Musical Instrument Identification based on a Short-Time Spectral Analysis

Christoph M. Moll & Tim. J. Ellis
Department of Electrical, Electronic & Information Engineering
City University, London EC1V 0HB, U.K.

Abstract: - In this paper, an implementation of an audio recognition system for personal computers is presented, combining methodologies of digital signal processing, machine perception, and statistical decision models. In particular, attention was given to musical tones, harmonics, and MIDI notes, that build the musical context to identify two musical instruments from their corresponding musical notes. Altogether nine musical instruments from electronic and acoustic devices have been sampled, with an 89% correct performance for single-mode identification and 67% correct performance for double-mode identification.

Key-Words: - pattern recognition, Bayes decision theory, statistical classifiers, musical analysis, harmonics

1 Introduction

Perception and recognition are something everyone experiences but nobody completely understands. A practical example is the sense of hearing, with such a complexity, that still there is no sufficient human understanding to build computer systems which are capable of understanding music at the level of an average five-year-old [1].

On the whole there are at least two different categories of listeners concerned with opposite modes of listening process. An "expert" listener who may be interested in particular aspects of the music, e.g. key sense, harmonic perception, and pitch equivalence represents the first. Alternatively, the "non-expert" listener (or non-musician) is not in a position to understand the functional properties in musical sound and does not recognize common musical structures. Our interest lies in making the computer function like an expert. Hence, analyzing the structure of musical sound means identifying musical instruments, musical tones, MIDI notes, chords, pitch, and musical transcription, in a piece of music.

This problem demands detailed knowledge of acoustics and determining a set of characteristic features that a recognition system is capable of analyzing. A desirable aim of computational *auditory scene analysis* [2] is to create computer systems that handle acoustic features that often betray physical properties of their sources. This paper presents just one specific feature set, supplied by the harmonics of musical instruments (and therefore is only applicable to melodic instruments – this excludes drum instruments which produce musical noise) that are to be identified. On the other hand, Martin [3] explores several additional features to be used for the purpose of sound source recognition. Just naming them briefly, the *Pitch, Frequency Modulation, Spectral Envelope, Spectral Centroid, Intensity, Amplitude Envelope, Amplitude Modulation, Onset Asynchrony, and Inharmonicity* can be observed and extracted either within the correlogram of sound, that explores the distribution of energy given a time and a frequency axis, or the *PCM* signal of sound, that represents temporal properties of sound waves. An understanding of DSP (digital signal processing), especially ADC (analogue-to-digital conversion), PCM (pulse code modulation), FFT (fast fourier transform), window method, Normal Density Functions, and Bayes theory, is recommended.

2 Problem Formulation

The goal of the work described in this paper was to build a system to attempt the separation of simultaneous musical sounds and to identify the musical instruments. Additionally, the extraction of musical notes and MIDI notes in the sound signal has been realized to approach *automatic music transcription*. The necessary technique to obtain musical notes demands *polyphonic pitch tracking* [4]. Moorer was the first in the literature to attempt separation of two simultaneous musical sounds [5]. His system demonstrated *pitch tracking*, given that the voices do not cross, the pitches are piece-wise constant (i.e. no vibrato or jitter), and the fundamental frequencies of the tones are not in a 1:N relationship (unison, octave, twelfth, etc).

The methodology introduced in this paper works directly from a short-time spectral analysis, and the tool developed could be implemented into a real-time recognition system.

3 Problem Solution

While constructing a pattern recognition system, knowledge in different areas is needed. It can be categorized into musical content analysis, DSP on audio signals and classification and decision systems.

3.1 Musical Content Analysis

Sound is a phenomenon prescribed by the laws of nature emerging as fluctuations in pressure, which exists in the path of a sound wave. Some dictionaries explain musical sound as sound produced by mechanical vibrations, which is not wrong, but doesn't satisfy common instruments like some electronic devices. A particular part of sound is the *pure musical tone* that is a plain, steady single note of constant pitch and intensity, or in other words, it is a sine wave of constant vibration[6]. Thus, we expect a single peak in the frequency spectrum indicating the frequency of specific vibration.

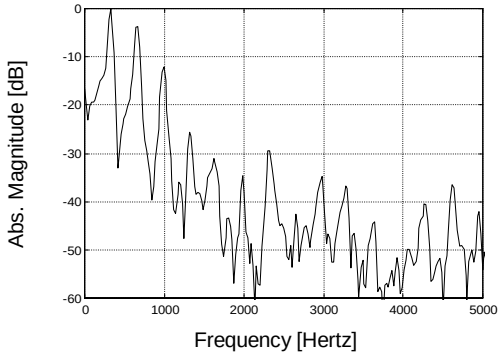


Fig.1 Power spectrum density of a single guitar note 'E4' (sampling frequency 8192 Hz, 1024-point FFT, frequency range 0 – 5000 Hz, energy normalized to zero dB)

Analysis of a guitar note (Fig.1) turns into a process of splitting a complex sound wave into its fundamental (lowest tone) and various overtones (*harmonics*). This property has been proved by Pythagoras (around 570BC), a Greek philosopher and inventor of the *Musical Pythagorean Scale*, who investigated the ratios of a vibrating string, which are found to be inversely proportional to the length. This accumulation of pure musical tones is called a *harmonic spectrum*. As a rule we might say the overtones exhibit frequencies higher than integer multiples of the fundamental frequency [7], but often inharmonicity can be observed. Because of mechanical stiffness, freely vibrating strings produce overtones in the vicinity of the estimated pitch period [3].

To come to the point, the fundamental and its overtone frequencies are very characteristic features to be used by a recognition system. But what is the number of overtones musical instruments provide? James Boyk's study on strings, woodwinds, brass, and percussion instruments shows that at least one member of these instrument families produces energy to 40 kHz [8]. For example, the 'ultrasound' of a French horn can extend to above 90 kHz. Dependent on the noise in sound it becomes difficult to measure spectral components of higher frequencies, as the proportion of energy above 20 kHz comprises only approximately 2%.

3.1.1 Overlapping Harmonics

Assuming that a piano is striking two keys at the same time, then the power spectrum density will show harmonics of both keys overlapped (Fig.2)

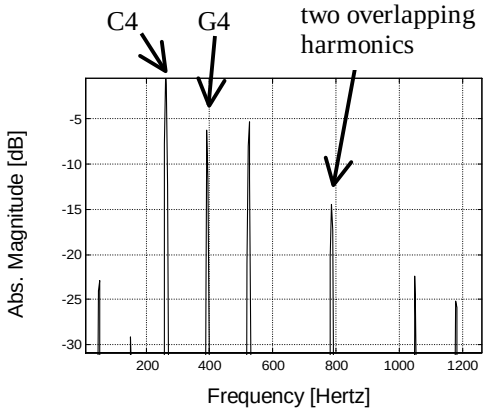


Fig.2 Power spectrum density of two piano keys, note C4 at 261.6 Hz and note G4 at 392.0 Hz (sampling frequency 8192 Hz, 1024-point FFT, frequency range 0 – 13000 Hz, energy normalized to zero dB)

Reconsider, the harmonics belong to an integer multiple of the fundamental frequency. It turns out that the peak around 800 Hz in Fig.1 is the sum of the 2nd harmonic of note C4 and the 1st harmonic of note G4. When proceeding further the 5th harmonic of note C4 overlaps with the 3rd harmonic of note G4 at approximately 1568 Hz, then the 8th harmonic with the 5th harmonic at approximately 2353 Hz, and so on.

3.2 Pattern Recognition System & Bayes Classifier

Three components are used to build a pattern recognition system, the *transducer*, a *feature extractor* and a *classifier,* illustrated in Fig.3.

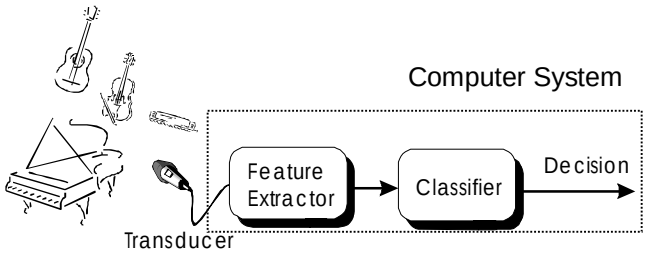


Fig.3 Real-time pattern recognition system for musical sound

The job of the transducer is to record a sound signal in analogue realm and to pass it through an ADC into digital realm.

The feature extractor is the next component in the pattern recognition chain, the starting point for pattern classification that is often the computation of features that are stored in *feature vector x*. There are two important aspects while programming the feature extractor. One aspect is the selection of features, which is responsible for performance of the classifier. Both, the physical interpretation and the correlation of different

features must be considered. The second aspect is the selection of feature space corresponding to the total number of features.

Pattern classification entails the analysis of the measurement features to distinguish an input pattern amongst a number of *'i'* possible pattern classes ω_i. To identify any combination of pattern various techniques can be used to make judicious inferences from the available information. Some possibilities without careful attention are to explore *FUZZY logic* and *neural networks*. In this work a classifier using Bayes model and discriminant functions has been designed. Bayes classifier [10] will require storing the *covariance matrices* Σ, the *mean vector* μ, and the *a priory probability* $P(\omega_i)$, by the following functions for each class ω_i respectively

(1) $$W_i = -\frac{1}{2}\sum_i^{-1}$$

(2) $$w_i = \sum_i^{-1}\mu_i$$

(3) $$\boldsymbol{w_{i0}} = -\frac{1}{2}\mu_i^{\,t}\sum_i^{-1}\mu_i - \frac{1}{2}\log|\sum_i| + \log P(\omega_i)$$

To determine the decision the discriminant function

(4) $$g_i(\boldsymbol{x}) = \boldsymbol{x}^t W_i \boldsymbol{x} + \boldsymbol{w_i}^t \boldsymbol{x} + \boldsymbol{w_{i0}}$$

must be adapted to all classes ω_i of instruments respectively given an input feature vector *x*. The lowest value of $g_i(x)$ will indicate this particular class to represent the identified instrument.

3.3 System Arrangement

The complete programming work was done on two personal computers, a Pentium II-400 MHz and less powered Pentium I-75 MHz. The tool for programming was MATLAB including the Image Processing and a Signal Processing Toolbox, which support the *FFT* and *Windowing*.

When developing a pattern classification system there are two independent steps to run until a decision can be made. One is the *teaching phase* with an objective to find appropriate features of musical instrument and to create the classifier, or in other words, the decision surfaces. The *classification phase*, i.e. the second step, proceeds by extracting features and classifying them into one of several instrument classes ω_i.

Altogether 9 musical instruments have been applied in the teaching phase: the class electric 4-string bass guitar, electric 6-string guitar, flute, harmonica and violin sampled with a microphone that was connected directly to the 16-bit soundboard of the computer; and the class electric piano, acoustic piano, harmonica, and pipe organ provided by a sampler keyboard connected to the same soundboard. All samples have been recorded from a single note of approximately 2 seconds (30,000 – 80,000 sample values using a 22,050 Hz sampling frequency at 8-bits). For multimode instrument identification these samples have been mixed via an audio software tool and stored separately. Furthermore, we require that pitches be identified with the actual note relative to the equal tempered scale relating to A4 being 440 Hz. This demands the capability of *absolute pitch* [6] that has been provided by all musical instruments.

3.3.1 Extractor

The problem is to extract feature vector *x* required for the discriminant functions (equation 1-4) of the classifier. The extractor is an essential part for both, the teaching phase and the classification/testing phase. However, this particular component seems to be very tricky to realize and demands different skills and a lot of experience, especially if the signal contains more than one musical instrument.

The presented implementation implies the windowing, *N-point* FFT, calculation of absolute magnitude, and normalization (based on normalizing the largest coefficient to zero dB) to be applied to the discrete audio signal. Subsequently the data obtained passes some numeric maximum functions that deliver the peaks in the power spectrum, i.e. the harmonics of the sound signal. Dependent on the instrument type and whether the note is low or high (which determines the richness of harmonics) the extractor normally returned between 3 and 40 peaks.

It goes without saying that the number of sample values taken from a digitized music signal must be the same as the number of samples applied to the *N-point* FFT, i.e. *N* values. Two difficulties may arise when *N* is not chosen correctly. S*ide-band leakage* [9] can cause errors, which affects the extraction of the fundamental frequency if the integer value *N* was chosen too small. The lower the fundamental frequency the larger value *N* must be chosen to minimize the leakage. The second aspect upon selection of *N* is the envelope of the sound signal itself. It makes no sense to extract the power spectral density of a deterministic signal if the signal is changing its spectral properties over time (i.e. non-stationary). Rather it is meaningful to extract a steady-state portion of sound, supported in a short sequence of data. As shown in Fig.4 the loudness of a piano note is not constant for all three sections (see envelope), therefore it is necessary to set *N* to quite a small value. In practice the best results were achieved for *N=8192*, giving a portion of sound within Δt seconds.

(5) $$\Delta t = \frac{N}{F_s}s = \frac{8{,}192}{22{,}050}s = 0.372s$$

Given this selection of value *N* the boundary for the spectral leakage is set to be

(6) $$\Delta B_1 = \frac{F_s}{N} = \frac{22{,}050Hz}{8192} = 2.6917Hz$$

Spectral leakage plays a prominent role for the detection of the fundamental frequencies. The difference in frequency of two adjacent semi-tones Δf must satisfy the boundary of spectral leakage to serve correct identification of musical notes or MIDI notes, hence $\Delta B < \Delta f$ must be provided. Adapting these results the lowest note that can be identified without inclination to errors is given by

$N = 8192$: Note Gb1 = 46.24930 Hz
Note G1 = 48.99943 Hz
Δf = 2.75013 Hz, ΔB = 2.6917 Hz
→ Lowest musical note is G1

3.3.2 Classifier & Teaching phase

In the teaching phase 68 files of sampled tones were used to build the classifiers.

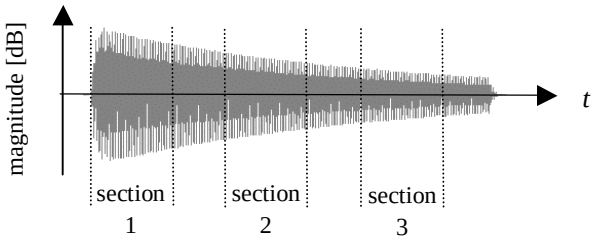


Fig.4 Sampled piano tone. When building the classifier all three sections per sampled note with 4096 or 8192 discrete values must be considered.

The envelope characterizes the energy for short-time segments of sound as illustrated in Fig.4. In musical context this envelope is divided into *attack*, *decay*, *release*, and *sustain*. Only a few musical instruments like violin and harmonica do not show sustain varying, i.e. the envelope over a long-time period is almost constant. But since the properties change with time the harmonics and hence the feature values also change. Better results for classification can be achieved when samples in the teaching phase were partitioned into three sections. Each section includes the signal that is to be extracted, relating to three sets of features per sampled note. Consequently the extracted data contained 204 different feature vectors of 9 different musical instruments within a 3-section classifier (3 sections x 68 samples in the teaching phase). An 8-section classifier was also tested, using 544 feature vectors. Given these feature vectors the *covariance matrices* Σ and *mean vectors* μ have been calculated for each instrument class using equations (1) to (3).

To become more familiar with statistical distribution of harmonics Fig.5 a) illustrates the *univariate density function* for the 4th harmonic of all piano samples given the *mean* by $\mu = -26.60$ dB and the *standard devi*ation σ by the square root of variance $c = 58.90$. Plot b) shows the distribution of the 4th harmonic given all sampled piano sounds. On closer inspection some equivalence in the shape of both plots can be seen. To obtain a more homogeneous shape of plot b) in comparison to plot a) more samples of the piano would be required.

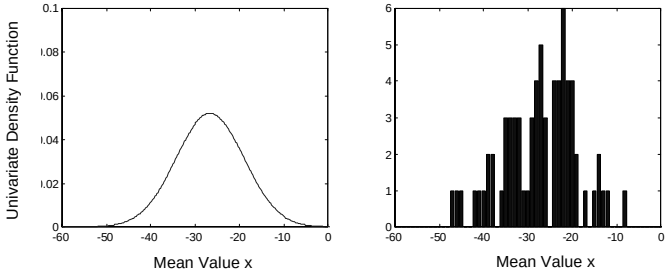


Fig.5 Piano samples: a) gaussian distribution of 4th harmonic given the values of mean and variance; b) real distribution of 4th harmonic

The reason why the system is expected to distinguish between the instruments provided can be observed in the next plot. It seems very clear that properties of the 5th harmonic dictate different *mean* values and slightly changing *standard deviations*. Statistical analysis would detect different decision boundaries for this peaky feature.

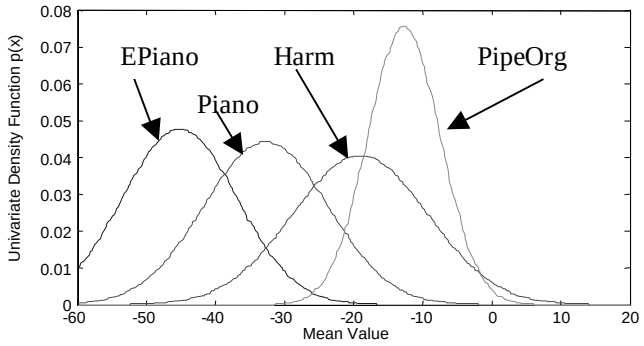


Fig.6 *Mean* and *Standard Deviation* plots of 5th harmonic given the musical instruments electric piano, acoustic piano, harmonica, and pipe organ.

3.3.3 Single Mode Identification

Firstly the classifier has been tested in *single mode*. Single notes corresponding to instruments playing in isolation are sampled and input to the feature extractor and then the classifier. The sampling rate is dictated by the maximum bandwidth of signal that we need to analyze. Considering the highest musical note we wish to represent in the system (i.e. C6), with a frequency value 1047 Hz, and the 9th harmonic – i.e. the 10th feature – is 10,470 Hz, at most 10 harmonic features would be used within the decision model. A *sample rate* of 22,050 Hz was chosen to satisfy *the Nyquist rate* [9], resulting in a spectral range from 0 Hz to 11,025 Hz.

In using the discriminant functions, a mandatory rule must be obeyed, which says that *absolute magnitude of covariance matrix* $|\Sigma|$ is zero and *multivariate normal density function* p(x) is degenerate if sample vectors drawn from a normal population are confident to a linear subspace [10]. This happens, for example, when one

component of feature vector *x* has zero variance, or when two components are identical. After normalization of the power spectral density data it turns out that the fundamental of the electric bass guitar yields always the energy of zero dB, thus the fundamental frequency cannot be used as a feature because of its zero variance.
Feature numbers 2–10 correspond to the first nine harmonics, hence define the feature vector of the classification system given by

(7) $\boldsymbol{x}=\{ 1^{st} \text{ harmonic}, 2^{nd} \text{ harmonic}, \ldots , 9^{th} \text{ harmonic} \}$

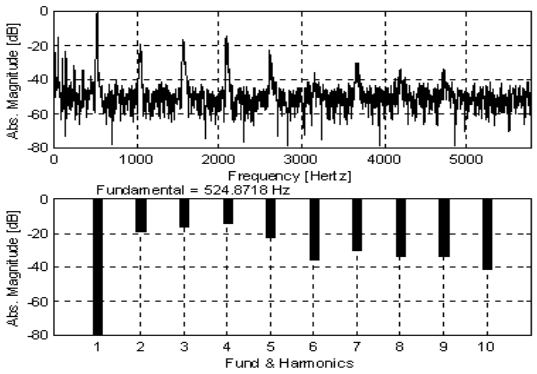


Fig.7 Extracted features of violin note C5 (523 Hz); a) power spectral density, b) representation of features skipping the fundamental (feature 1) but appending 9 successive harmonics (pointed out in feature 2 to 10)

Since the 68 samples served in the teaching phase do not include extrinsic musical sounds, 28 additional samples delivered by the sampler keyboard have been added, giving the electric Bass_1 (8-bit), electric Bass_2 (8-bit), and violin (16-bit). It is also possible to re-use the samples employed in the teaching phase for classification if a different section within the same sample is used.
The first system tested was the 3-section classifier, which operated on 4 separate sections per sample giving a total of 96 samples. Each section comprised 8192 samples offset at sampling positions 111; 4,444; 8,888; and 20,000 into the PCM sound signal. Regarding these instrument samples that do support the covariance matrices the identification rate of musical notes and instruments reaches mostly 90 to 100 percent, excepting harmonica and violin with 70% and 75% respectively. Concerning the samples that do not support the covariance matrices, the classifier assigned 85% of electric Bass_1 samples to the class ‘electric piano’ or 'acoustic piano'. Electric Bass_2 samples have been assigned to class ‘electric bass’ 73% correctly, violin 16-bit samples have been assigned to the ‘violin’ class 88% correctly. We may summarize the harmonics of electric Bass_1 – at least the first nine – indicate similar behavior comparing the harmonics of electric or acoustic piano. When this instrument was excluded in the classification, the overall identification rate rose to 89.3%, otherwise it showed 82.9%. The 8-section classifier returned an overall identification rate of 85.3% given an advantage of only 2.4%.
Figure 8 shows a scatter of the distribution of two particular features, for a harmonica and a piano, given by the fourth and fifth harmonic. It clarifies two distinct clusters, and a statistical decision boundary can be drawn (approximately) into this plot to indicate the operation of the classifier.

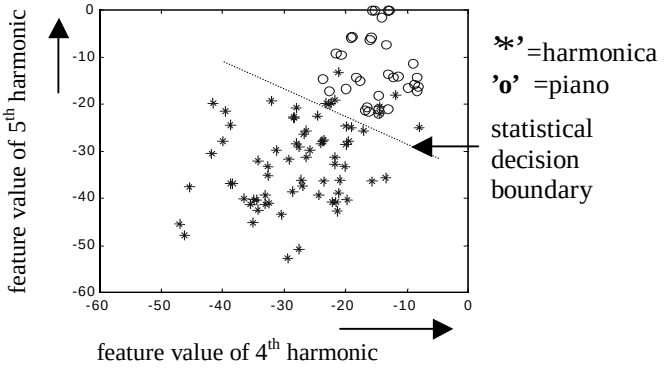


Fig.8 Scatter diagram for two features, i.e. 4th and 5th harmonic, of a harmonica and a piano inherited from various samples of different musical notes

3.3.4 Double Mode Identification

Given the theory on musical acoustics introduced in section 3.1.1 we know that overlapping harmonics are not permitted to be loaded into the classifier. Hence, a pre-selection of harmonics (features) must be applied to the sound samples we want to categorize.

	Fundamental	1st Harmonic	2nd Harmonic	3rd Harmonic	4th Harmonic	5th Harmonic	6th Harmonic	7th Harmonic	8th Harmonic	9th Harmonic
	Feature No.									
Interval	**---**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
+ 1	o	x	x	x	x	x	x	x	x	x
- 1	o	x	x	x	x	x	x	x	x	x
+ 2	o	x	x	x	x	x	x			
- 2	o	x	x	x	x	x				x
+ 3	o	x	x	x	x			x	x	x
- 3	o	x	x	x			x	x	x	x
+ 4	o	x	x	x		x	x	x		
- 4	o	x	x		x	x			x	x

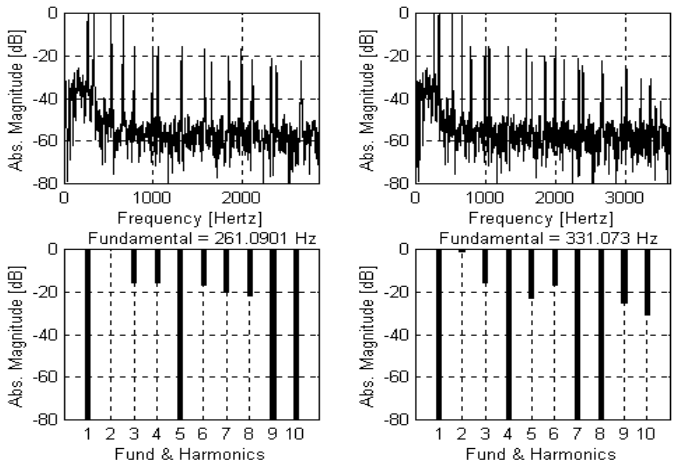
Table 1. Sample of overlapping harmonics for two simultaneous musical notes. Left column represents upwards and downwards intervals of semi-tones (MIDI code). A cross indicates this particular harmonic does not overlap while an empty field means the opposite. The fundamental, represented by circles, is not used.

Unfortunately, present covariance matrices and mean vectors used for single-mode identification cannot be used for double-mode identification, excepting for some special cases. As a result several different classifiers

must be built that will operate on specially selected harmonics. The selection of harmonics comes from Table 1, which indicates the valid harmonics for all intervals between *first* and *third*. Assuming that a given musical sound includes a large variety of intervals, Table 1 must be extended to include these intervals.

The extraction procedure before classification is very significant and works as follows: After obtaining the lowest fundamental frequency the first nine successive harmonics can be extracted. If the sound sample includes a second musical note that is not an interval of *unison* or *multiple octave* a different harmonic/fundamental should be detected in between the harmonics/fundamental of the lower note. Utilizing this particular feature leads back to the fundamental of the second musical note. The required musical note and MIDI note can be read out from tables given both fundamental frequencies. The expected MIDI code contains integer values starting from 0 up to 119; the difference in MIDI code for both fundamentals represents the musical interval.

Figure 9 illustrates the power spectral density of two simultaneous piano keys, sorting out the overlapping harmonics indicated as –80 dB bars.



Fi**g.9** Feature extraction of two simultaneously sounding piano notes C4 (261 Hz) and E4 (329 Hz); a) power spectrum of note C4 (0-2,871 Hz), b) power spectrum of note E4 (0-3,619 Hz), c) extracted features of note C4, d) extracted features of note E4

The testing procedure operated in a similar way as for single-mode identification given mixed instrument samples. The hits for correct identification of two musical notes, two MIDI notes, and two musical instruments proved to be 61.7% for 12 samples including 2 or even 3 musical tones, and 67.5% for 11 samples of two simultaneous stroked piano keys. Wrong decisions could be observed to have two key causes. First, the selection of features was sometimes incorrect, second the number of features, which depends on the interval between both fundamentals, caused a wrong decision if it was too small.

4 Conclusion

In respect to real-time systems, short-time spectral analysis provides several harmonics of melodic instruments that can be used to categorize instruments and their musical notes in a piece of music. Musical theory and methodologies on DSP supply a dominant point of view for musical signal extraction that here supported 9 harmonic frequencies adapted to the feature space. The classification obeying a statistical decision model returned at maximum two correct identifications of musical sound even if the sound included three different sources. Identification of two instruments within approximately 65% correct performance confirmed the idea that correct separation of harmonics is a potential investment for multiple instrument identification.

Acknowledgment.

The first author acknowledges the support of an advanced course studentship from the UK Engineering and Physical Science Research Council (EPSRC) that supported this work.

References:

[1] K. D. Martin, E. D. Scheirer & B. L. Vercoe, *Music Content through Models of Audition*, MIT Media Laboratory Machine Listening Group, Cambridge MA USA, 1998

[2] A.S. Bregman, *Auditory Scene Analysis*, Cambridge: MIT Press, 1990

[3] K. D. Martin, *Toward Automatic Sound Source Recognition: Identifying Musical Instruments*, MIT Media Laboratory Machine Listening Group, Cambridge MA 02139 USA, 1998

[4] E. D. Scheirer, *Music Perception Systems*, a proposal for a Ph.D. dissertation, MIT Media Laboratory, 22 Oct. 1998, pp. 24

[5] J. A. Moorer, On the transcription of Musical Sound by Computer, Computer Music Journal 1/4, 1977, pp. 32-38

[6] J. Askill, *Physics of Musical Sound*, D. Van Nostrand Company, New York, 1979 pp. 67-87

[7] A. H. Benade, *Fundamentals of Musical Acoustics*, Second Edition, New York: Dover, 1990

[8] J. Boyk, *There's life Above 20 Kilohertz, A Survey of Musical Instrument Spectra to 102.4 kHz*, Music Lab, California Institute of Technology, Pasadena, CA 91125 USA, May 2000

[9] P. A. Lynn, *Digital Signals, Processors and Noise*, First edition, The Macmillan Press LTD, 1992, pp. 6-33, 90-122, 127

[10] R. O. Duda, P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley, New York, pp. 1-38