the Department of Electronic Engineering at Queen Mary, University of London.

1-4244-0469-X/06/$20.00

2.1. Features

In a source-filter configuration, the short segment of a signal is assumed to be generated as the output of an all-poles filter $H(z) = 1/A(z)$, where $A(z)$ is the inverse filter given by

$$A(z) = 1 + a_1 z^{-1} + ... + a_p z^{-p}, \quad (1)$$

p is the order of the LPC analysis and $\{a_i\}, i = 1, ..., p$ the filter coefficients.

Introduced by Itakura in [8], the Line Spectrum Pairs (LSP) are the roots of two polynomials $P(z)$ and $Q(z)$ defined as

$$\begin{cases} P(z) = A(z) + z^{-p+1}A(z^{-1}) \\ Q(z) = A(z) - z^{-p+1}A(z^{-1}) \end{cases} \quad (2)$$

It can be shown that they are interlaced and lay on the unit circle. Their corresponding angular frequencies are called the Line Spectrum Frequencies. An interesting property of the LSF is that they are relatively robust to quantisation noise. In other words, a slight variation of one coefficient value only affects the spectral envelope around its angular frequency. This property is exploited here for the calculation of characteristic spectral shapes of the instruments using the K-means algorithm.

Recent research works highlighted the importance of temporal properties (e.g. onsets and transients location [9]) in the identification of sounds by machines. As the uniqueness of the spectral envelope cannot be absolutely guaranteed across instruments, extra information about the signal temporal behaviour are often needed in order to increase the systems performance. Features such as the attack and decay time or the duration of transient can be considered when building instrument identification systems.

However, the main difficulty in extracting such temporal features resides in the fact that robust automated pre-processing techniques for onsets or transients detection are difficult to design, especially in the case of pitched musical sounds. For this reason, a more general approach is preferred. It consists of appending the delta (speed) and delta-deltas (acceleration) coefficients to the feature vector in order to include information about its evolution with time. The augmentation of the LSF feature vectors with their delta (first derivative as a function of time) as it is commonly performed with cepstral coefficients will be experimented in section 4.

2.2. Instrument modelling

In the following, it is assumed that the training data set consists of T observations $X = \{\vec{x}_t\}_{t=1,...,T}$ of LSF coefficients extracted from the training data set of one instrument. By performing a K-means on the training data set, X is represented by a codebook $C = \{\vec{c}_k\}_{k=1,...,K}$ of K codevectors. The algorithm implementation described in [10] has been used. In practice, the initial codebook is obtained by the splitting technique. Starting with one codeword (the mean of the whole data set), each vector in the dictionary is iteratively split into two vectors until the largest smaller power of two of K is reached. If needed, the maximally populated cluster is then split into two new clusters until the desired number of centroids is attained. The two-stage procedure described above is then iterated until the average total distortion reaches a local minimum.

In figure 1 are represented 16 characteristic spectral shapes calculated in this way for four instruments, the clarinet, the piano, the cello and the oboe. The feature data set consists of 24 LSF extracted from 2 minutes recordings. It can be noticed that most of the spectra are globally low-pass, except for the piano. Due to the transient nature of a piano attack, the frequency responses exhibit a consequent amount of energy in the high frequencies. Although not being sufficient as such for a perfect discrimination between the other classes of instruments, the transient nature of the piano sound is nevertheless carried by the model.

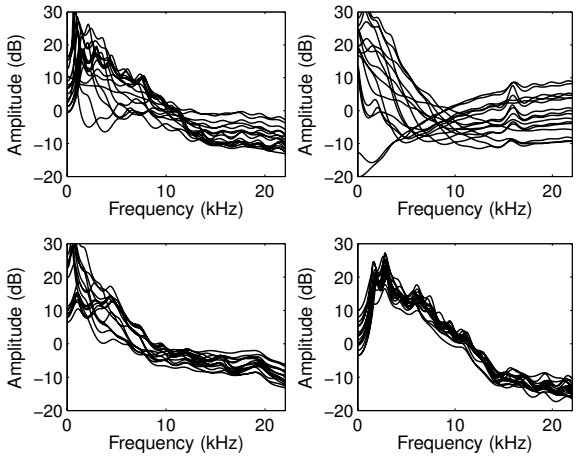


Fig. 1. Frequency responses of 16 characteristic short-term spectral shapes corresponding to an optimised codebook of LSF vectors. The codevectors have been obtained using two minutes of solo phrases. Instruments are (clockwise from top left): clarinet, piano, cello and oboe.

2.3. Instrument identification

In the following, we will assume that the instrument to identify is represented by a sequence $Y = \{\vec{y}_s\}_{s=1,...,S}$ of LSF coefficients and that the N instruments in the database are represented by their codebooks $C_n, n = 1, ..., N$ of K codewords. During the identification phase, a K-means is used to build a codebook $\tilde{C} = \{\vec{\tilde{c}}_k\}_{k=1,...,K}$ which models the observation Y. The idea is to fully take advantage of the available observation prior to the classification. The similarity measure between two codebooks, $\tilde{C}$ and one codebook C_n taken from the models in the database is defined as:

$$d_{\tilde{C},C_n} = \sum_{k=1}^{K} \left[\min_{1 \le k \le K} d(\vec{\tilde{c}}_k, \vec{c}_{n,k}) \right], \quad (3)$$

in which d is the Euclidean distance. The identity of the unknown codebook is retrieved by finding n^* such that:

$$n^* = \arg \min_{1 \le n \le N} d_{\tilde{C},C_n} \quad (4)$$

3. SUPPORT VECTOR MACHINES

Support Vector Machines are becoming increasingly popular for data classification problems. Their use in an instrument identification context has been studied in [2] and [11].

3.1. Principle

In the binary decision case and given a labelled training set, the principle of the SVM is to find the optimum hyper-plane (i.e. the one

that maximises the margin) separating the two data sets. In the case of non-linearly separable data, a mapping between data space and a higher dimensional feature space is performed using a kernel function. A *one-against-one* approach [12] can be used to extend the algorithm for multi-class classification purposes.

A radial basis function of the form

$$k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma||\mathbf{x} - \mathbf{y}||^{\mathbf{2}}), \quad \gamma > \mathbf{0} \tag{5}$$

has been chosen for the experiments. γ is set to $1/N$, where N is the number of instruments in the database and C is empirically determined using a trial-error procedure on the training data set. Prior to the SVM optimisation, data are normalised to lay in the range [-1,+1].

During the identification stage, each individual frame $\vec{y}_s$ of Y is tested against the model which returns a possible identity. The final identity of the observation Y is the one that has been the most often retrieved over the S tested frames.

4. EXPERIMENTS AND RESULTS

The purpose of the experiments is to evaluate the systems performance for the identification of musical phrases in a realistic context. Two sets of experiments have been conducted. The first one focuses on the use of melodic phrases for both training and testing. In this case, models have been trained using 50% of the available data while the other 50% were retained for testing. The second one investigates the system robustness when isolated notes from a different database are used to train the models. In both cases, segments of 5s in duration are presented to the classifiers which returns an identity.

4.1. Databases

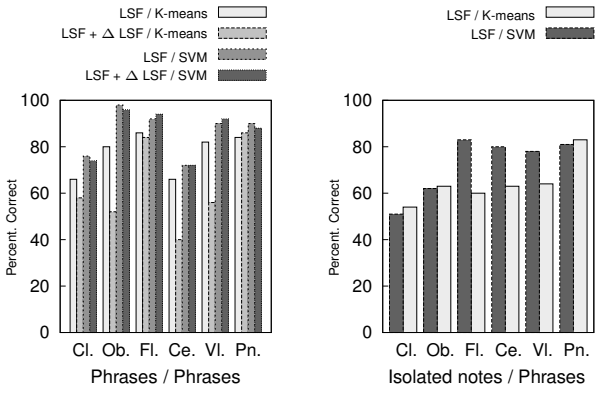
The first database consists of melodic phrases extracted from commercial recordings. The following instruments are considered: the clarinet (Cl.), the oboe (Ob.), the flute (Fl.), the cello (Ce.), the violin (Vl.) and the piano (Pn.). Each data set contains 300s of sounds originating from 10 different sources. The second database consists of isolated notes taken from the RWC [13] collection. For each class, 3 instances of each note corresponding to various instrument brands have been retained, thus totalling an average of 250 notes per instrument.

4.2. Feature extraction

Audio excerpts are resampled at 22050 Hz prior to any processing. Silence and very low level segments are firstly discarded and the DC bias is removed. Amplitudes are then normalised to the 0 dB level and a pre-emphasis is applied. This step is particularly useful when LPC-based models are used as it helps the algorithm to better pick the envelope high frequency structure. Next, features are extracted every 34.5ms within frames of 46.43ms in duration that have been previously weighted by a Hanning window. For each frame, 16 LSF are derived from the linear predictive coefficients using the algorithm described in [14].

4.3. Results

The results are summarised in figure 2. Comparative performances of the two classifiers are presented in figure 2(a) where solo phrases are used for both training and testing. The base system described in section 2, using a K-means and 40 codevectors, is able to correctly identify 78% of the testing samples, almost 15% less than



(a) (b)

	Cl.	Ob.	Fl.	Ce.	Vl.	Pn.
Cl.	**76**%		8%	2%	12%	2%
Ob.		**98**%	2%			
Fl.	2%		**92**%			6%
Ce.	2%		6%	**72**%	16%	4%
Vl.	8%			2%	**90**%	
Pn.	8%		2%			**90**%

(c)

Fig. 2. (a) Individual correct identification rates for the 2 systems using melodic phrases for both training and testing, (b) Comparative individual correct identification rates for the two systems when isolated notes are used for training and phrases for testing, (c) confusion matrix for a system using a SVM in the configuration in which it performs the best (86.3% average correct identification).

when an isolated notes database was used [7]. This can be explained by the fact that in previous experiments, although training and testing samples were different, some of them came from the same database where there is a consistent and *dry* acoustic. In the current configuration, all the training and testing excerpts were extracted from different recordings. Individual correct identification rates range from 66% (cello) to 86% (flute). It has been found during the experiments that the most important confusions involved the pairs oboe-flute (14%) and piano-clarinet (10%). In terms of instrument families identification, bowed strings (cello and violin) are recognised 88% of the time and reeds (oboe and clarinet) 77% of the time. The global performance of this system is comparable to the one described in [1] where 77% average correct identification has been achieved using ten cepstral coefficients and a Gaussian mixture model. Finally, the addition of the delta coefficients to the feature vectors seriously degraded the performance which dropped by more than 15% down to 62.5%.

Using a Support Vector Machine allows 86.3% of the testing samples to be correctly recognised, an improvement of 8% compared to the previous system. The corresponding confusion matrix is shown in figure 2(c). The best individual correct identification rates are achieved for the oboe class (98%), followed by the flute class (92%). Again, among the 6 instruments in the database, the cello samples are the least correctly identified (72%). Most important confusions involve the pairs cello-violin (16% of the cello samples were recognised as being violin) and clarinet-violin (12%). Finally,

bowed strings and reeds families are identified 90% and 87% of the time respectively. It can also be noticed that appending the delta did not affect the average percentage of correct identification (86%). The overall performance is comparable to the systems described in [5] and [15] where respectively 90% and 84% correct identification was achieved for similar sets of instruments.

In figure 2(b) is summarised the performance when isolated notes are used for training. In this configuration, the correct average identification rates dropped to 72.5% and 64.5% for the SVM and K-means classifiers respectively. For both systems, only half of the clarinet excerpts are correctly identified and the piano samples are correctly classified roughly 80% of the time. Using a SVM helps to increase the identification rates for the flute, cello, and violin classes by 10%, 8% and 6% respectively compared to the K-means.

5. DISCUSSION

Musical instruments have a *natural* frequency range that can be defined by the range of notes that are played in *realistic* conditions. This obviously depends on the style of music being played. Training the models using musical phrases inherently fit the model for its purpose. That is: recognising an instrument that will be played within the same frequency range and with similar variations.

A perfect timbre model should in theory be able to *adapt* to the pitch. This is what cross-databases isolated notes/phrases experiments attempt to evaluate. However, when asking to professional musicians to recognise isolated notes out of musical context, confusions between a high pitched violin note and a high pitched oboe sound, amongst others, are not rare. Further, a consequent amount of the timbral information is contained in the waveforms variations and changes with time. By training the models using isolated notes, the essential information present in musical phrases is not taken into account, thus resulting in lower identification rates during the experiments.

These observations lead to the following concept of pitch abstraction or dependence in musical instrument identification system design. The spectral envelopes of a musical instrument sound – as have been used here – are related in some ways to the pitch, and in a similar manner, timbre is related to the pitch. On one hand, an ideal system should model the timbre independently of the pitch whereas on the other hand, another model should be able to quantify the timbral variation as a function of the pitch.

The challenge in timbre modelling is to capture salient properties characterising the tone colour of a sound. Whereas multi-descriptors based approaches including automatic feature selection algorithms can partially achieve this goal, the *a-posteriori* interpretation of the results, that is trying to understand why a particular set of features is performing better than others is often arduous. Direction of future research is deeply oriented towards the retrieval of the physical mechanisms of sound production from the signal observation. To this end, and in the same vein as the technique presented here, systems consisting of extracting features characterising both the spectral envelope and the residual signal obtained after a linear predictive analysis stage are currently being studied.

6. CONCLUSION

The performance of two systems for identifying musical instruments using the Line Spectrum Frequencies as unique features has been evaluated. We have focused on the classification of solo musical phrases taken from commercial recordings and shown how well our approach works under semi-arbitrary acoustic environments. Average correct identification rates of 78%, using the characteristic short-term spectral envelopes based models, and 86% using the SVM, can be achieved for the classification of one unknown excerpt amongst 6 instruments. Finally, it has been shown that for recognising solo phrases, better performance is achieved if models are trained using solo phrases that have been recorded in various acoustic conditions.

7. REFERENCES

[1] J. C. Brown, O. Houix, and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instruments," *J. Acoust. Soc. Amer.*, vol. 109, no. 3, pp. 1064–1072, 2001.

[2] J. Marques and P. J. Moreno, "A study of musical instrument classification using gaussian mixtures models and support vector machines," *Compaq Cambridge Research Laboratory*, vol. Tech. Report 99-4, 1999.

[3] G. Peeters, "Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization," *Proc. AES 115th Convention*, 2003.

[4] S. Essid, G. Richard, and B. David, "Musical instrument recognition based on class pairwise feature selection," *Proc. ISMIR*, 2004.

[5] E. Vincent and X. Rodet, "Instrument identification in solo and ensemble music using independent subspace analysis," *Proc. ISMIR*, 2004.

[6] S. Dubnov and X. Rodet, "Investigation of phase coupling phenomena in sustained portion of musical instruments sound," *J. Acoust. Soc. Amer.*, vol. 112, no. 6, December 2002.

[7] N. Chétry, M. Davies, and M. Sandler, "Musical instrument identification using LSF and K-means," *Proc. AES 118th Convention*, 2005.

[8] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Amer.*, vol. 57, pp. S35, 1975.

[9] S. Essid, P. Leveau, G. Richard, L. Daudet, and B. David, "On the usefulness of differentiated transient/steady-state processing in machine recognition of musical instruments," *Proc. AES 118th Convention*, 2005.

[10] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Communications*, vol. 28, pp. 702–710, 1980.

[11] O. Gillet and G. Richard, "Automatic transcription of drum loops," *Proc. ICASSP*, 2004.

[12] C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[13] RWC Music Database, "Music genre database and musical instrument sound database," `http://staff.aist.go.jp/m.goto/RWC-MDB`.

[14] P. Kabal and R.P. Ramachandran, "The computation of line spectral frequencies using Chebyshev polynomials," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 6, pp. 1419–1426, 1986.

[15] J. Eggink and G. J. Brown, "Instrument recognition in accompanied sonatas and concertos," *Proc. ICASSP*, 2004.