ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/2986289

Musical Instrument Classification and Duet Analysis Employing Music Information Retrieval Techniques

Article *in* Proceedings of the IEEE · May 2004

DOI: 10.1109/JPROC.2004.825903 · Source: IEEE Xplore

CITATIONS
68

READS
313

1 author:

Bozena Kostek
Gdansk University of Technology

276 PUBLICATIONS **1,313** CITATIONS

SEE PROFILE

All content following this page was uploaded by Bozena Kostek on 17 September 2014.

The user has requested enhancement of the downloaded file.

Musical Instrument Classification and Duet Analysis Employing Music Information Retrieval Techniques

BOZENA KOSTEK, MEMBER, IEEE

Invited Paper

The aim of this paper is to present solutions related to identifying musical data. These are discussed mainly on the basis of experiments carried out at the Multimedia Systems Department, Gdansk University of Technology, Gdansk, Poland. The topics presented in this paper include automatic recognition of musical instruments and separation of duet sounds. The classification process is shown as a three-layer process consisting of pitch extraction, parametrization, and pattern recognition. These three stages are discussed on the basis of experimental examples. Artificial neural networks (ANNs) are employed as a decision system and they are trained with a set of feature vectors (FVs) extracted from musical sounds recorded at the Multimedia Systems Department. The frequency envelope distribution (FED) algorithm is presented, which was introduced to musical duet separation. For the purpose of checking the efficiency of the FED algorithm, ANNs are also used. They are tested on FVs derived from musical sounds after the separation process is performed. The experimental results are shown and discussed.

Keywords*—*Duet separation, MPEG-7, music information retrieval (MIR), musical content processing, musical data management.

I. Introduction

There are many problems related to the management of musical data that have not yet been solved. These are now being extensively considered in the field of music information retrieval (MIR) [1], [2]. Topics that should be included within the scope of this discussion include the problem of automatically classifying musical instrument sounds and musical phrases/styles, music representation and indexing, estimating similarity of music using both perceptual and musicology criteria, problems of recognizing music using audio or semantic description, building up musical databases, evaluation of MIR systems, intellectual property right issues, user interfaces, issues related to musical styles and genres, language modeling for music, user needs and expectations, auditory scene analysis, gesture control over musical works, and others. Some of topics contained within the notion of MIR are covered by the MPEG-7 standard [3]–[5], which defines description of the multimedia content in order to support better interpretation of information. It should be stressed that solving these problems needs human assistance and management. These issues will be explained further on.

One of the aims of this paper is to show examples of MIR domain applications, namely, automatic recognition of musical instruments and separation of musical duets. Together they resulted from the experiments conducted for several years in the Multimedia Systems Department, Gdansk University of Technology, Gdansk, Poland. The first mentioned process consists of several stages, such as preprocessing, feature extraction, and classification. The preprocessing stage is needed to estimate musical sound pitch. The feature extraction changes redundant information contained in the signal into descriptors. The main stage is the classification process, which can be executed based on various techniques, both statistical and soft computing.

For the purpose of the study conducted, a database of musical instrument sounds engineered earlier at the Multimedia Systems Department was adopted. This database is a multimedia application containing information on musical instruments, along with sound analyses. It encompasses sound samples of the whole musical scale of individual instruments and exemplary musical phrases played by a given instrument, time- and frequency-domain representation of sounds, tables containing sound parameter values, program

Manuscript received February 4, 2003; revised November 10, 2003. This work is supported in part by the Committee for Scientific Research, Warsaw, Poland, under Grant 4 T11D 014 22 and in part by the Foundation for Polish Science.

The author is with the Department of Multimedia Systems, Faculty of Electronics Telecommunications and Informatics, Gdansk University of Technology, Gdansk 80-952, Poland (e-mail: bozenka@sound.eti.pg.gda.pl).

Digital Object Identifier 10.1109/JPROC.2004.825903

0018-9219/04$20.00 © 2004 IEEE

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

help, the possibility of creating various kinds of charts and printed reports for all parameters, Structured Query Language (SQL)-based query, descriptive information concerning the SQL-based help, and selected HTML link pages.

The separation of musical duets requires a more complex approach in comparison to the recognition of musical instrument classes based on monophonic sound samples [6], [7]. The proposed technique for the decomposition of duet sounds is based on the modified frequency envelope distribution (FED) analysis. The recently introduced FED algorithm decomposes signal into linear expansion of waveforms, called envelope modulated oscillations (EMOs), providing a combination of complex exponential signals modulated by complex amplitude envelopes. These waveforms are chosen to best match the harmonic parts of the signal; however, nonharmonic structures can also be represented by EMOs. The first step of the engineered algorithm is the estimation of the fundamental frequency of the lower pitched instrument. Pitch estimation is carried out in block processing. The input signal is divided into short overlapping blocks, and pitch is estimated for each block separately, to deliver the pitch contour signal (PCS). Then harmonics of the second sound are searched for in the residual signal. Therefore, in this approach based on the FED algorithm, the multipitch detection is not needed. Results of the performed experiments are shown in the following sections.

The experiments described in the paper show only some of the applications related to management of musical data. It would be very valuable if the reader of this paper would refer to the rich literature related to these topics, examples of which are given in [8]–[29]. More detailed description of some of the topics mentioned is available through the cited author and her team's papers.

II. Expert Knowledge in MIR Systems

In the theory of information introduced by Shannon, information is discussed in terms of quantity. It is assumed that information reduces uncertainty on the basis of knowledge of *a priori* probabilities. This theory serves well when dealing with well-specified problems of data transmission and communication; however, it can be observed that information is not a notion of content and furthermore of knowledge [30]. In the knowledge discovery and data management, human assistance is needed. Decision systems may produce numerous rules generated in the mining process. This makes it necessary to provide for postprocessing the generated rules. Another problem which needs attention is processing unknown or missing attribute values when acquiring knowledge from databases. Real data usually contain a certain percentage of missing values. Even the choice of attributes is not an arbitrary process. When preparing a format description, either a numerical or a categorical one, it is done on the basis of understanding of the problem domain. Information retrieval systems are presupposed to do exact matches of documents involving the same cues to the user query. However, operations which are behind the query do not always provide good responses to the user's interest. To improve information retrieval quality, various strategies were proposed and used, such as probabilistic, vector space, clustering, and intelligent retrieval. The latter technique often uses concept analysis requiring semantic calculations.

The MPEG-7 standard refers to metadata information contained in the Internet archives. This notion is very often applied to the value-added information created to describe and track objects, and to allow access to those information objects [31]. In this context, descriptors that are well defined allow for better computing, and improved user interfacing and data management. In the context of the MPEG-7 standard, higher level information is defined as textual information on audio such as titles of songs, signers' names, composers' names, duration of music excerpt, etc. One should keep in mind the fact that music can be described in a number of ways, and the musical sounds include polyphonic sounds and human voice sounds (speech and singing). A musical signal, music, scores (graphical form), MIDI code, or verbal description each comes as a different representation. Provided within the MPEG-7 standard are also low-level descriptors for musical data, organized in groups of parameters such as timbral temporal, basic spectral, basic, timbral spectral, spectral basis, and signal parameters [3]. The so-called audio framework that contains all these parameter groups includes 17 vector and scalar quantities. They represent log(attack time), temporal centroid, audio spectrum envelope, audio spectrum centroid, audio spectrum spread, audio spectrum flatness, audio waveform and power, harmonic spectral centroid, harmonic spectral deviation, harmonic spectral spread, harmonic spectral variation, spectral centroid, audio spectrum basis, audio spectrum projection, audio harmonicity, and audio fundamental frequency [3], [4]. These low-level descriptors provide information for higher-level application, namely, sound recognition, musical instrument timbre similarity, melody and melodic contour recognition, robust audio matching, and spoken content recognition. It can easily be observed that these low-level descriptors are more data oriented than human oriented. This is because the idea behind this standard is to have data defined and linked in such a way as to be able to use it for more effective automatic discovery, integration, and reuse in various applications. The most ambitious task is, however, to provide seamless meaning to low- and high-level descriptors. In such a way data can be processed and shared by both systems and people.

It seems there exists a way to change primitives into higher abstraction level, namely, semantics. Assessing timbre or quality of musical instrument sounds, humans use criteria that are rarely quantitative but most often qualitative. Therefore, there is a need to find methods that make it possible to find a relationship between objectively extracted information from sound and subjective notions of timbre. Especially interesting seems a "computing with words" concept introduced by Zadeh [32], [33], which refers to the fact that humans employ words in computing and reasoning, arriving at conclusions expressed as words from premises expressed in a natural language. Computing with words can be a necessity when the available information is too imprecise to justify the use of numbers and can be justified when it is in a better rapport with reality. It seems that this

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

new paradigm of computing can be used with success in MIR by offering better processing of subjective descriptors of musical instrument sounds and enabling the analysis of data that would result in a new way of describing musical instrument sounds. An example of such processing was recently introduced by the author [34]. It was proposed that categorical notions would be quantities partitioned using fuzzy logic.

One can name such parameters both subjective and objective ones such as pitch (frequency in hertz or barks), brightness (spectral centroid), tone/noise-like quality (spectral flatness measure), attack asymmetry (skewness), overshoot or inharmonicity (log ratio of the first harmonic to second harmonic or, more generally, higher frequency harmonics to the fundamental frequency ratio), vibrato (periodic fluctuation of pitch), tremolo (periodic change of sound level), nasality (formants' position if they exist), synchronicity (delay of higher harmonics with relation to the fundamental during the attack), etc., that have double interpretation. The relationship between the objectively measured parameters of musical instrument sounds and their subjective quality can be assessed by listeners in subjective listening tests. The subject's task is to assign ranges to such parameters and to associate presented stimuli with a set of semantic scales. In further analysis, in order to secure better generalization properties, it was thought that processing should be based on learning algorithms.

It may be observed that musical object classification using learning algorithms mimics the way of human reasoning. These algorithms are especially valuable in domains in which there is a problem of imprecision and a need of knowledge mining; thus, they are a way to handle uncertainties in musical data. Such algorithms often need human supervisory control. This remark refers both to rule-based systems and neural networks (NNs) in which an expert controls the algorithm settings and the choice of feature vectors (FVs). The approach mentioned is still at the experimental stage [34]; thus, here this idea was only roughly presented.

III. Automatic Classification of Musical Instruments

Automatic classification of musical instruments can generally be viewed as a three-layer process. The first layer consists in preprocessing, which may be identified as pitch extraction. The subsequent stage of preprocessing provides information on frequency components. This information will then be used in the parametrization, which presents the second layer in the automatic classification process. Parametrization provides data on musical sounds in the form of FVs. Fundamental frequency and its harmonics allow for calculation of numerous parameters. The FVs are then used in the pattern recognition, which is the last layer in the automatic classification flow.

A. *Preprocessing*

Since the starting point of the classification process is pitch detection of a musical sound, some methods known from the literature were first reviewed and then implemented. Pitch detection is one of the most difficult tasks in speech and musical signals processing and has been studied in many publications for many years. It is due to the fact that acoustic signals are nonstationary, i.e., their pitch and amplitudes of harmonics are varying in time. In many cases significant noise disturbances are contaminating analyzed signals, making pitch estimation even more difficult. Due to these facts, a universal solution for the problem does not seem to exist, and pitch detection algorithms (PDAs) vary often in accordance to different requirements and applications. An even more difficult problem is the case of pitch tracking in polyphonic music. It is worth noticing that generally such procedures are called PDAs, even if they aim at extracting a signal-related physical variable such as the fundamental frequency and not the perceptual notion such as pitch. Also, the error estimation is often calculated using a semitone precision.

For the purpose of pitch extraction, it is possible to use spectral estimation procedures such as parametric and/or nonparametric ones. Other numerous methods operate on time, frequency, cepstral, and time-frequency domains. Among others, the autocorrelation, Schroeder's histogram, cepstral, *zero crossing* (ZXABE), *threshold crossing* (TABE), *two-threshold crossing* (TTABE), or *average magnitude difference function* (AMDF) methods can be cited [18], [35]–[47]. Pitch can also be estimated using such time-frequency methods as subband processing based on the Meddis–Hewitt model [37], [41] and the McAulay–Quatieri method [41].

There are two major problems, namely, octave errors and pitch estimation accuracy [38], [40], that most PDAs are susceptible to. Octave error problems seem to be present in all known pitch tracking algorithms; however, these errors are caused by different input signal properties in the estimation process. In time domain-based algorithms [44]–[46] octave errors may result from the low energy content of odd harmonics. In the frequency domain, errors are mostly caused by the low energy content of the lower order harmonics. In cepstral [40], as well as in autocorrelation of log spectrum (ACOLS) [48] analyses, such problems may appear due to the high energy content in higher frequency parts of the signal. On the other hand, the estimation accuracy problem for all mentioned domains is caused by a limited number of samples representing analyzed peaks related to the fundamental frequency.

In the study conducted here, both the AMDF method and a modified Schroeder's histogram were implemented and further analyzed. The AMDF method was chosen because of its relatively low computational cost. On the other hand, the Schroeder's method may be used in cases where some harmonics are missing or are not properly extracted, because on the basis of the statistical properties of the histogram, it is still possible to determine the most probable fundamental frequency. That is why both these algorithms may be used for large musical database analyses.

The AMDF method consists of searching the zeros of the following function:

$$AMDF_n(m) = \frac{1}{N}\sum_{i=1}^{N}|x_n(i) - x_n(i+m)| \qquad (1)$$

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

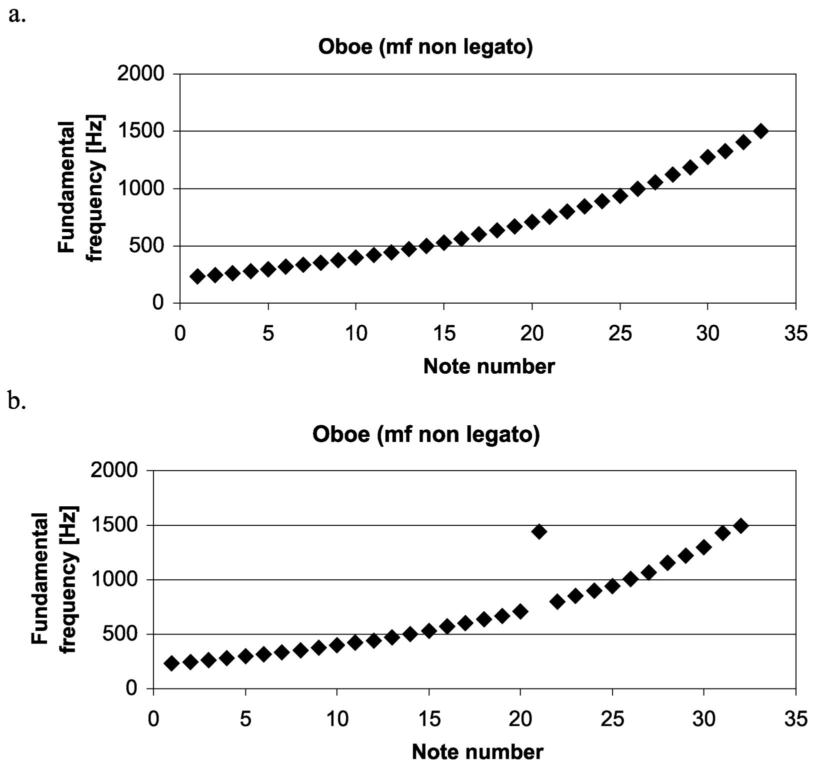

Fig. 1. Pitch detection performed for oboe sounds by: (a) the AMDF algorithm and (b) Schroeder's histogram.

where N is length of the analysis window, and $x_n(i)$ represents the analyzed sound.

In the case of quasi-periodic sound, this method does not secure a proper calculation of zeros. However, the estimation process can be carried out with sufficient quality provided that the search will be for the local minima of this function.

The AMDF method detects the fundamental frequency quite efficiently, whereas the second method is susceptible to octave errors. Generally, Schroeder's algorithm consists in building a histogram over the signal spectrum. For each peak, all its potential fundamentals are noted and their locations accumulated within the histogram. The largest peak within the histogram is assigned to the fundamental frequency. The engineered algorithm consists of three main steps. The first step is the fast Fourier transform (FFT) analysis, the second one aims at identifying sound harmonics, and the last step is the estimation of the fundamental frequency. Apart from the standard FFT and log procedures, the first step employs a low-pass filter (FIR) of the fifth order and the cutoff frequency equal to $\pi/80$. The second step assigns a threshold value corresponding to a certain level, above which spectrum partials are treated as signal harmonics and others are discarded. As a result of this operation, a spectrum consisting of partials is created. The third stage of the modified Schroeder's PDA consists in observation of the spectral distribution of partials. The assignment of the threshold value was done experimentally after performing dozens of sound analyses.

The above-presented pitch estimation methods were implemented and their performance tested using numerous musical sounds. The following instruments were used, namely, bassoon, clarinet, English horn, saxophone, French horn, trombone, trumpet, violin, viola, and cello. In addition, several representations of these instruments, i.e., sounds played forte, mezzoforte, piano, staccato, portato, and nonlegato were employed. All files were monophonic, sampled using a 44.1-kHz sample rate and 16-b resolution. Pitch estimator performance was evaluated using the mean absolute error

$$e = \frac{1}{N}\sum_{n=1}^{N} |\hat{f}_0 - f_0| \tag{2}$$

where $\hat{f}_0$ is the estimated and f_0 is real fundamental frequency.

However, from the practical point of view, in the recognition process, it is sufficient to estimate the pitch with semitone precision; therefore, the following formula was used:

$$df_{\mathrm{pr}} = \frac{100}{N}\sum_{n=1}^{N} r(n) \tag{3}$$

where

$$r(n) = \begin{cases} 0, & \text{for } \hat{f}_0 < f_0 \sqrt[12]{2^{-1}} \vee \hat{f}_0 > f_0 \sqrt[12]{2} \\ 1, & \text{for } f_0 \sqrt[12]{2^{-1}} < \hat{f}_0 < f_0 \sqrt[12]{2} \end{cases} \tag{4}$$

returning the percentage of correctly estimated fundamental frequencies with semitone precision.

In Fig. 1, an example of pitch detection process results obtained for oboe sounds employing AMDF and Schroeder algorithms can be observed. Using the AMDF algorithm, all oboe sounds of this particular articulation were properly estimated; on the other hand, Schroeder's histogram showed one octave error and one sound not being estimated at all. In

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

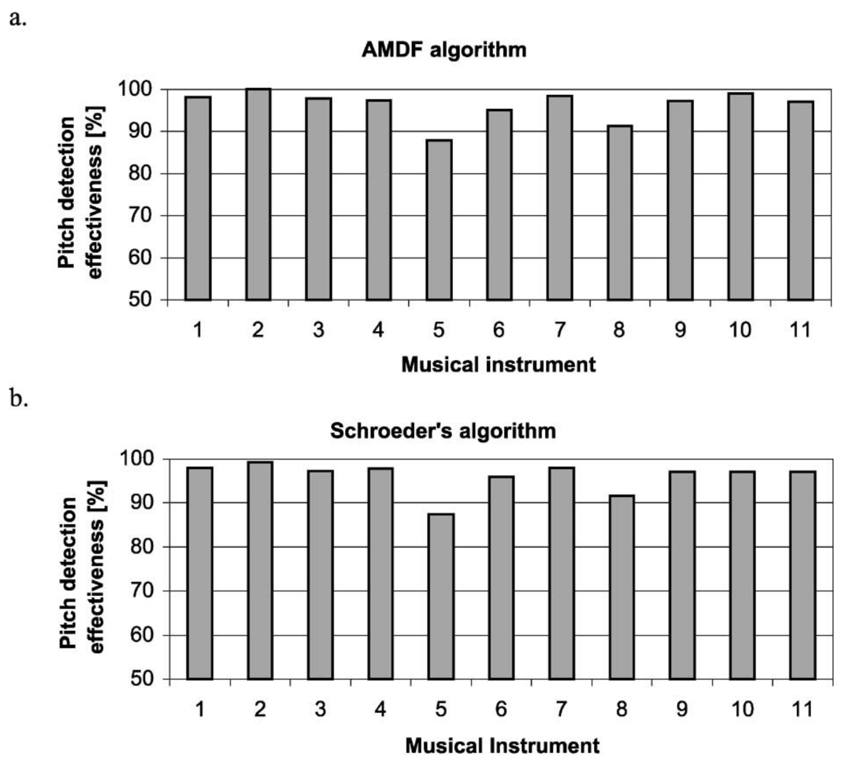

Fig. 2. Overall effectiveness of the PDAs. (a) AMDF method. (b) Modified Schroeder's histogram. Denotations assigned to the $x$axis are as follows: 1—trumpet, 2—trombone, 3—saxophone, 4—French horn, 5—bassoon, 6—oboe, 7—clarinet, 8—English horn, 9—violin, 10—viola, 11—cello.

Fig. 2, results of the implemented algorithm performances can be seen for various instruments and diverse articulation and dynamics. As seen from Fig. 2, the implemented algorithms give rise to some difficulties in proper recognition of the pitch of the individual instrument. The occurred errors were in most cases related to recognizing the fundamental frequency one octave higher or lower than the actual pitch. However, in the case of a bassoon instrument, some additional errors, such as misclassification, occurred. This happened when the difference between estimated frequency and the nominal for a particular note was larger than 100 cents. The best results of pitch detection were obtained for a trombone and viola, regardless of the articulation used. In this case, the precision of pitch tracking was 99%, and in the case of the AMDF algorithm, all trombone sounds were properly recognized. On the other hand, based on the performed analysis, it is difficult to state which of the implemented algorithms is more efficient. However, the overall effectiveness of both implemented algorithms was higher than 95% (see Fig. 2) and this is sufficient for recognition purposes.

B. Parametrization

The purpose of the parametrization process is to build FVs in order to diminish the information redundancy of the musical signal and to provide well-defined descriptors of musical data. On this basis, it is then possible to automatically classify musical instruments or musical phrases or to find a musical object that matches a given pattern. It should, however, be remembered that feature extraction is a heuristic process.

There are many parameters defined for musical instrument sounds that can be found in the literature. As mentioned before, some of these parameters were contained in the MPEG-7 standard. However, it happened that these descriptors were exploited to some extent only, meaning that in different studies various combinations of descriptors were used and tested as to their significance for the automatic search of musical objects within databases. They were derived from time, spectral, and cepstral, as well as time-frequency domains (see, e.g., [9], [10], [12]–[19], [22], [28], [29], [49], [50]). The applicability of various parameters has been verified extensively by the author; thus, only a review of the FFT-based FVs will be shown below. Of importance is the fact that there is further need to test the quality of the parameters contained in the MPEG-7 standard in a process of automatic identification of musical instrument classes.

The parametrization, exploited in the experiments carried out, consists in calculating various types of FVs and checking them statistically first, and then applying them to the automatic recognition of musical instrument sounds. They were called FFT-, MPEG-7-, and wavelet-based FVs. The FFT basis for the vector of sound description consists of 14 parameters, which are as follows: Tristimulus parameters proposed by Pollard and Jansson [51], rising time of Tristimulus parameters for both sound attack and steady state, delays of higher order harmonics in relation to the fundamental, contents of odd and even harmonics, brightness, and fundamental frequency [18], [19]. FVs based on the MPEG-7 standard description contain the following parameters: brightness, harmonic spectral spread,

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

harmonic spectral centroid, audio spectral flatness, attack time, temporal centroid, fundamental frequency, and, in addition, content of odd harmonics.

Since descriptors contained in the MPEG-7 standard and those derived on the FFT analysis were exploited very extensively and their definitions could be easily found in the literature [4], [12], [13], [19] or in some cases, such as contents of odd or even harmonics, are self-explanatory; thus, only the wavelet-based FVs analysis will be discussed more thoroughly.

Wavelet Analysis: One of the main advantages of wavelets is that they offer a simultaneous localization in time and frequency domain. This is also simply an alternative way of describing a signal in the frequency domain. Such a description in the frequency domain provides a more parsimonious representation than the usual one on a grid in the time domain. In order to define features that may be derived from the wavelet analysis, some extensive experiments were performed by the author and her team [18], [50], [52]. Frames consisting of 2048 samples taken from the transient of a sound were analyzed. Several filters such those as proposed by Daubechies, Coifman, Haar, Meyer, Shannon, etc., were used in analyses and their order was varied from two up to eight. It was found that Daubechies filters (second order) have the computational load considerably lower than while employing other types of filters; therefore, they were used in the analysis.

For the purpose of the study, several parameters were calculated. They were derived by observing both energy and time relations within the wavelet subbands. Below, energy-related parameters are shown. They are as follows [18]:

$$E_n = \frac{E_i}{E_{\text{total}}} \tag{5}$$

$$E_{\text{total}} = \sum_{i=1}^{10} E_i \tag{6}$$

where E_n denotes partial energy parameters, $E_i = E_1 \ldots E_{10}$ refers to energy computed for the wavelet spectrum subbands normalized with regard to the overall energy E_{total} of the parameterized frame corresponding to the starting transient, where:

$i = 1 \rightarrow$ energy in the frequency band 21.53–43.066 Hz;
$i = 2 \rightarrow$ energy in the frequency band 43.066–86.13 Hz;
$\ldots$;
$i = 10 \rightarrow$ energy in the frequency band 11 025–22 050 Hz, and c_k are consecutive wavelet coefficients.

In Fig. 3, sample results of the wavelet-based feature extraction (E_n) are shown for some chosen instruments. In all cases a frame consisting of 2048 sound samples was analyzed. In Fig. 3, energy values are presented for ten wavelet spectrum subbands. The whole instrument range can be seen within each subband. Left-side lines within each subband correspond to the lowest sounds, whereas the right-side lines correspond to the highest ones. It can be observed that energy distribution pattern within the wavelet spectrum subbands is quite similar for wind instruments. On the other hand, such a parametric representation of a violin differs greatly from wind instruments. Although this parameter is sensitive both to type of instrument and sound pitch, it is also, in a way, characteristic for wind and string instruments. On the other hand, such a sound representation of a viola instrument is much more similar to wind instruments than to the string group.

In the experiments, several time-related parameters were also explored. Two of them, the most significant from the statistical point of view, were included in the FV. The idea behind trying various parameters is as follows. If there are certain parameters that allow for easier distinguishing between particular instrument classes and others that will do the same for other classes, it is, thus, possible to design a system consisting of a few preprocessing blocks that will first separate for example groups of parameters. Thus, two additional relations are defined in the wavelet discrete-time domain. They are presented by (7) and (8), as follows:

$$e = \left(\frac{\sum\limits_{n=n_0}^{N} \left(|c_n| > 0.2 \cdot |c_{\max}| \right)}{512 - n_0} \right) \cdot \frac{1}{s} \tag{7}$$

where e is the time-related parameter allowing for characterization of the wavelet pattern, calculated for each wavelet spectrum subband, c_n are wavelet coefficients, n_o is the first wavelet coefficient that exceeds the assigned threshold, and s refers to sound pitch.

$$f = \text{var}\left(f' \left[|c_n| \right]\right) \tag{8}$$

where f is variance of the first derivative of the absolute value of the wavelet coefficient sequence.

The parameter e from (7) refers to the number of coefficients that have exceeded the given threshold. This threshold helps to differentiate between “tonelike” and “noiselike” characteristics of the wavelet spectrum. The value of such a threshold was assigned experimentally to 0.2. It then returns the associated sample number, which is illustrated in Fig. 4 for eighth wavelet spectrum subband. In Fig. 4, the assignment of threshold values is illustrated for two chosen instruments belonging to two subclasses (brass and woodwind). The x axis shows the number of samples, and the y axis corresponds to wavelet coefficient values normalized over maximum value $(c_{\max})$ of c_n. It takes approximately 180 samples for a trumpet sound and 220 samples for a clarinet to attain this threshold. The meaning of the parameter f is the variance estimation of the derivative of the sequence of coefficients.

Parameter Analysis: All calculated parameters were checked statistically for all pairs of instruments on the basis of the Fisher statistic (FS) [19], [53]. The FS is a useful tool for checking the separability of two classes for a given parameter p. The choice of FS for musical sound analysis was determined by the fact that the compared sets may consist of a different number of elements, such as when comparing the musical scale of a particular instrument. The basic assumption is that of equal mean values in two normally distributed populations [53].

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

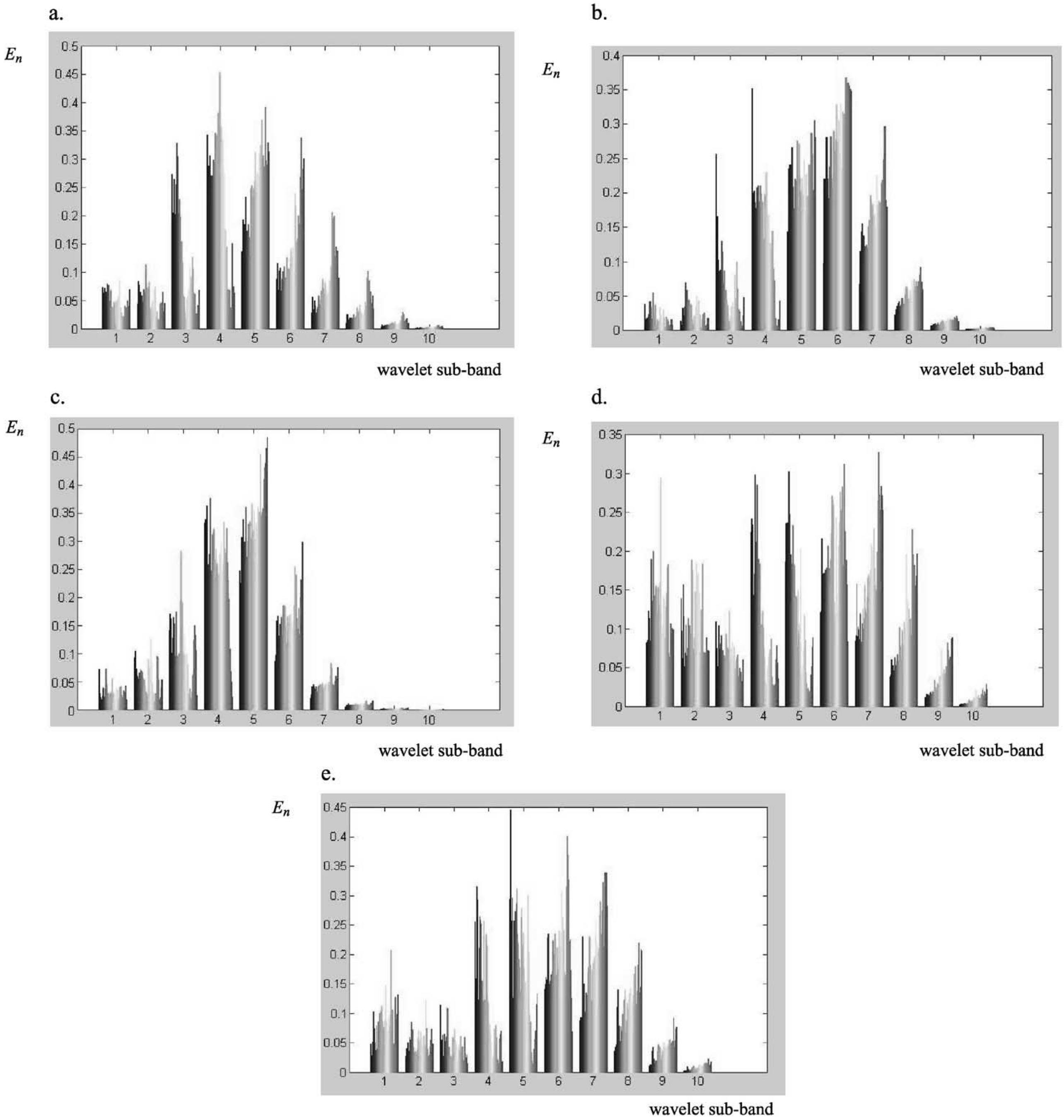

Fig. 3. Values of E_n parameters for selected instruments. (a) Saxophone. (b) Trumpet. (c) French horn. (d) Violin. (e) Viola. The x axis refers to consecutive wavelet subbands; the y axis corresponds to E_n values normalized with regard to the overall energy of the parameterized frame. The left-side lines within each subband correspond to the lowest sounds; the right-side lines, to the highest ones.

Therefore, the value M, based on the FS $|V|$ was calculated for every parameter p of two classes (instruments) X and Y, as defined below [19]

$$M = \min_{i,j}\left(\max_{p}|V(X_i, X_j, p)|\right) \tag{9}$$

where $|V|$ is FS applied to parameter p for the pair of instruments X and Y

$$V(X,Y) = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} \tag{10}$$

where $\overline{X}$, $\overline{Y}$ are mean parameter values for instruments X and Y, n, m are the cardinality of two sets of sound parameters, and S_1^2, S_2^2 are variance estimators

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{11a}$$

$$S_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2. \tag{11b}$$

The greater the absolute value of this statistic for the selected parameter for a chosen pair of instruments is, the easier it is to distinguish between objects representing these

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

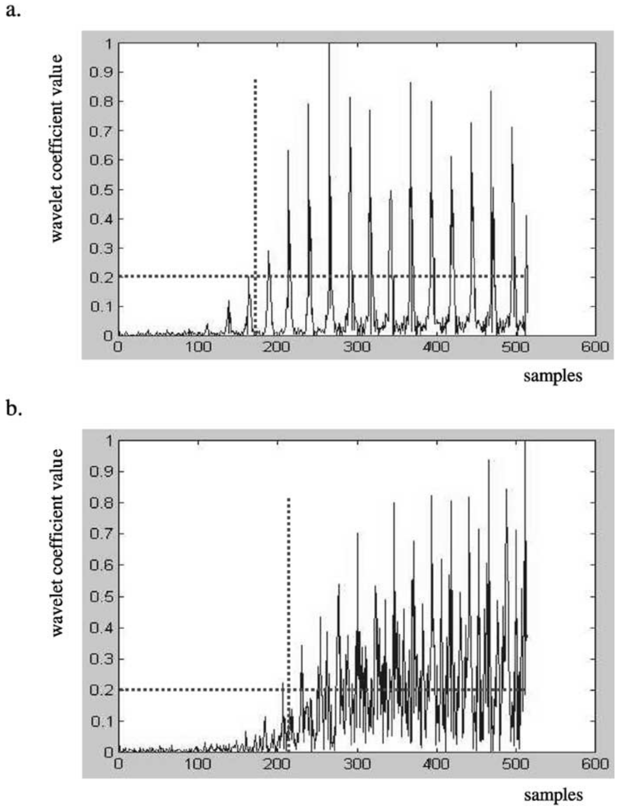

Fig. 4. Time-related parameter e allowing for discrimination between brass and woodwind instruments. (a) Trumpet. (b) Clarinet. The x axis shows the number of samples, and the y axis corresponds to wavelet coefficient values normalized over maximum value ($c_{\max}$) of c_n.

two instruments on the basis of the given parameter p. This implies that instruments will be discernible on the basis of the selected parameter if their mean values are definitely different, variances are small, and examples are numerous. In Table 1, absolute values of the FS for the selected instruments are shown for the parameter e calculated for the eighth wavelet subband. These instruments belong to different groups: woodwind, brass, and string. As seen from Table 1, the greatest values of this statistic were obtained in the case of instrument pairs: clarinet and trombone, and clarinet and trumpet. This means that on the basis of this parameter, these instruments might easily be distinguished from each other. On the other hand, it would be more difficult to properly distinguish between clarinet–violin and trombone–trumpet pairs, since the FS values are very low in these cases.

C. Pattern Recognition

The author's decision was to use one of the soft computing techniques in the recognition process. These techniques combine artificial NNs (ANNs), fuzzy and rough sets, evolutionary computing, genetic algorithms, and others. It has emerged that these techniques can be used with success in many fields of science and engineering, and they have proven to be convenient tools for handling uncertainty by offering better generalization properties and/or enabling the analysis of data that would not have been otherwise possible or explained. The latter remark refers to the rule-based systems, where the knowledge is contained in the rules derived. In the study presented in this paper, ANN-based classifiers were employed taking advantage of the generalization properties of such systems.

In experiments, multilayer NNs of a feedforward type were used. They contain only one hidden layer. The number of the neurons in the input layer depended on the structure of the FV. Thus, the input layer was changed according to the number of parameters included in FVs. After some optimization experiments based on the pruning method, it was decided that the hidden layer should contain 14 neurons. On the other hand, the number of neurons at the output depended on the number of instrument classes being recognized. During the learning phase of the NNs two methods were used, namely: the error back-propagation (EBP) method based on the delta learning rule (with *momentum*) and a so-called reduced memory Levenberg–Marquardt (LM) algorithm [54]. The latter method is based on Newton's method. The main difference between these two methods consists in the number of iterations needed for the error convergence and speed of learning. The LM method needs fewer iterations in order to obtain the desired error value; however, the cost of each iteration is much greater than in the EBP method.

Testing FV Effectiveness: To test FVs for their usefulness for musical instruments classification process, they were fed to the network input. Sounds were divided into two groups of the same size. The first one was a training set for the NN, whereas the remaining FVs were used in testing. Activation functions were unipolar sigmoidal, and the neurons at their outputs had values of from zero to one. In the classification process, the winning neuron had to be determined, i.e., one whose output value was the highest. The mean-square error was calculated during the training process, and this process was stopped when the error would continue to increase in five consecutive training cycles. In some cases the training took approximately 500 iterations, the learning error was reaching 0.02, and the validation error was 0.04. It should be remembered that such a procedure is heuristic and algorithm settings were established basing on experiments carried out very extensively [18], [19].

It should be stressed that experiments based on ANNs depend on the type of training, structure of the NN, and number of descriptors contained in FVs. Only a great load of experiments performed may result in a so-called optimum structure and algorithm settings. FVs containing FFT descriptors served such a role for gaining knowledge and some intuition on classification strategies. Various configurations of ANNs and almost all known learning schemes along with pruning techniques were employed during such experiments. However, in these preliminary classification experiments, only four musical instruments in various tested configurations were used. In Fig. 5, results are gathered for various configurations of musical instruments. They refer to various classes of musical instruments. In some cases, very high success rates were found, even for instruments belonging to the same instrument class such as woodwinds and brass. It should, however, be remembered that FFT-based parameters describe mainly harmonic relationship in the

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

Table 1
Separability Between Pairs of Musical Instruments Based on the Fisher Statistic Calculated for a Parameter Expressing Energy of the Eighth Wavelet Subband

	Trumpet	Trombone	Bassoon	Clarinet	Saxophone	English horn	French horn
Trombone	**1.3**						
Bassoon	9.2	9.7					
Clarinet	**18.8**	**19**	4.6				
Saxophone	2.8	3.6	5.3	11.2			
English horn	1.4	2.4	9.3	14.7	1.5		
French horn	16.2	17	5.9	2.3	11	13.9	
Violin	13.5	14	3	**1.2**	8.7	11.1	3.1

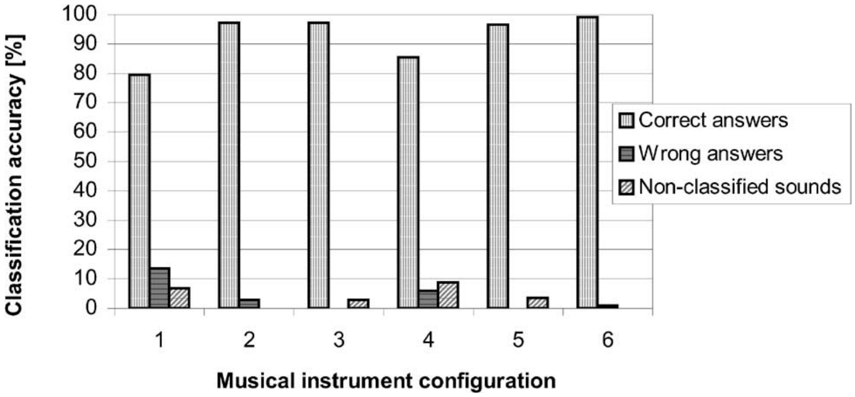

Fig. 5. Classification accuracy rates obtained for various configurations of musical instrument sounds: 1—double bass, cello, viola, violin (vibrato); 2—flute, tuba, violin, double bass; 3—bassoon, contrabassoon, C trumpet, trombone (tenor); 4—E-flat clarinet, bass clarinet, French horn, muted French horn; 5—oboe, bass clarinet, bassoon, bass trombone; 6—bass trombone, trombone, English horn, bassoon.

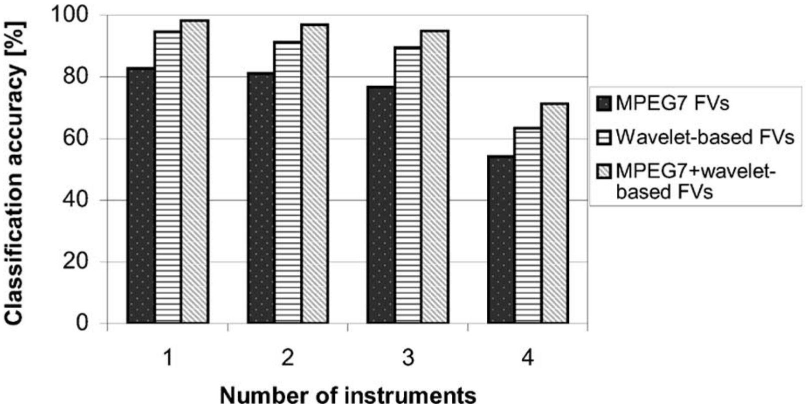

Fig. 6. Classification accuracy rates obtained respectively for: (1) four; (2) five; (3) six; and (4) twelve instrument classes for three types of FVs.

spectrum, whereas information on inharmonic parts of spectrum as well as more time-related features are needed in the classification. That is why, in further tests, MPEG-7- and wavelet-based FVs were employed; however, all performed experiments allowed for easier choice of ANNs structure and algorithm settings when the number of instruments was extended to 12 and diverse articulation was used.

Classification tests employing MPEG-7 and wavelet-based parameters were performed for four, five, and six instruments [55]. Accuracy rates were the highest for the combined representation; on the other hand, for five and six instruments, the effectiveness of the ANN classifier decreased for all FV representations in comparison to four instrument classes (see Fig. 6). After these preliminary

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

Table 2
Recognition Effectiveness of the MPEG-7-Based FVs

[%]	TRU	TRO	BASS	CLA	SAX	FRH	ENH	VI	VIOLA	CEL	PIA	OBOE
TRU	66.67	7.14	0.00	2.38	19.05	0.00	0.00	2.38	0.00	0.00	0.00	2.38
TRO	13.21	45.28	5.66	1.89	3.77	20.75	3.77	3.77	0.00	1.89	0.00	0.00
BASS	0.00	3.45	60.34	3.45	1.72	6.90	3.45	3.45	0.00	3.45	3.45	10.34
CLA	3.33	1.67	3.33	71.67	1.67	1.67	0.00	3.33	1.67	5.00	3.33	3.33
SAX	9.09	6.82	2.27	2.27	45.45	20.45	2.27	2.27	0.00	0.00	6.82	2.27
FRH	3.92	19.61	13.73	5.88	1.96	50.98	0.00	0.00	0.00	1.96	1.96	0.00
ENH	2.38	4.76	19.05	11.90	0.00	9.52	16.67	4.76	2.38	4.76	14.29	9.52
VI	6.82	0.00	0.00	9.09	0.00	0.00	0.00	54.55	4.55	15.91	0.00	9.09
VIOLA	4.29	2.86	4.29	2.86	10.00	4.29	2.86	5.71	58.57	0.00	1.43	2.86
CEL	3.03	1.52	12.12	6.06	0.00	4.55	3.03	7.58	3.03	59.09	0.00	0.00
PIA	0.00	0.00	1.54	0.00	1.54	1.54	3.08	0.00	4.62	0.00	87.69	0.00
OBOE	0.00	4.08	4.08	10.20	16.33	10.20	12.24	2.04	2.04	0.00	6.12	32.65

Table 3
Recognition Effectiveness of the Wavelet-Based FVs

[%]	TRU	TRO	BASS	CLA	SAX	FRH	ENH	VI	VIOLA	CEL	PIA	OBOE
TRU	83.33	2.38	0.00	0.00	4.76	0.00	0.00	0.00	0.00	0.00	0.00	9.52
TRO	1.89	58.49	7.55	0.00	0.00	18.87	0.00	0.00	0.00	0.00	13.21	0.00
BASS	0.00	5.17	87.93	1.72	0.00	1.72	1.72	0.00	0.00	1.72	0.00	0.00
CLA	0.00	3.33	1.67	51.67	8.33	0.00	8.33	0.00	5.00	11.67	5.00	5.00
SAX	0.00	6.82	0.00	6.82	81.82	0.00	2.27	0.00	2.27	0.00	0.00	0.00
FRH	0.00	19.61	13.73	3.92	1.96	60.78	0.00	0.00	0.00	0.00	0.00	0.00
ENH	0.00	4.76	0.00	14.29	2.38	0.00	61.90	2.38	0.00	9.52	2.38	2.38
VI	0.00	0.00	0.00	6.82	2.27	0.00	4.55	34.09	45.45	6.82	0.00	0.00
VIOLA	0.00	0.00	0.00	8.57	5.71	0.00	2.86	8.57	52.86	18.57	0.00	2.86
CEL	0.00	0.00	0.00	3.03	3.03	0.00	4.55	0.00	16.67	60.61	10.61	1.52
PIA	1.54	4.62	7.69	6.15	0.00	3.08	3.08	0.00	0.00	13.85	58.46	1.54
OBOE	2.04	0.00	0.00	12.24	2.04	0.00	2.04	4.08	6.12	2.04	2.04	67.35

tests the main experiments were performed and the number of classes was extended to 12. Sounds recorded at the Multimedia Systems Department as well as monophonic sounds from the McGill University, Montreal, QB, Canada, collection were used in the validation stage. Generally, in comparison to the case of 12 instruments being recognized, all results obtained for a smaller number of classes were much higher and at the same time all instruments were properly classified, though in the case of 12 instruments all but English horn and violin instrument classes were recognized properly. In Tables 2–4, results of automatic classification of FVs containing respectively MPEG-7-based, wavelet-based, and the combined MPEG-7- and wavelet-based descriptors are shown. Each line in the table stands for the recognized input object, and the columns show the ANN recognition effectiveness. The scores obtained for a particular instrument are given by the table diagonal (correct recognition of the class of the instrument).

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

Table 4
Recognition Effectiveness of the Combined MPEG-7 and Wavelet-Based FVs

[%]	TRU	TRO	BASS	CLA	SAX	FRH	ENH	VI	VIOLA	CEL	PIA	OBOE
TRU	90.48	0.00	0.00	0.00	2.38	0.00	0.00	0.00	0.00	0.00	0.00	7.14
TRO	1.89	73.58	7.55	0.00	0.00	16.98	0.00	0.00	0.00	0.00	0.00	0.00
BASS	0.00	3.45	79.31	3.45	0.00	3.45	3.45	0.00	0.00	5.17	1.72	0.00
CLA	0.00	0.00	1.67	60.00	1.67	0.00	8.33	0.00	5.00	11.67	5.00	6.67
SAX	4.55	2.27	0.00	6.82	81.82	0.00	0.00	0.00	0.00	0.00	2.27	2.27
FRH	0.00	23.53	11.76	0.00	0.00	62.75	0.00	0.00	0.00	0.00	1.96	0.00
ENH	0.00	4.76	7.14	9.52	2.38	0.00	54.76	7.14	2.38	4.76	0.00	7.14
VI	0.00	0.00	0.00	9.09	2.27	0.00	2.27	56.82	18.18	11.36	0.00	0.00
VIOLA	0.00	0.00	0.00	4.29	5.71	0.00	2.86	24.29	54.29	7.14	0.00	1.43
CEL	1.52	0.00	1.52	1.52	1.52	1.52	3.03	10.61	13.64	63.64	0.00	1.52
PIA	0.00	0.00	0.00	0.00	0.00	0.00	1.54	0.00	0.00	0.00	98.46	0.00
OBOE	0.00	0.00	0.00	10.20	0.00	0.00	2.04	2.04	6.12	0.00	0.00	79.59

Table 2 shows recognition effectiveness of the MPEG-7-based descriptors in percentages. As seen from Table 2, the English horn was not properly recognized, and apparently there were many problems with proper recognition of most instruments. Generally, it happened that ANN could not discern between instruments of the same group. In many, cases instrument sounds were confused within the same instrument group, such as woodwinds, brass, and strings.

Table 3 presents results of testing the effectiveness of the wavelet-based FVs. It can be observed that the highest scores are seen in the diagonal of the table. This means that for a particular instrument this was the winning output neuron. From such a point of view, most sounds were properly assigned to the appropriate instrument classes. There were some errors in recognition of violin and piano sounds, but overall better results were obtained than in the case of the MPEG-7-based FVs; this could be especially observed for saxophone, trumpet, and bassoon.

As seen from Table 4, recognition scores improved for most instruments. For example, piano sounds were recognized nearly errorless. Therefore, it can be said that the combined MPEG-7 and wavelet representation ensures much better recognition results. This is because the parameterization process covers both the transient and steady state of sound, and the resulting features of the FV complement each other.

IV. Musical Duet Separation

The musical duet separation is based on a modified FED analysis algorithm [50]. This algorithm decomposes the signal into linear expansion of waveforms, called EMO, that are a combination of complex exponential signals modulated by complex amplitude envelopes. These waveforms are sinusoidal signals with time-varying phase and amplitude and are chosen to best match the harmonic parts of the signal [50]. Such a representation is similar to the one introduced by Serra [7]; however, here inner products representing the decomposition frequency are directly related to the decomposition frequencies, whereas Serra's approach is based on retrieving partials of the signal from the spectrogram matrix. The presented solution works faster, since inner products are calculated only for chosen frequencies, and the retrieving phase is based on windowed inner product. In addition, nonharmonic structures can also be represented by such waveforms; this is possible when amplitude and phase changes have frequency higher than sinusoidal signal frequency. In practice, it means that in such analysis, we set our attention to consecutive spectral lines assuming that each line changes its amplitude and phase versus discrete time. Since the aim is to reconstruct intelligible audio data after the separation process in order to perform listening tests, that is why phase is retained in calculations. Oppenheim and Lim [56] and McAulay and Quatieri [57] pointed out in their studies the importance of phase in signal reconstruction.

The input signal can be represented as a sum of EMO structures (represented by amplitude envelope and phase) and a residual signal [58]

$$S[n]=\sum_{i=1}^{K} A_i[n]\cos\left(\frac{2n\pi f_i}{f_s}+\Theta_i[n]\right)+R_s[n] \qquad (12)$$

where $S[n]$ is the input signal, K is the number of decomposition iterations, A_i refers to the amplitude envelope for the ith iteration, $\mathit{\Theta}_i$ denotes the phase envelope for the ith iteration, and R_s is the residual signal.

The first step of the FED algorithm is the power spectrum density (PSD) estimation of the input signal using Welch's

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

averaged, modified periodogram method [41]. The frequency of the maximum value of the PSD $(f_{\max})$ is treated as the frequency of the most energy carrying EMO structure. Next is the calculation of nodes that represent the amplitude envelope of the real and imaginary part of related to the $f_{\max}$ complex exponential. Such an operation can be viewed as calculating inner products of the signal and the complex exponential divided into frames, where the inner product of each frame represents amplitude value. First, signals are multiplied sample by sample

$$S_m[n] = S[n]e^{j\frac{2n\pi f_{\max}}{f_s}} \tag{13}$$

where $S[n]$ is the input signal and $S_m[n]$ refers to signal multiplied sample by sample by complex exponential of the frequency $f_{\max}$.

Signal S_m is divided into frames of the same length as that of the complex exponential period, and for each block frame the value is calculated

$$K_i = \sum_{m=1}^{w(f_{\max})} S_m^i[n] \tag{14}$$

where K_i is the amplitude value for the ith block, $w(f_{\max})$ refers to frame length related to $f_{\max}$, and S_m^i is the ith frame of the S_m signal.

The node value for the ith frame is an inner product of the input signal in the ith frame and the complex exponential in the ith frame. To obtain amplitude signals of the same size as the input signal one, appropriate interpolation has to be performed. Cubic spline approximation provides interpolating curves that do not exhibit large oscillations associated with high-degree interpolating polynomials [42] and thanks to its low computational complexity, seems to be the perfect tool for the task of amplitude envelope interpolation. In the next algorithmic step, cubic spline interpolation is performed. It is also used to overcome the problem with phase unwrapping.

The first decomposition step is then performed

$$R_s[n] = S[n] - A_1[n]\cos\left(\frac{2n\pi f_1}{f_s} + \Theta_1[n]\right) \tag{15}$$

where R_s is the residual signal and f_1 refers to frequency $f_{\max}$ for the first iteration.

Each iteration is computed identically assuming that a residual signal of the previous iteration becomes the input signal for the next iteration. However, if the same $f_{\max}$ is detected, a significantly shorter amplitude calculation frame has to be applied and the iteration is then repeated, assuming that most of the energy carrying frequencies phase is disturbed and does not preserve harmonic properties. In this case, the EMO structure represents the nonharmonic part of the signal. Decomposition frequencies are chosen *a priori* for the FED. The first decomposition frequency is the fundamental frequency of the lower pitched instrument. Therefore, it is necessary to first employ a pitch estimation algorithm.

Since multipitch detection is not needed at this stage and one is interested in the lower instrument pitch only, an algorithm based on the correlation analysis seems to be well suited to carry out this task [42], [50]. However several modifications were applied to improve the accuracy of the algorithm according to the research done by Masuda–Katsuse [59].

The average pitch of a chosen segment results in the first decomposition frequency. It is assumed that this frequency is the fundamental frequency of the lower pitched instrument. Frequencies of the first ten harmonics are then calculated and FED iterations are performed for those frequencies. Since FED iterations can represent harmonic or inharmonic parts of the signal, a necessary modification of the FED was necessary in order to decompose only harmonic parts. Such modification is achieved by allowing only relatively large windows for calculating envelopes for each EMO.

The first K harmonics of the lower pitched instrument, within each segment can be represented as a sum of EMO structures and can be written in a simplified way as

$$\begin{aligned} I_1[n] = \sum_{i=1}^{K} &\left(\mathrm{Re}\left(\mathrm{EMO}(S_m)^{\mathrm{fi}}{}_i[n]\right)\right. \\ &\left.+ \mathrm{Im}\left(\mathrm{EMO}(S_m)^{\mathrm{fi}}{}_i[n]\right)\right) + R_{I_1}(S_m)[n] \end{aligned} \tag{16}$$

where S_m is the mth segment of the input signal, I_1 is the extracted signal containing harmonic components of the lower pitched instrument, K is the number of iterations or the number of harmonics to be decomposed, fi is the frequency corresponding to the ith harmonic, $\mathrm{EMO}(S_m)^{\mathrm{fi}}$ refers to the ith EMO corresponding to the ith harmonic frequency, and $R_{I_1}(S_m)$ is the residual signal containing inharmonic components of both instruments and harmonics of the higher pitched instrument.

The pitch detection procedure is repeated for $R_{I_1}(S_m)$ resulting in PCS. Further segmentation of S_m is carried out if necessary. FED decomposition is repeated for each segment of S_m. The first K harmonics of the higher pitched instrument can be represented as a sum of EMO structures

$$\begin{aligned} I_2[n] = \sum_{i=1}^{K} &\left(\mathrm{Re}\left(\mathrm{EMO}\left(S_{m_p}\right)^{\mathrm{fi}}{}_i[n]\right)\right. \\ &\left.+ \mathrm{Im}\left(\mathrm{EMO}\left(S_{m_p}\right)^{\mathrm{fi}}{}_i[n]\right)\right) + R_{I_2}\left(S_{m_p}\right)[n] \end{aligned} \tag{17}$$

where S_{m_p} is the pth segment of S_m, I_2 is the extracted signal containing harmonic components of the higher pitched instrument, K refers to the number of iterations or the number of harmonics to be decomposed, fi is the frequency corresponding to the ith harmonic, $\mathrm{EMO}(S_{m_p})^{\mathrm{fi}}$ denotes the ith EMO corresponding to the ith harmonic frequency, and $R_{I_2}(S_{m_p})$ is the residual signal containing inharmonic components of both instruments and harmonics of the lower pitched instrument.

A. Signal Decomposition

The segmentation of a sound based on PCS allows for small fluctuations of pitch. Pitch for each segment is actually an average pitch within such a segment. This generalization does not produce large errors in algorithm, since each EMO structure, thanks to the envelope frequency modulation properties, adapts itself to such small fluctuations.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

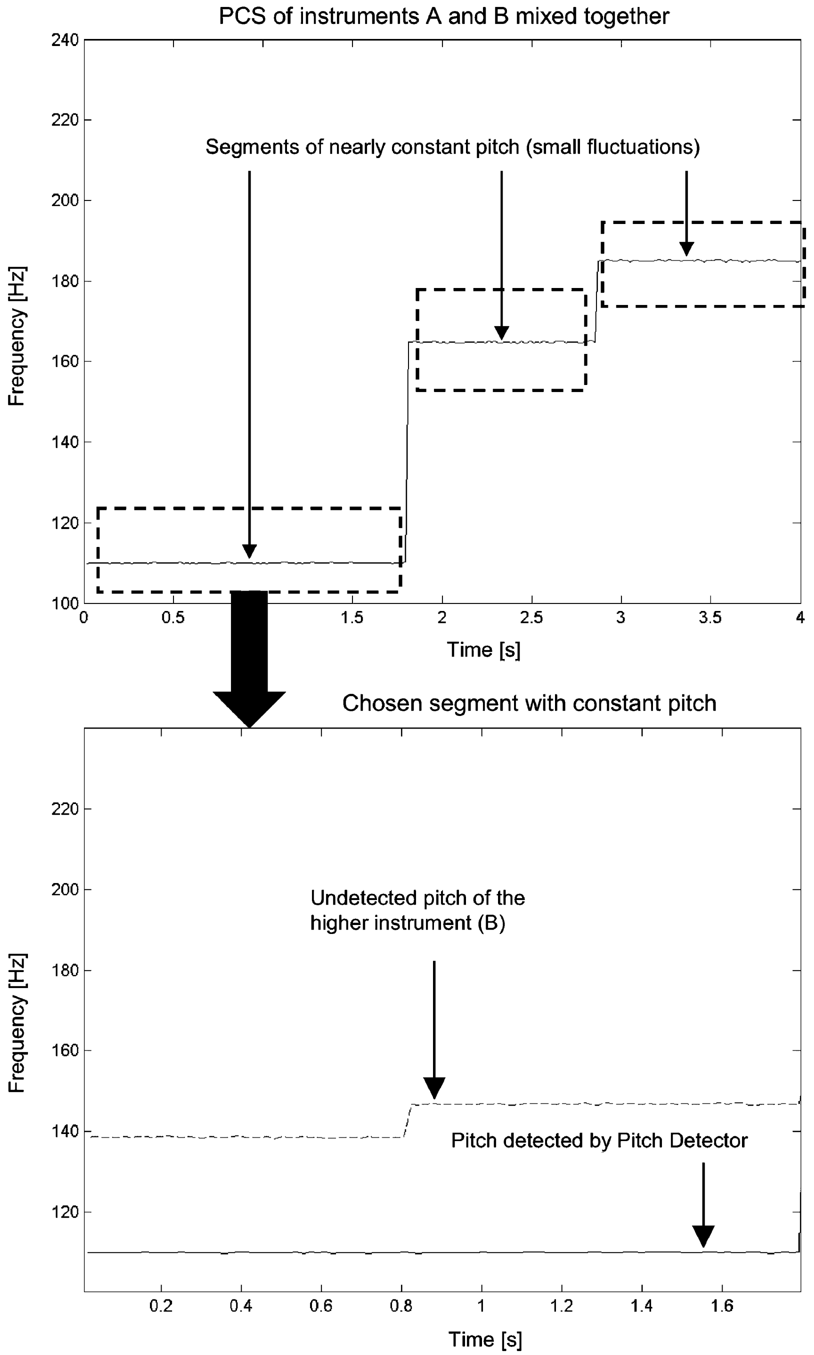

Fig. 7. Example of PCS for instruments A and B mixed together.

In Fig. 7, an example of the PCS is shown for instruments A and B mixed together. One segment of the input signal with constant pitch becomes the input signal for the FED decomposition algorithm. FED removes harmonics related to the detected pitch. The residual signal consists of harmonics from the higher pitched instrument. Based on the residual signal, the pitch contour of the remaining instrument can be calculated. Since the pitch of the higher instrument was not constant, further segmentation in this case would be required.

B. Harmonics Detection

Since fundamental frequencies of both instruments can be in harmonic relation, some of the harmonics from both instruments might share the same frequency. Frequencies of the coinciding harmonics can be easily calculated and eliminated for the sound recognition task if pitch of both instruments is known and eliminated for sound recognition tasks. Additionally, FED decomposition can be carried out for $R_{I_2}(S_{m_p})$ and for $R_{I_1}(S_{m_p})$, since both residual signals do not contain coinciding frequencies.

The FED of the residual signals can be expressed in a simplified way as

$$\begin{aligned} I'{}_2[n] = \sum_{i=1}^{K} \Big(\mathrm{Re}\Big(\mathrm{EMO}\Big(R_{I_1}(S_{m_p})^{\mathrm{fi}}{}_i[n]\Big) \\ + \mathrm{Im}\Big(\mathrm{EMO}\Big(R_{I_1}(S_{m_p})^{\mathrm{fi}}{}_i[n]\Big)\Big) \end{aligned} \tag{18}$$

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

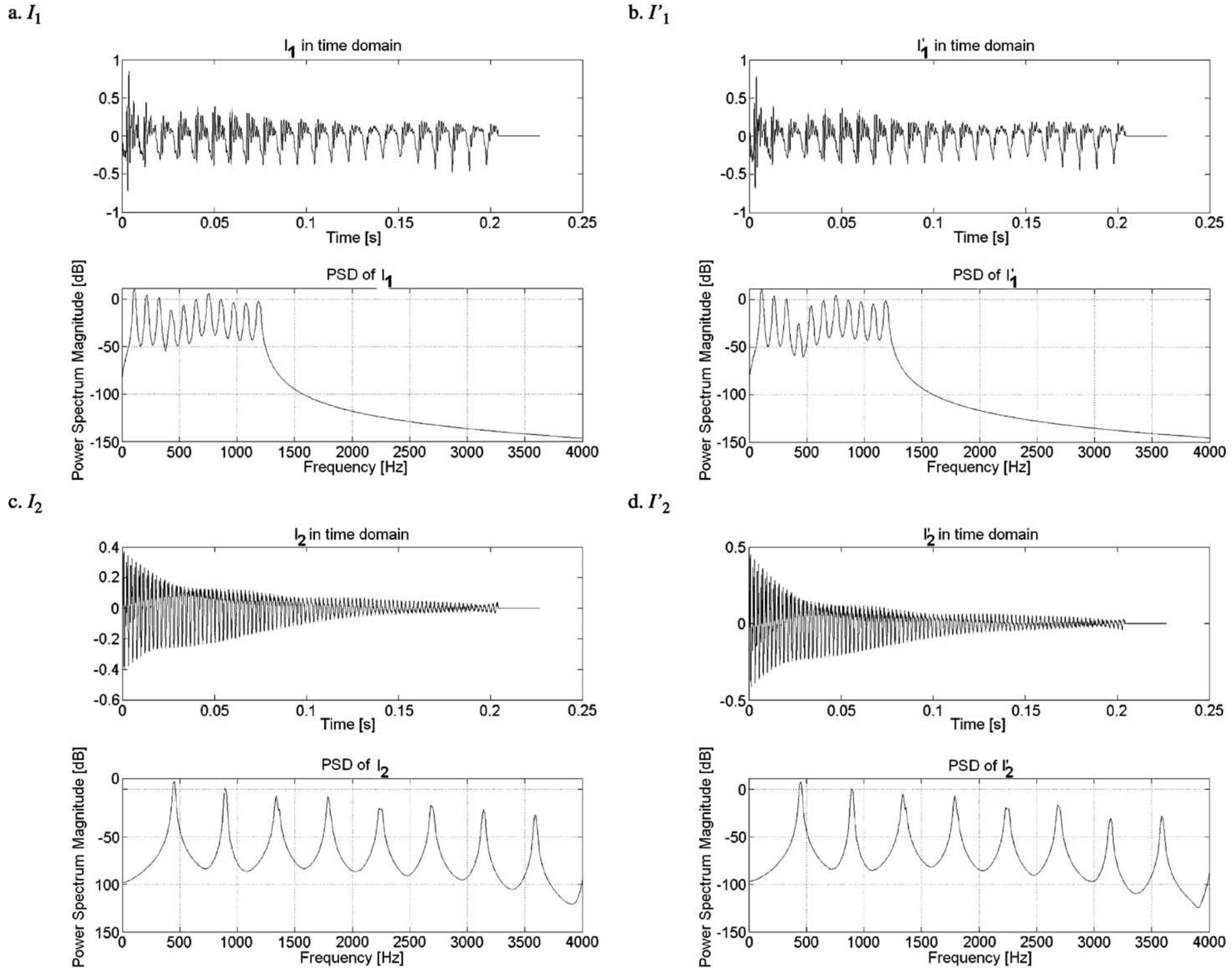

Fig. 8. Separated I_1, I_2, I'_1, and I'_2 signals (I_1, I'_1 refer to cello sound and, correspondingly, I_2, I'_2 to saxophone sound).

$$I'_1[n] = \sum_{i=1}^{K} \Big(\mathrm{Re}\Big(\mathrm{EMO}\Big(R_{I_2}(S_{m_p})^{\mathrm{fi}}{}_i[n]\Big) + \mathrm{Im}\Big(\mathrm{EMO}\Big(R_{I_2}(S_{m_p})^{\mathrm{fi}}{}_i[n]\Big)\Big) \tag{19}$$

where I'_2 is the higher pitched instrument signal in the pth segment, containing noncoinciding harmonics, and I'_1 is the lower pitched instrument signal in the pth segment, containing noncoinciding harmonics.

Fig. 8 shows I_1, I_2, I'_1, and I'_2 EMO representations resulting from one segment of the signal consisting of the mixed 448.8-Hz saxophone sound with 108.1-Hz cello sound. Fig. 8 contains both time- and frequency-domain plots of sounds after separation.

V. Automatic Separation of Musical Duets

For the purpose of checking the efficiency of the FED algorithm devoted to the task of musical duet separation, some musical instrument sound recognition experiments based on the ANN were used. The structure of this ANN was three-layer, consisting of correspondingly 8 (input layer), 20 (hidden layer), and 8 (output layer) neurons. A unipolar sigmoidal transfer function was used and a back-propagation training algorithm with momentum was applied during the training phase. The ANN was trained employing about 400 sound excerpts containing sounds of eight instruments of differentiated musical articulation [52]. In addition, an optimization process of generalization properties was performed using another set of 400 sounds. It consisted in stopping the ANN training every time when the mean-square error appeared to increase.

Sounds that were separated employing the FED algorithm were then parametrized. I_1, I_2, I'_1, and I'_2 signal representations after separation were used for FV calculation and these FVs were then fed to the ANN. The sample answers of the NN for pairs of musical sounds: clarinet and trumpet and English horn and viola are presented in Tables 5–8. In Tables 5 and 6, the ANN answers are given for testing FVs derived from pairs of sounds before mixing (original sounds), and Tables 7 and 8 contain results for testing FVs resulted from parametrization of sounds after separation based on the FED algorithm. The consecutive columns refer to instruments on which the ANN was trained. Rows correspond to sounds that were used in the ANN testing. Values contained

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

Table 5
ANN Output Neuron Answers for Trumpet and Clarinet Sounds (Original Sounds)

Musical Instr./	CLA	CLA	CLA	CLA	TRU	TRU	TRU
ANN output	A#3	A4	A#4	A#5	A3	B3	A4
VIOLA	0.0005	0	0.0003	0.0127	0	0	0.0034
ENG. HORN	0.032	0	0	0.3089	0.0047	0.0045	0.0039
FR. HORN	0.0053	0.0347	0.0453	0	0.0001	0.0001	0.0001
SAX	0.0176	0.2743	0.1415	0	0.0146	0.0343	0.0141
CLARINET	**0.9888**	**0.9484**	**0.9178**	**0.7783**	0.0006	0.0018	0.0001
BASSOON	0	0.0001	0.0001	0.0002	0	0	0
TROMBONE	0	0.0462	0.0176	0	0.0129	0.0011	0
TRUMPET	0	0	0	0	**0.99**	**0.9877**	**0.9987**

Table 6
ANN Output Neuron Answers for Viola and English Horn Sounds (Original Sounds)

Musical Instr./	ENH	ENH	ENH	ENH	VIOLA	VIOLA	VIOLA
ANN output	A3	A#3	C#4	B4	A3	A#3	A4
VIOLA	0.6678	0.2621	0.4945	0.3158	**0.9712**	**0.872**	**0.9916**
ENG. HORN	**0.9101**	**0.8726**	**0.808**	**0.892**	0.0105	0.5759	0.0082
FR. HORN	0.001	0.0006	0.0007	0	0.0003	0.0003	0.0002
SAX	0	0	0	0	0	0	0
CLARINET	0.0024	0.0082	0.0003	0.3842	0.0102	0.0042	0
BASSOON	0	0	0	0	0	0	0
TROMBONE	0	0	0	0	0	0	0
TRUMPET	0	0	0	0	0	0	0

in the tables refer to the ANN output while testing particular sounds and they are highlighted in bold when correctly classified by the ANN. The following sounds were mixed in two groups in pairs:

- trumpet A4—clarinet A#4;
- trumpet B3—clarinet A4;
- trumpet A3—clarinet A#5;
- trumpet B3—clarinet A#3;

and

- English horn A3—viola A#3;
- English horn A#3—viola A#4;
- English horn B4—viola A3;
- English horn C#4—viola A#4.

Remark: B3 trumpet and viola A#4 sounds were used twice in the mix of sounds.

As seen from Tables 5 and 6, values at the output neurons corresponding to sounds being recognized were very close to one, whereas output of remaining neurons was close to zero. On the other hand, as seen from Tables 7 and 8, values at the output neurons corresponding to sounds being recognized were in some cases not so close to the value of one; however, these neurons were the winning neurons in the output layer. In addition, some sounds were not recognized properly. The residual signal containing both inharmonic spectrum content and overlapping harmonics of the other sounds from the duet caused the erroneous answer, still allowing for the recognition of one of the mixed sounds. For example; it can be observed that recognition of sounds of the viola and English horn was much easier than trumpet and clarinet for the ANN-based algorithm. In the first case, the recognition system had some problems with octave-related sounds,

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

Table 7
ANN Output Neuron Answers for Trumpet and Clarinet Sounds (Sounds After FED Separation)

Musical Instr./	TRU	TRU	TRU	TRU	CLA	CLA	CLA	CLA
ANN output	A4	B3	A3	B3	A#4	A4	A#5	A#3
VIOLA	0.647	0.657	0	0	0.013	0.888	0.446	0.002
ENG. HORN	0.006	0.002	0.013	0	0	0.044	0.06	0.0001
FR. HORN	0.0001	0.0001	0.0009	0.491	0.099	0	0.0003	0.083
SAX	0.019	0.615	0.006	0.179	0.02	0.02	0	0.251
CLARINET	0	0	0.009	0.0001	**0.866**	0	0.008	**0.942**
BASSOON	0	0	0	0.0002	0	0	0	0
TROMBONE	0	0	0.0002	0.484	0.005	0	0	0
TRUMPET	**0.956**	0.01	**0.931**	**0.762**	0.138	0.224	0.313	0.085

Table 8
ANN Output Neuron Answers for Viola and English Horn Sounds (Sounds After FED Separation)

Musical Instr./	ENH	ENH	ENH	ENH	VIOLA	VIOLA	VIOLA	VIOLA
ANN output	A3	A#3	B4	C#4	A#3	A#4	A3	A#4
VIOLA	0.003	0	0.053	0.012	**0.778**	**0.776**	**0.999**	**0.724**
ENG. HORN	**0.597**	0.0002	**0.848**	**0.97**	0.0008	0.051	0.008	0.602
FR. HORN	0.001	0.115	0.0005	0.0001	0	0.0007	0.0002	0.004
SAX	0.0001	0.002	0.0001	0	0	0	0	0
CLARINET	0.011	0.021	0.003	0.004	0	0.184	0.0009	0.0006
BASSOON	0	0	0	0	0	0	0	0
TROMBONE	0	0.0008	0	0	0	0	0	0
TRUMPET	0.003	0	0.0004	0.0001	0	0	0	0

whereas in the second case, only sounds of the first and fourth pairs were recognized properly.

VI. Conclusion

In this paper, some sample results of the musical instrument class recognition process and musical duet separation were shown. Such experiments belong to the MIR field, and they are only a part of this domain, which aims at automatic retrieval of complex information from musical databases. It was shown that the soft computing approach to music instrument classification is justified with recognition scores. Usually scores obtained for a small number of instruments are very high, and for a larger number of instruments, in most cases, despite the decision vagueness, the system indicates the appropriate instrument. In addition, as seen from results obtained, the FED algorithm separates musical duets quite efficiently. This is especially true for instruments belonging to different musical instrument groups. Still, the problem may remain when two sounds are octave related, but in such a case, one of the mixed sounds is chosen by the ANN as the most probable to happen. It is worth noting that after the processing, human experts recognized separated sounds without any difficulties. Further experiments will include some optimization of the FED algorithm in order to improve the ANN-based recognition process results.

Acknowledgment

The author would like to thank the anonymous reviewers for providing insightful comments and suggestions, which contributed to the improvement of this manuscript.

References

[1] J. S. Downie, "Music information retrieval," in *Annual Review of Information Science and Technology*. Medford, NJ: Information Today, 2003, vol. 37, ch. 7, pp. 295–340.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

[2] ISMIR: The international conferences on music information retrieval and related activities [Online]. Available: http://www.ismir.net/

[3] Meta-labs.com [Online]. Available: http://www.meta-labs.com /mpeg-7-aud

[4] *International Organisation for Standardisation, Coding of Moving Pictures and Audio: MPEG-7 overview*, ISO/IEC JTC1/ SC29/WG11. [Online]. Available: http://www.chiariglione.org/ mpeg/standards/mpeg-7/mpeg-7 .htm.

[5] A. T. Lindsay and J. Herre, "MPEG-7 and MPEG-7 audio—an overview," *J. Audio Eng. Soc.*, vol. 49, no. 7/8, pp. 589–594, 2001.

[6] N. Laurenti and G. de Poli, "A method for spectrum separation and envelope estimation of the residual in spectrum modeling of musical sound," presented at the COST G6 Conf. on Digital Audio Effects (DAFX-00), vol. 2000, Verona, Italy.

[7] X. Serra, *Musical Sound Modeling With Sinusoids Plus Noise*. Lisse, The Netherlands: Swets & Zeitlinger Publishers, 1997.

[8] S. McAdams, "Concurrent sound segregation I: Effects of frequency modulation coherence," *J. Acoust. Soc. Amer.*, vol. 86, no. 6, pp. 2148–2159, 1989.

[9] J. C. Brown, "Computer indentification of musical instruments using pattern recognition with cepstral coefficients as features," *J. Acoust. Soc. Amer.*, vol. 105, pp. 1933–1941, 1999.

[10] 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.

[11] P. Cosi, G. de Poli, and G. Lauzzana, "Auditory modeling and self-organizing neural networks for timbre classification," *J. New Music Res.*, vol. 23, no. 1, pp. 71–98, 1994.

[12] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing 2000*, pp. 753–756.

[13] P. Herrera, X. Amatriain, E. Battle, and X. Serra. Toward instrument segmentation for music content description: a critical review of instrument classification techniques. presented at Int. Symp. Music Information Retrieval (ISMIR 2000). [Online]. Available: http://ismir2000.indiana.edu/

[14] P. Herrera, X. Serra, and G. Peeters, "A proposal for the description of audio in the context of MPEG-7," in *Proc. Eur. Workshop Content-Based Multimedia Indexing (CBMI'99)*, pp. 81–88.

[15] I. Kaminskyj, "Multi-feature musical instrument sound classifier v/user determined generalization performance," in *Proc. ACMC 2002*, 2002, pp. 53–62.

[16] ——, "Multi-feature musical instrument sound classifier," in *Proc. ACMC*, 2000, pp. 46–54.

[17] K. Kashino and H. Tanaka, "A sound source separation system with the ability of automatic tone modeling," in *Proc. ICMC*, 1993, pp. 248–255.

[18] B. Kostek and A. Czyzewski, "Representing musical instrument sounds for their automatic classification," *J. Audio Eng. Soc.*, vol. 49, no. 9, pp. 768–785, 2001.

[19] B. Kostek, "Soft computing in acoustics, applications of neural networks, fuzzy logic and rough sets to musical acoustics," in *Studies in Fuzziness and Soft Computing*. Heidelberg, Germany: Physica Verlag, 1999.

[20] ——, "Computer based recognition of musical phrases using the rough set approach," *J. Inf. Sci.*, vol. 104, pp. 15–30, 1998.

[21] C. Papaodysseus, G. Roussopoulos, D. Fragoulis, T. Panagopoulos, and C. Alexiou, "A new approach to the automatic recognition of musical recordings," *J. Audio Eng. Soc.*, no. 1/2, pp. 23–35, 2001.

[22] G. de Poli and P. Prandoni, "Sonological models for timbre characterization," *J. New Music Res.*, vol. 26, no. 2, pp. 170–197, 1997.

[23] European IST project 'RAA' (2000). [Online]. Available: http://www.joanneum.ac.at

[24] P. Y. Rolland, "Discovering patterns in musical sequences," *J. New Music Res.*, vol. 28, no. 4, pp. 334–350, 1999.

[25] ——, "Adaptive user modeling in a content-based music retrieval system," in *Proc. 2nd Int. Symp. Music Information Retrieval ISMIR'01*, Bloomington, IN, 2001.

[26] G. Tzanetakis and P. Cook, "Audio information retrieval (AIR) tools," presented at the Int. Symp. Music Information Retrieval (ISMIR'00), Plymouth, MA.

[27] M. Ueda and S. Hashimoto, "Blind decomposition of concurrent sounds," in *Proc. ICMC*, 1994, pp. 311–318.

[28] A. Wieczorkowska and A. Czyzewski, "Rough set based approach to automatic classification of musical instrument sounds," *Electron. Notes Theor. Comput. Sci.*, vol. 82, no. 4, pp. 1–12 , 2003.

[29] A. Wieczorkowska, J. Wroblewski, D. Slezak, and P. Synak, "Application of temporal descriptors to musical instrument sound recognition," *J. Intell. Inf. Syst.*, vol. 21, no. 1, pp. 71–93, 2003.

[30] 20th WCP: Philosophy and cognitive science [Online]. Available: http://www.bu.edu/wcp/MainCogn.htm

[31] [Online]. Available: http://www.w3.org/Metdadata

[32] L. Zadeh, "Fuzzy logic=computing with words," *IEEE Trans. Fuzzy Syst.*, vol. 2, pp. 103–111, May 1996.

[33] ——, "From computing with numbers to computing with words—from manipulation of measurements to manipulation of perceptions," *IEEE Trans. Circuits Syst.*, vol. 45, pp. 105–119, Jan. 1999.

[34] B. Kostek, "'Computing with words' concept applied to musical information retrieval," presented at the Int. Workshop RSKD, Warsaw, Poland, 2003.

[35] J. W. Beauchamp, "Detection of musical pitch from recorded solo performances," presented at the 94th. Audio Engineering Soc. Convention, Berlin, Germany, 1993.

[36] J. C. Brown, "Musical fundamental frequency tracking using a pattern recognition method," *J. Acoust. Soc. Amer.*, vol. 92, no. 3, pp. 1394–1402, 1992.

[37] A. Czyzewski, A. M. Szczerba, and B. Kostek, "Pitch estimation assisted by the neural network-based prediction algorithm," in *Proc. ISMA 2002*, pp. 246–255.

[38] W. Hess, *Pitch Determination of Speech Signals, Algorithms and Devices*. Berlin, Germany: Springer-Verlag, 1983.

[39] A. Klapuri, "Wide-band pitch estimation for natural sound sources with inharmonicities," presented at the 106th Audio Engineering Soc. Convention, Munich, Germany, 1999.

[40] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 14, pp. 293–309, 1967.

[41] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing. Principles, Algorithms and Applications*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[42] L. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. Gonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 399–418, Oct. 1976.

[43] M. R. Schroeder, "Period histogram and product spectrum: new methods for fundamental frequency measurement," *J. Acoust. Soc. Amer.*, vol. 43, pp. 829–834, 1968.

[44] X. Quian and R. Kimaresan, "A variable frame pitch estimator and test results," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, 1996, pp. 228–231.

[45] D. Talkin, *A Robust Algorithm for Pitch Tracking (RAPT). Speech Coding and Synthesis*. New York: Elsevier, 1995, pp. 495–518.

[46] A probabilistic approach to AMDF pitch detection, G. S. Ying, L. H. Jamieson, and C. D. Michell. [Online]. Available: http://purcell.ecn.purdue.edu/~speechg

[47] J. D. Wize, J. R. Caprio, and T. W. Parks, "Maximum-likelihood pitch estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 418–423, Oct. 1976.

[48] N. Kunieda, T. Shimamura, and J. Suzuki, "Robust method of measurement of fundamental frequency by ACOLS-autocorrelation of log spectrum," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, 1996, pp. 232–235.

[49] B. Kostek and P. Zwan, "Wavelet-based automatic recognition of musical instruments," in *Proc. 142nd Meeting Acoustical Soc. Amer.*, vol. 110, 2001, p. 2754.

[50] B. Kostek, P. Zwan, and M. Dziubinski, "Statistical analysis of musical sound features derived from wavelet representation," presented at the 112th Audio Engineering Soc. Convention, Munich, Germany, 2002.

[51] H. F. Pollard and E. V. Jansson, "A tristimulus method for the specification of musical timbre," *Acustica*, vol. 51, pp. 162–171, 1982.

[52] B. Kostek, M. Dziubinski, and P. Zwan, "Further developments of methods for searching optimum musical and rhythmic feature vectors," presented at the 21st Audio Engineering Soc. Conf., St. Petersburg, Russia, 2002.

[53] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes*. Cambridge, U.K.: Cambridge Univ. Press, 1992.

[54] J. Zurada, *Introduction to Artificial Neural Systems*. St. Paul, MN: West, 1992.

[55] B. Kostek, P. Zwan, and M. Dziubinski, "Musical sound parameters revisited," presented at the Stockholm Music Acoustics Conf. (SMAC'03), Stockholm, Sweden.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.

[56] A. V. Oppenheim and J. S. Lin, “The importance of phase in signals,” *Proc. IEEE*, vol. 69, pp. 529–541, May 1981.

[57] R. J. McAualay and T. F. Quatieri, “Pitch estimation and voicing detection based on sinusoidal speech model,” in *Proc. ICASSP*, vol. 1, 1990, pp. 249–252.

[58] M. Dziubinski, “Evaluation of musical instrument sound separation method effectiveness in polyphonic recordings by means of soft computing methods,” Ph.D. dissertation, Multimedia Syst. Dept., Gdansk Univ. Technol., Gdansk, Poland, to be published.

[59] I. Masuda-Katsuse, “A new method for speech recognition in the presence of non stationary, unpredictable and high level noise,” in *Proc. Eurospeech 2001*, pp. 1119–1122.

Bozena Kostek (Member, IEEE) received the M.Sc. degree in sound engineering and the M.Sc. degree in organization and management from the Gdansk University of Technology (GUT), Gdansk, Poland, in 1983 and 1986, respectively, the Diplome d’Acoustique degree from the Paul Sabatier University of Toulouse, France, in 1988, the Ph.D. degree from GUT in 1992, and the D.Sc. degree at the Warsaw Institute of Research Systems, Polish Academy of Sciences, Warsaw, Poland, in 2000.

She is an Associate Professor with the Department of Multimedia Systems, GUT. She has also led a number of research projects sponsored by the Polish State Committee for Scientific Research. She has presented more than 200 scientific papers in journals and international conferences. Her main scientific interests are musical acoustics, music information retrieval, the psychophysiology of hearing, and applications of soft computing methods to the mentioned domains.

Prof. Kostek is a Member of the Audio Engineering Society, Acoustical Society of America, and others. In 2000, she received an award for scientific achievements from the Prime Minister of Poland. In 2003, she was elected the AES Vice-President for central Europe.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 29,2010 at 17:13:01 UTC from IEEE Xplore. Restrictions apply.
View publication stats