Consiglio Nazionale delle Ricerche
ISTI – Istituto di Scienza e Tecnologie della Informazione

AN IDENTIFICATION SYSTEM OF MUSICAL TONES

Bertini Graziano, Di Salvo Vincenzo Giovanni

B4 - 20

September 2003

PISA

Consiglio Nazionale delle Ricerche
ISTI – Istituto di Scienza e Tecnologie della Informazione
Area della ricerca di PISA-S.CATALDO

AN IDENTIFICATION SYSTEM OF MUSICAL TONES

Bertini Graziano(*), Di Salvo Vincenzo Giovanni(**)

(*) ISTI-CNR
(**) MOSART Project Young Researcher

INDEX

September 2003

Abstract

The aim of this paper is to present a system for the automatic identification of musical tones from a monophonic music melody in progress, among available alternatives of a library of the same previously recorded tones.

This study was born by the demand to have a tool based on a comparison criterion to measure the fidelity of reproduction of some musical tones during a musical piece in execution.

The algorithm is realized in two main distinct steps. At first, digital processing techniques are used with the purpose to obtain a pattern vectors from the original waveform. Thanks to the analysis techniques of Short-Time Fourier Transform, it has been possible to extract these patterns, so that they could to reflect precise energy-dependent features of the original signal, relevant to the identification.

This resulting patterns are, subsequently, elaborated using the Theory of the Least Squares Optimal Filtering. The Least Squares Criterion is, here, regarded as purely deterministic, that is there is no presumed knowledge of the statistical properties.

Therefore the algorithm has several desirable features. There is no upper limit on the frequency search range, so the algorithm is suited for high-pitched tones. The algorithm is relatively simple and may be implemented efficiently and with low latency on DSP processors.

A preliminary investigation of the problem was developed in cooperation between the Norwegian University of Science and Technology of Trondheim and the National Research Council of Pisa, in Italy, in the framework of a stage at NTNU within the European Project "Mosart" (2003).

* * *

1. Introduction.

It is well known that sounds of a musical instrument are listened and identified on the basis of musical characteristic attributes, like loudness, pitch, timbre.

The task to build a mathematical model, deciding which physical parameters are suitable to be identified by an automatic recogniser and how to combine them is not straightforward.

It is important to estimate attributes that reflect unique characteristic of musical signal, that are more measurable and easy to identify by an automatic identification machine.

This cognitive task is performed by high-skilled human musicians.

The solution proposed in this article attempts to recognize boundaries corresponding to the presence of the individual tones in a given melody in progress.

The work is divided in two main parts:

- To analyse the energy-dependent features of the corresponding waveform of the musical tones, using the spectrographic method of Short-Time Fourier Transform.
- To identify the musical tones.

From a careful study, based on the Short-Time Fourier Transform, the authors have decided to define a model based on the amplitude and frequency deviation of the energy samples available by a spectrographic analysis of the tones.

The choice of this strategy has been confirmed by the fact that the human ear is sensitive to the energy changes of the acoustic field near it.

The extracted *deviation pattern vectors* is compared among available alternatives of a library of previously recorded sounds, using the *.Least Squares Criterion.*

The solution proposed in this article has been developed considering the tones of three wind instruments, flute [1,2], clarino, oboe.

The source material analysed in this article is anechoic monophonic.

The characteristics of the musical tones analysed are shown in fig.1, fig. 2, fig. 3.

2. Preliminary Theoretical Issues.

2.1. – The Short-Time Fourier Transform [3].

A visual inspection of a musical signals shows that the properties of the sound waveforms change markedly as a function of time.

Looking at fig.1, relative to a flute tone, it is possible see that:

- the waveforms changes between musical regions and plosive[1] events.
- there is a significant variation in the peak amplitude of the sound (envelope evolution).

This time-variation characteristics correspond to highly fluctuating spectral features over the time. A single Fourier Transform of the signal cannot capture this time-varying frequency content (non-stationary signal).

The Short-Time Fourier Transform is a valid tool for revealing the time-variation of the frequency.

The Short-Time Fourier Transform is given by

► $$\mathbf{X}(f, t_0) = \sum_{n=-\infty}^{\infty} s(n, t_0) \exp^{-j2\pi fn}$$

where

$$s(n, t_0) = s(n)\, w(n, t_0)$$

and $w(n,t_0)$ represent a temporal window centred at $t = t_0$ and $s(n, t_0)$ the windowed sound segment.

A peculiarity of the Short-Time Fourier Transform, is that the Square Magnitude, given by

► $$\boldsymbol{\Psi}(f, t_0) = |\, \mathbf{X}(f, t_0)\, |^2$$

can be thought of as a two dimensional (2-D) energy-density. As we move from plosive events to tone sounds, this energy-density describes the relative energy content in frequency at different times.

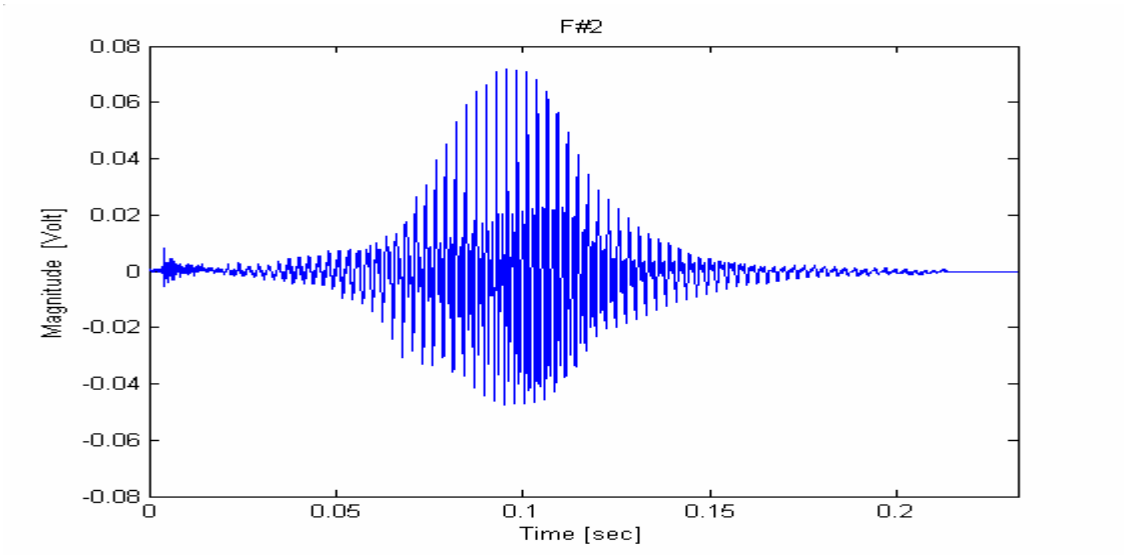
A **spectrogram** is a graphical display of the square magnitude $\boldsymbol{\Psi}(f, t_0)$.

It shows the trend of the frequency content of the signal over time.

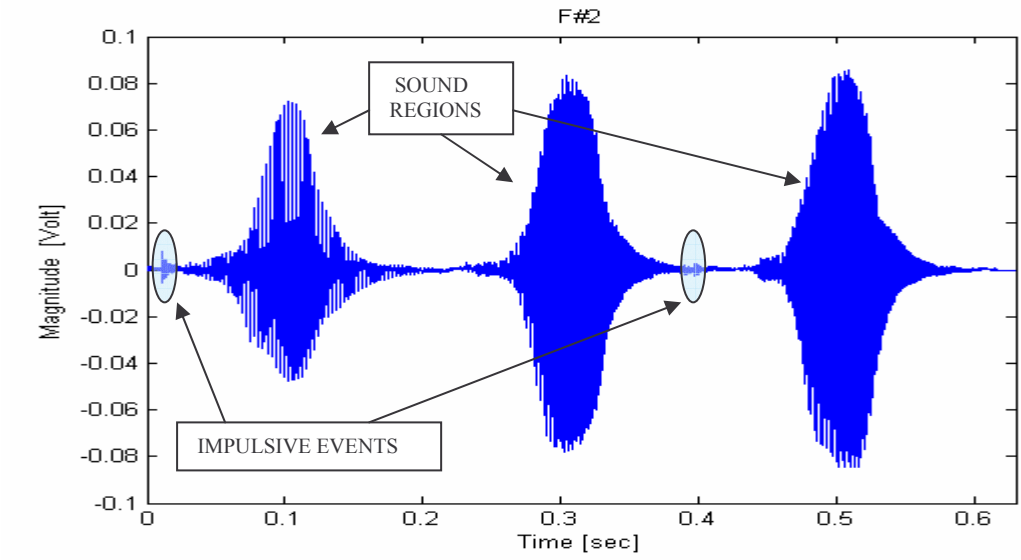
The possibilities to choose a window $w(n, t_0)$ among different shapes (Hamming, Hanning, Blackman, etc.), to choose the window $w(n, t_0)$ short in time (good time resolution) or longer in time (good frequency resolution) enables us to achieve the objectives in the best possible way [2].

[1] The plosive events can be a result of imperfect closing of the flute key. Note that plosive waveforms are characterized by high frequency content.

[2] Note: Time-Frequency objectives cannot be met simultaneously due to Heisenberg's uncertainty principle.



(a)



(b)

Fig. 1 – Flute Tone Signals

Two Flute anechoic tones have been analysed.
The first is the flute tone F#2 (~740 Hz);
The second is the flute pattern tones F#2, G2, A2, (~740 Hz) (~784 Hz) (~880 Hz)
The recording is monophonic (single channel), the sound signals are sampled at F_C = 44.1 KHz and quantized on Q_L = 16 bits (quantization level).
The size of the first file containing the tone F#2 has LENGHT = 9411 [Sample] so the time duration of the tones is D_T = LENGHT / F_C ≅ 213 msec.
The size of the file containing the three tones has LENGHT = 27159 [Sample] so the time duration of the tones is D_T = LENGHT / F_C ≅ 616 msec.

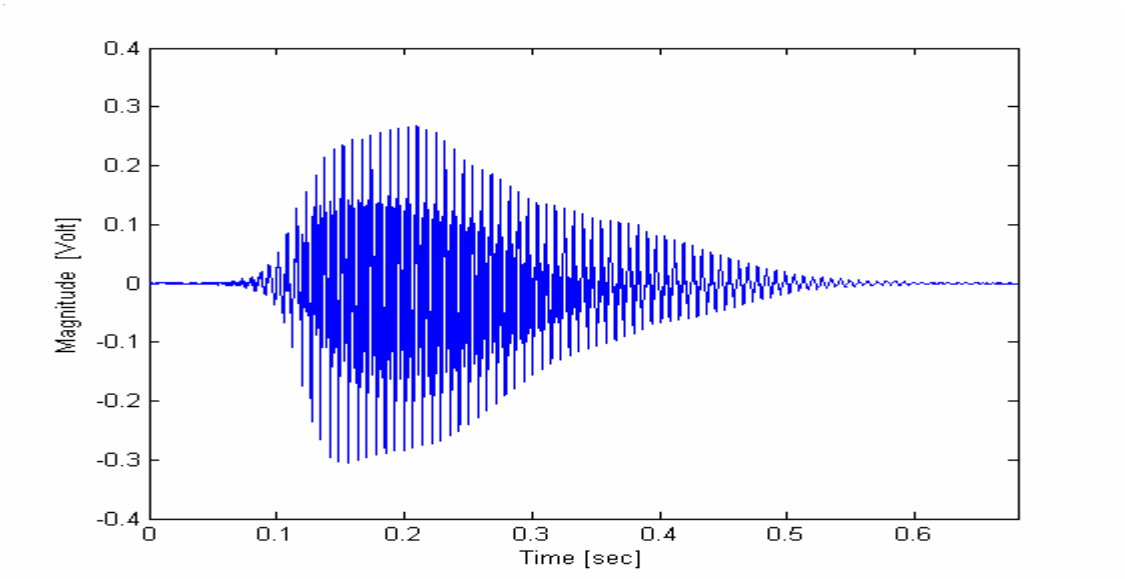


Fig. 2 – Clarino Tone Signal

One Clarino semi-anechoic tones have been analysed.
It is the clarino tone .... (~100 Hz);
The recording is monophonic (single channel), the sound signals are sampled at F_C = 44.1 KHz and quantized on Q_L = 16 bits (quantization level).
The size of the first file containing the tone .... has LENGHT = 30000 [Sample] so the time duration of the tones is D_T = LENGHT / F_C ≅ 680 msec.

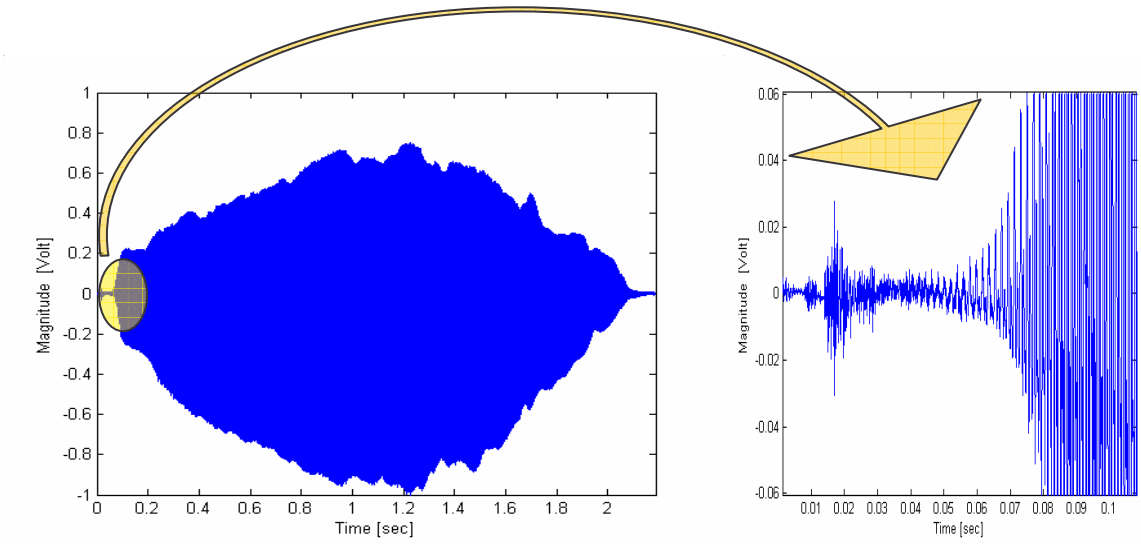


Fig. 3 – Oboe Tone Signal

One Oboe semi-anechoic tones have been analysed.
It is the oboe tone (~ 560 Hz);
The recording is monophonic (single channel), the sound signals are sampled at F_C = 44.1 KHz and quantized on Q_L = 16 bits. (quantization level).
The size of the first file containing the tone .... has LENGHT = ….. [Sample] so the time duration of the tones is D_T = LENGHT / F_C ≅ …… msec.

2.2. – Least Squares Optimal Filtering [4].

Optimal filtering deals with design of filters to process a class of signals with statistically similar characteristics (stochastic process). The topic is based on mean-square estimation as applied to signal processing.

There are various forms of optimal filtering. Linear prediction is a simple kind of optimal filtering problems. A more general type of optimal filtering is known as Wiener filtering.

The Optimal Filtering Theory is well lent to solve *system identification design* problems. In this case is required to make an absolute identification among N elements of a population [5].

The problem to estimate a desired sequence *s(n)* from a related sequence *x(n)* is essentially of the form:

$$\textit{choose the event } x_i(n) \textit{ such that } \boldsymbol{p}_i(x) > \boldsymbol{p}(s)\ ; \qquad i = 1,2,\ldots\ldots,N$$

In this section we consider the optimal filtering problem from a slightly different point of view. The problem is here regarded as purely deterministic. There is no presumed knowledge of the statistical properties beforehand.
It is assumed that a typical data sequence for both *s(n)* and *x(n)* has been measured and recorded and that these sequences can be used to design the filter.

If a causal FIR filter of length P is used, then the estimate for the given data sequence is

$$\hat{s}(n) = \sum_{k=0}^{P-1} h(n)\, x(n-k) \qquad (1)$$

and the error can be defined as

$$\varepsilon(n) = s(n) - \hat{s}(n). \qquad (2)$$

The approach here is to design the filter to minimize the sum of squared error

$$\boldsymbol{S} = \sum_{[nI,\, nF]} |\varepsilon(n)|^2 . \qquad (3)$$

where n_I and n_F are some initial and final values of *n* over which the minimization is performed.

The criterion (2) is called a Least Squares Criterion and could be regarded as purely deterministic.

Differentiating (3) [3] one finds the taps of the FIR filter:

$$\mathbf{h} = \mathbf{X}^{+}\mathbf{s} \qquad (4)$$

where

$$\mathbf{X}^{+} = (\mathbf{X}^{*T}\mathbf{X})^{-1}\mathbf{X}^{*T} \qquad (5)$$

is known as the *Moore-Penrose Pseudoinverse* matrix.

[3] The proof is not shown here.

3. Analysis and Results.

In the aim to find attributes that reflect unique characteristics of the musical tones the mathematical model adopted is based on two quantities:
Loudness Stability and **Tone Frequency Stability** (fig. 4).

The Loudness Stability refers to the energy variations of each harmonic of the stochastic process obtained with the Short-Time Fourier Transform.

The Tone Frequency Stability refers to the frequency deviation that occurs between two consecutive harmonics of the stochastic process obtained with the Short-Time Fourier Transform.

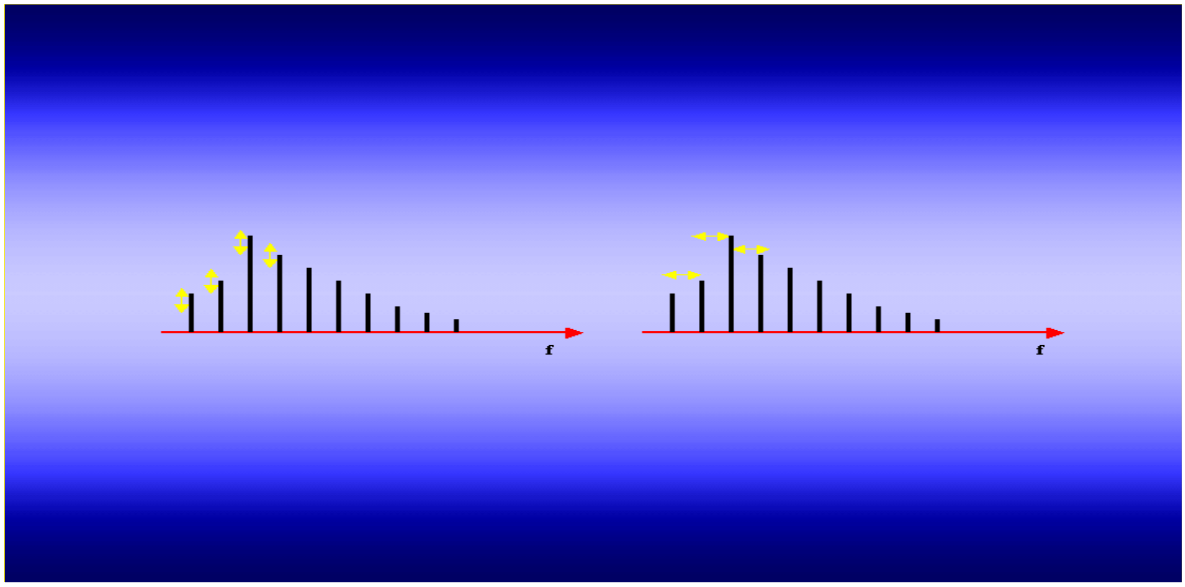


Fig. 4 – Meaning of *Loudness Stability* and *Tone Frequency Stability*.
In the aim to find attributes that reflect unique characteristics of the musical tones two quantities has been analysed:

- **Loudness Stability**
- **Tone Frequency Stability.**

The Timbre Stability refers to the energy variations of each harmonic of the stochastic process obtained with the Short-Time Fourier Transform.
The Tone Frequency Stability refers to the frequency deviation that occurs between two consecutive harmonics of the stochastic process obtained with the Short-Time Fourier Transform.

To measure the Timbre Stability of one pitch, a good choice seemed to consider the values along the row's elements in the spectrographic matrix (fig. 5).

The Tone Frequency Stability can be measured as the phase differences between two consecutive column elements (fig. 6).

The frequency analysis of the signals has been developed in Matlab. The Matlab code (Appendix A) calculate the *Spectrograms of the Musical Tones*, the *Strong Foundamental Frequency*, shows the *Analysis of the Timbre Stability* with the Short-Time Fourier Transform and with the Tristimulus Method [6] and the, the *Analysis of the Frequency Stability* with the Short-Time Fourier Transform.

The spectrograms of the musical tones are shown in fig. 7-9.

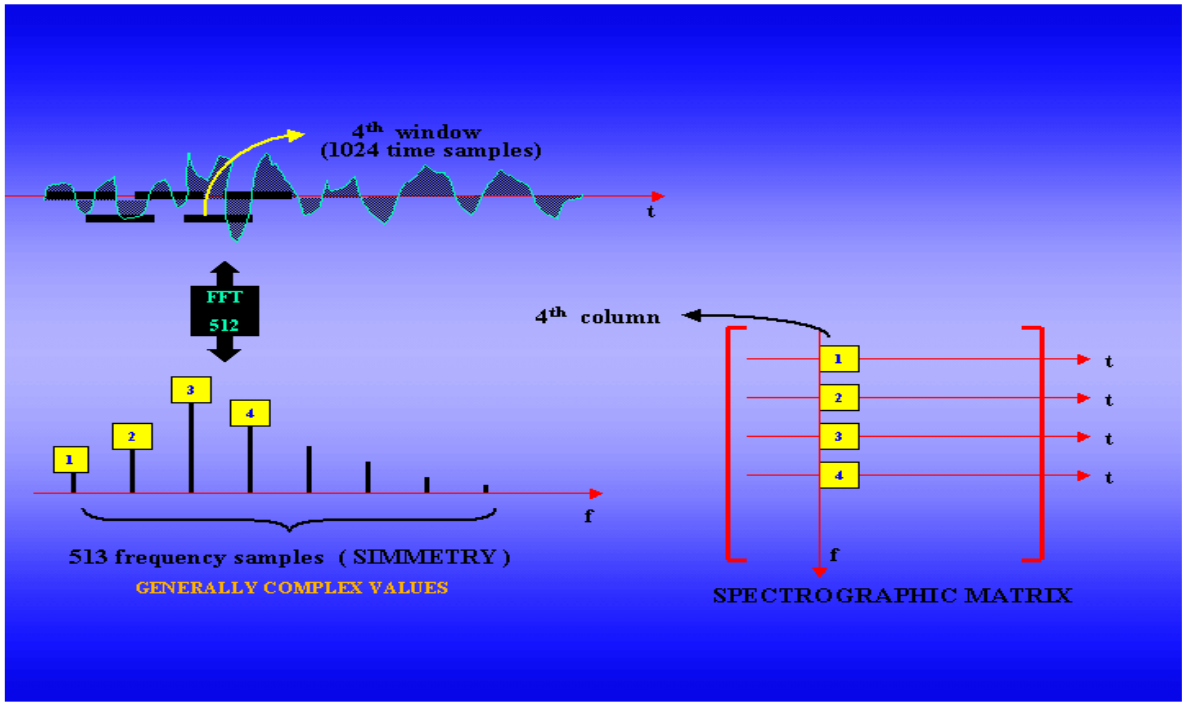


Fig. 5 – Loudness Stability on the spectrographic matrix

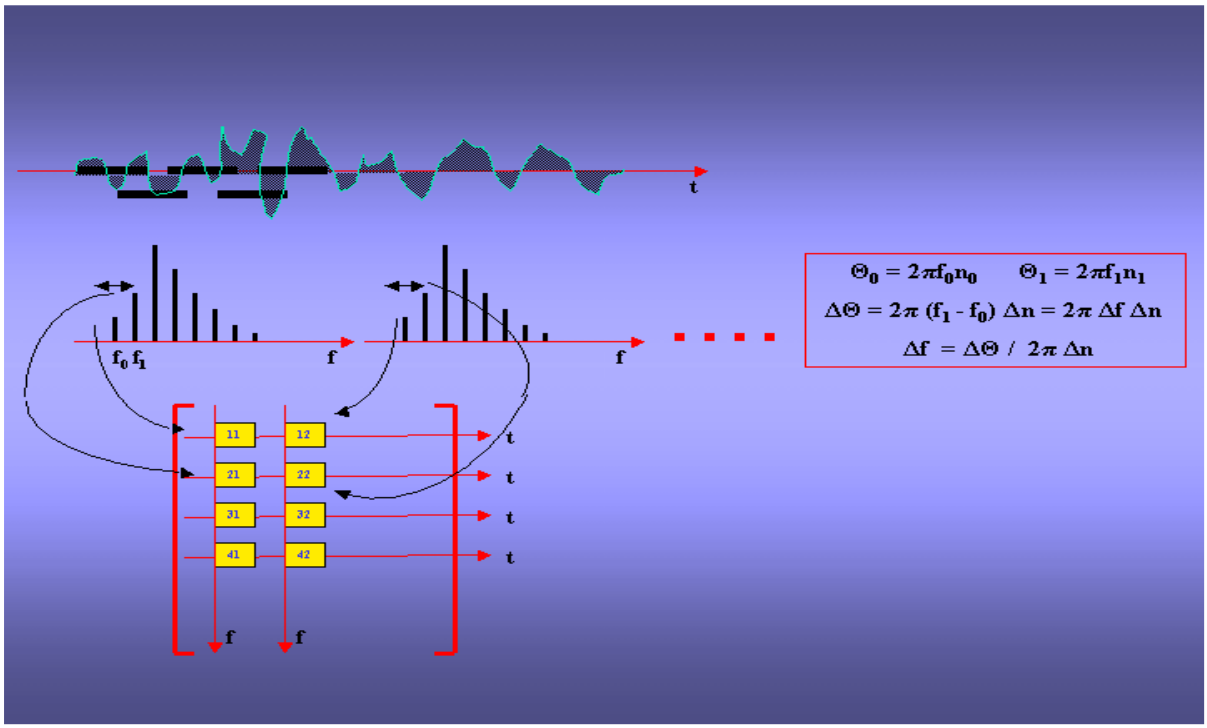
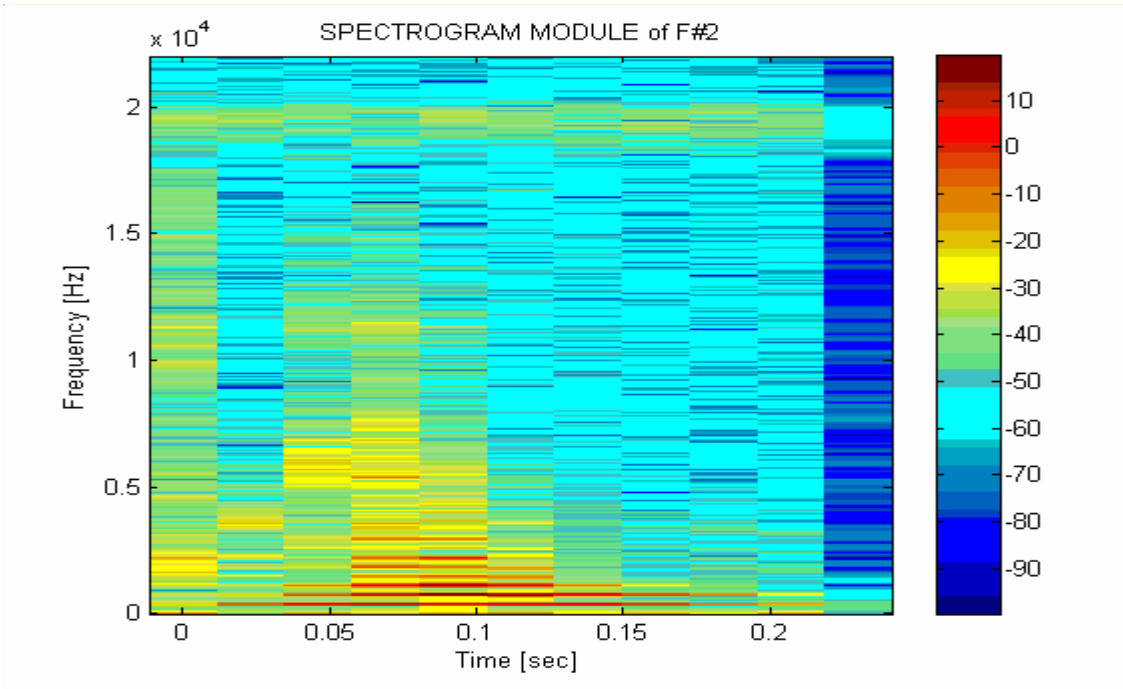
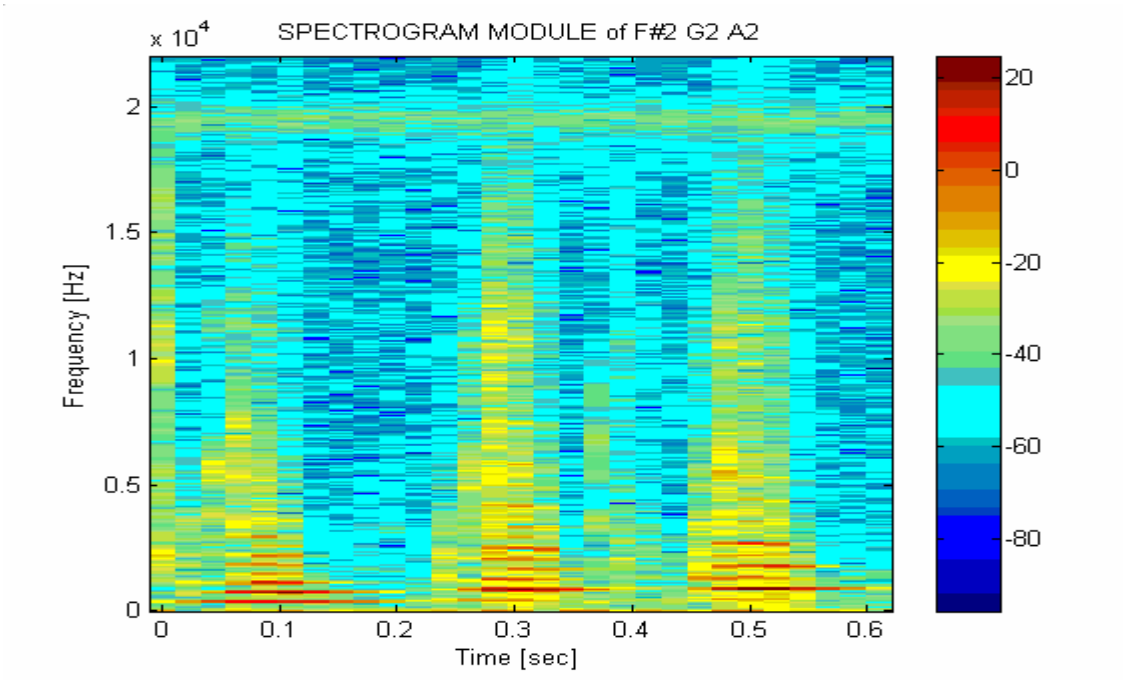


Fig. 6 – Frequency Tone Stability on the spectrographic matrix



(a)



(b)

Fig. 7 – Flute Tones Spectrograms

RESULTS : best results has been obtained using an Hamming Window of 1024 sample (~23 sec. @ 44.1 KHz), an overlap of 10 % and an FFT of 1024 sample. This choices guarantees that both the Timbre Stability and the Tone Frequency Stability are quite constant long the entire duration of the signals. An energetic measure showed that the first harmonic is the energy-strong fundamental harmonic.

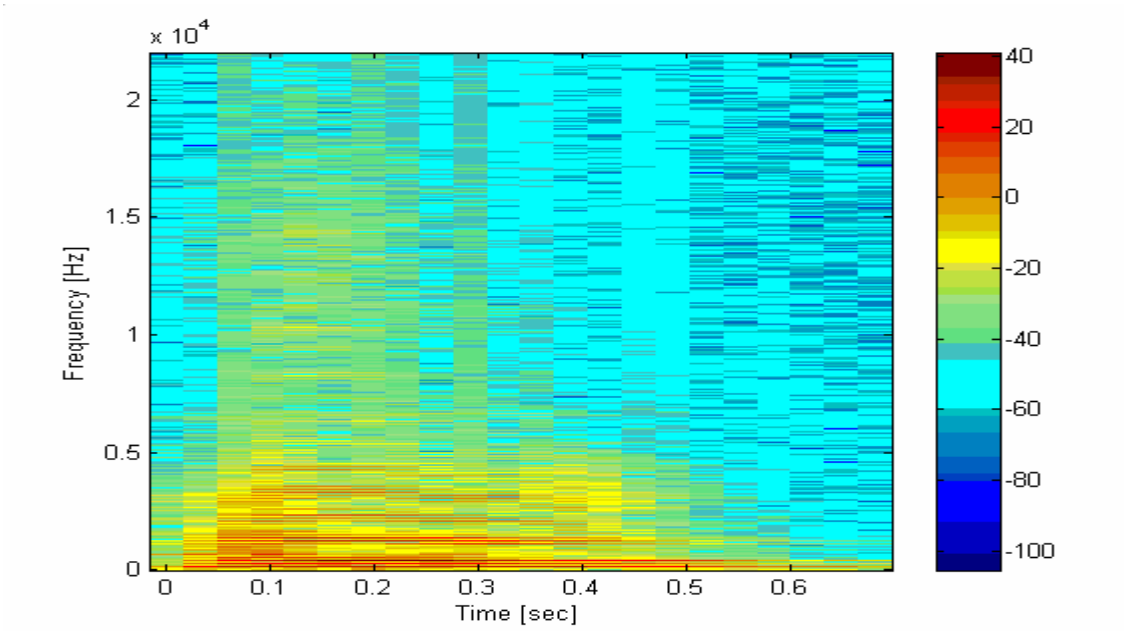


Fig. 8 – Clarino Tone Spectrograms

RESULTS : best results has been obtained using an Hamming Window of 1024 sample (~23 sec. @ 44.1 KHz), an overlap of 70 % and an FFT of 1024 sample. This choices guarantees that both the Timbre Stability and the Tone Frequency Stability are quite constant long the entire duration of the signals. An energetic measure showed that the first harmonic is the energy-strong fundamental harmonic.

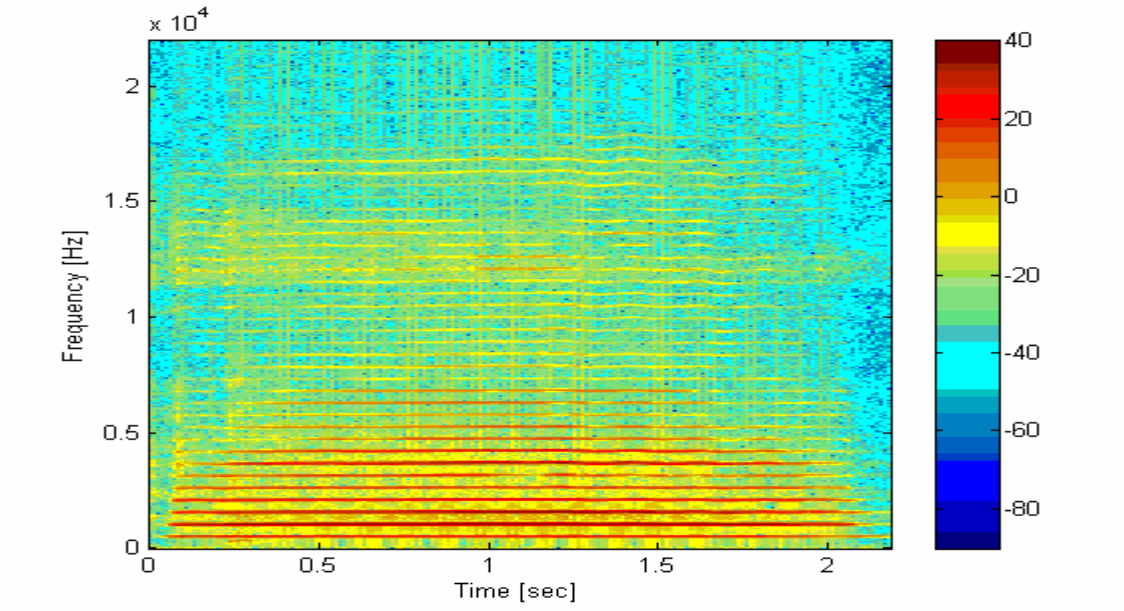


Fig. 9 – Oboe Tone Spectrograms

RESULTS : best results has been obtained using an Hamming Window of 1024 sample (~23 sec. @ 44.1 KHz), an overlap of 50% and an FFT of 1024 sample. This choices guarantees the Timbre Stability. The Frequency Stability improves to increase the FFT by points which is calculated up. An energetic measure showed that the first harmonic is the energy-strong fundamental harmonic.

It is important to observe how the signals are quite stationary. This result suggests recurring LMS algorithm to make the automatic recognition. This algorithm, in fact, assures a good convergence degree[4] if the signal is slowly changing time. that is the eigenvalues of the signal autocorrelation matrix are similar.

4. A proposal solution for the identification system.

A general representation of a system of sound recognition is shown in fig. 10.

As seen in this figure a representation of the sound signal is obtained using digital signal processing techniques which preserve the features of the sound signal that are relevant to tones identity.

The resulting pattern is compared to a reference database (library) of same previously prepared tones, using the LMS Algorithm [7] (see fig.11). A subsequent decision logic is used to make a choice among the available alternatives [8].

This section proposes the implementation of the automatic identification system relative to flute musical tones from a flute melody in progress. This is depicted in fig. 12.

As seen in this figure, *digital processing techniques* are the first step of the system.

In the light of the results obtained in the analysis shown in previous sections, the flute melody incoming, s(t), has been divided in Hamming windows of 10244 samples and an overlap of 80%.

A length of 10244 samples has been chosen because better match with the length of the signal of the library.

The windows are preliminary stored in a memory buffer of the DSP so to be available for the subsequent computation and also to allow the real-time elaboration.

For each frame of 10244 sample the spectrogram is computed using Hamming windows of 1024 sample, an overlap of 10% and FFT of 1024 sample compute the spectrogram.

The spectrogram input to a *features extraction block* so to extract the Loudness Stability of each frame.

At this point the objective is to decide which note model, from a known set of tone models (library), best correlate the actual note.

[4] The measure of the degree of convergence of the LMS algorithm consists in verifying that the eigenvalues of the autocorrelation matrix are similar.

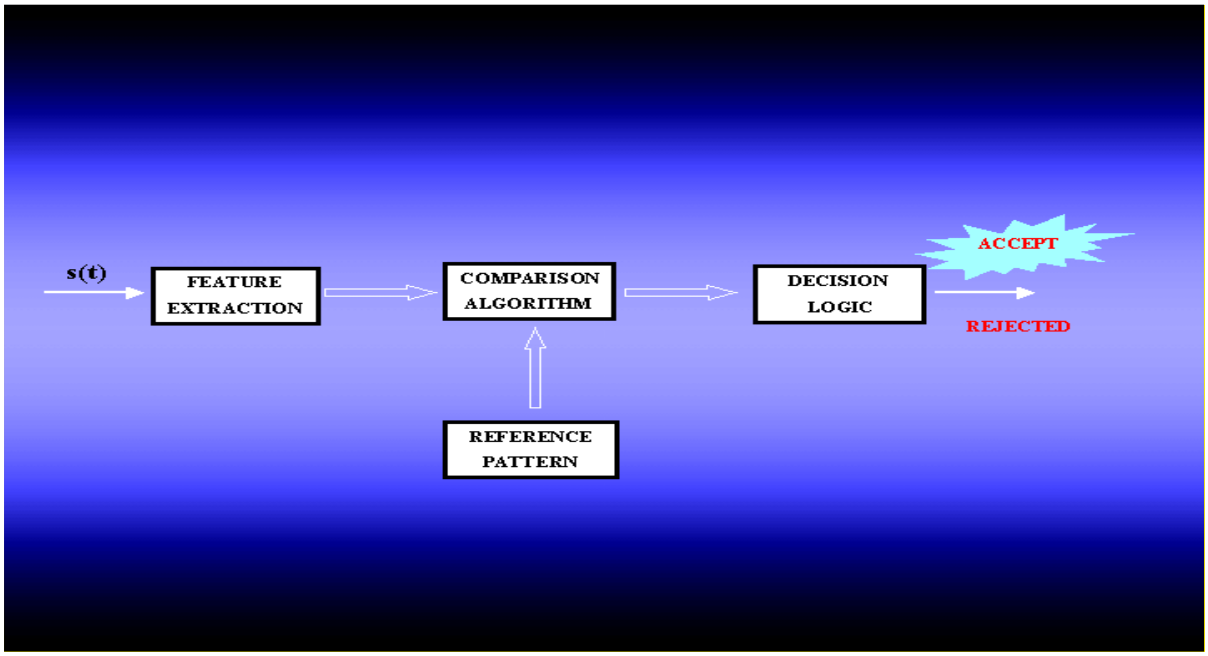


Fig. 10 - General Representation of a sound recognition system

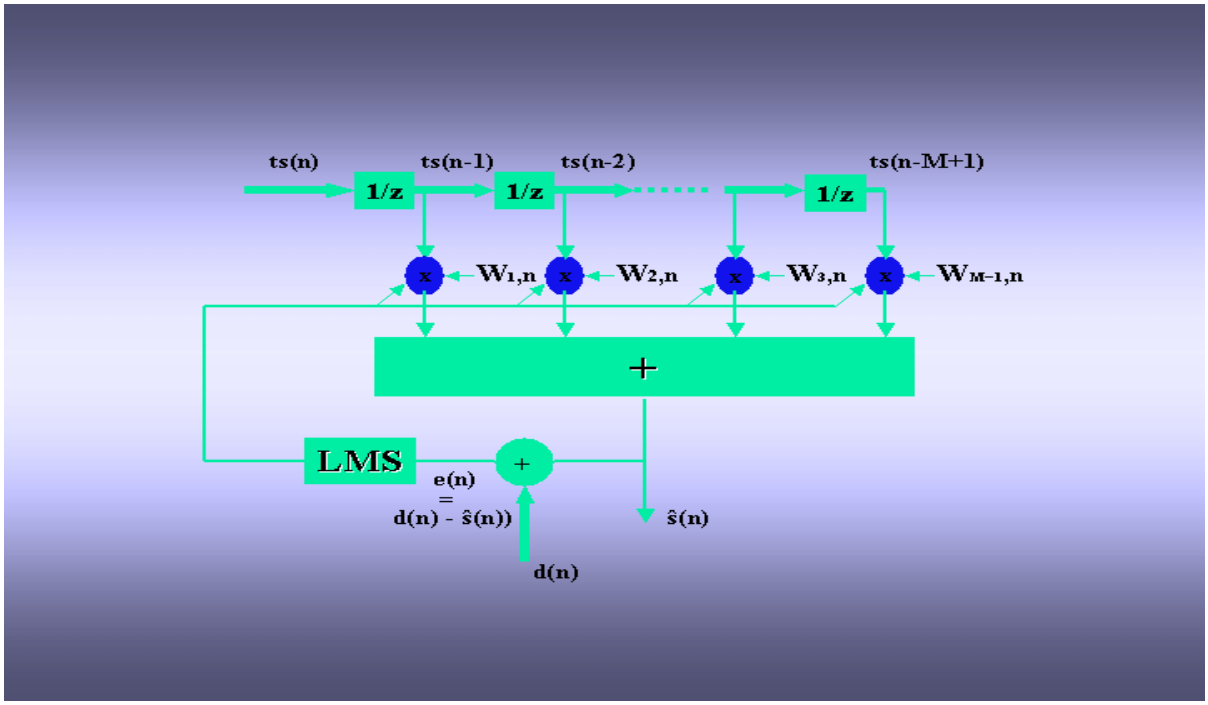


Fig. 11 - The Least Square Optimal Filtering

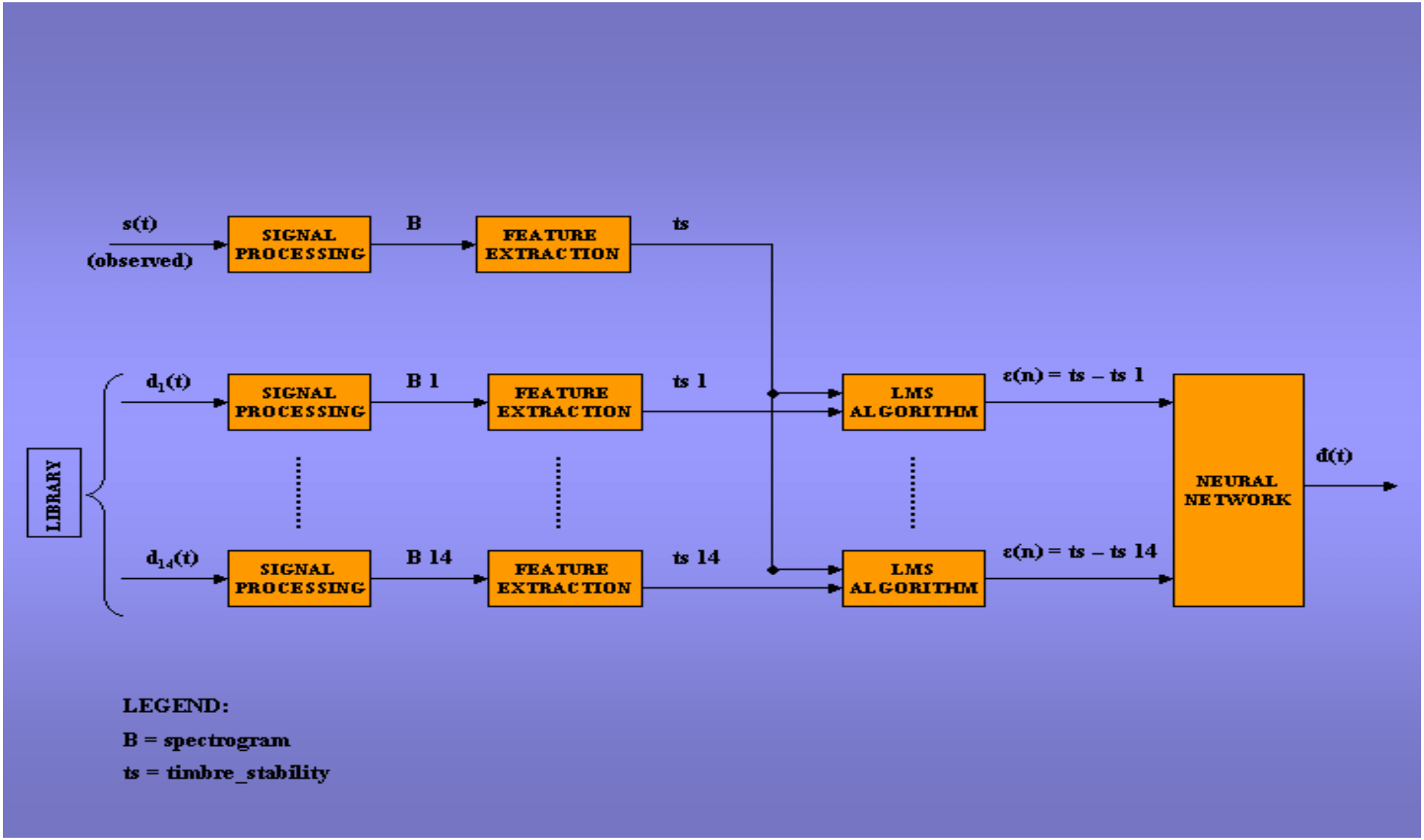


Fig. 12 - Representation of the proposed sound recognition system

4. Conclusion and Future Work

Audio identification system is still an active research area.
A least squares optimal filter has been used to estimate the Loudness Stability.
Much work needs to be done in different directions to optimise the present project.
The database of the reference tones can be easily implemented as multi-array form so to extend the identification on the entire scale of musical tones.
Improving the LMS algorithm for the detection of flute notes for signals recorded in normal or reverb chamber.
At the end, the identification strategy developing pre-processing techniques for the noise suppression can be optimised by an estimation of the background noise level and filtering of the noisy sources from the sound regions.

5. Thanks

The Authors wish to tanks the Prof. Jan Tro, Prof. Ulf Kristiansen and the Ph.D. students Magne Arthur Larsen, Bård Støfringsdal of the Acoustic Group of the Department of Telecommunications of the NTNU for their precious help and good advices.

6. References

[1] Sølvi Ystad – Identification and Modeling of a Flute Source Signal, Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99), NTNU, Trondheim, December 9-11, 1999
[2] http://www.phys.unsw.edu.au/
[3] Quatieri - Discrete Time Speech Signal Processing. Prentice-Hall.
[4] Therrien Charles W. - The Lee-Wiener Legacy. A History of the Statistical Theory of Communication. IEEE Signal Processing Magazine. November 2002.
[5] L.R.Rabiner/ R.W.Schafer - Digital Processing of Speech Signal, Prentice-Hall
[6] Pollard H.F. / Jansson E.V. – A Tristimulus Method for the Specification of Musical Timbre. Acoustica, Vol. 51;162-171
[7] Baher - Analog and digital signal processing; John Wiley & sons
[8] W.S.Hodgkiss / J.A.Presley – Adaptive Tracking of Multiple Sinusoids Whose Power Levels are Widely Separated, IEEE TRANSACTION ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, vol. ASSP 29, NO.3, June1981

APPENDIX A

```
%========================================================== %
%          NTNU- Norwigian University of Science and Technology          %
%                        MOSART PROJECT 2003                             %
%                                                                        %
%            FREQUENCY ANALYSIS OF MUSICAL TONES                         %
%                                                                        %
%   Developed by: Vincenzo Di Salvo   ( Vincenzo.DiSalvo@isti.cnr.it)    %
%                                                                        %
%   Supervision: Prof. Jan Tro ( Jan.Tro@tele.ntnu.no )                  %
%========================================================== %
```

% LOAD DATA FOR FILE

```
[y,Fs] = wavread('filename.wav');
%LENGTH = length(y);
%y = y';
%y = [y zeros(1,27762 - LENGTH)];
LENGTH = length(y);   % lenght of the signal (sample)
DT = LENGTH/Fs;       % lenght of the signal (seconds)
```

% MAKE SIGNAL PLOT

```
figure;
plot((1:LENGTH)/Fs,y)   % plot the signal in time domain
axis([0 DT -1 1]);
%title ('F#2 G2 A2')
xlabel('Time [sec]')
ylabel('Magnitude  [Volt]')
```

%%%%%%%%%% SPECTROGRAPHIC ANALYSIS %%%%%%%%%%%%%

```
% abs(fft(y);
% plot((0:floor(LENGTH/2-1)),Y(1:floor(LENGTH/2)))
```

% SET SPECTROGRAM PARAMETERS

```
    WINDOW_LENGTH = input('Give the WINDOW_LENGHT (in sample): ');
    OVERLAP_STEP = input('Give the OVERLAP_STEP: ');
    OVERLAP = round(WINDOW_LENGTH * OVERLAP_STEP);
    NFFT = input('Give the NFFT_LENGTH: ');
        WINDOW = hamming(WINDOW_LENGTH);

    % ATTENTION - must be: OVERLAP < WINDOW_LENGTH <= NFFT
```

% COMPUTE SPECTROGRAM

```
        [B,f,t] = specgram(y,NFFT,Fs,WINDOW,OVERLAP);
    % ATTENTION - B is a complex values matrix
```

% MAKE SPECTROGRAM PLOT

```
        figure;
        imagesc(0:.010:DT, 0:1000:(Fs/2), 20*log10(abs(B)));
    colormap('jet')
        axis xy
    colorbar('vert');
    %title('SPECTROGRAM MODULE')
        xlabel('Time [sec]')
        ylabel('Frequency [Hz]')
```

% FIND THE STRONG FOUNDAMENTAL HARMONIC

```
for ROW = 1 : 6
   LOUDNESS = B(ROW,:);
   LOUDNESS = abs(LOUDNESS);
   % LOUDNESS = (1/WINDOW_LENGTH)*(abs(LOUDNESS)).^2;
   LOUDNESS_MEAN = mean(LOUDNESS(:))
end
```

%%%%%%%%%%%%%% STABILITY ANALYSIS %%%%%%%%%%%%%%

% TIMBRE STABILITY

```
for ROW = 1 : 6
   LOUDNESS = B(ROW,:);                             %% make the stochastic variable
   LOUDNESS = (1/WINDOW_LENGTH)*(abs(LOUDNESS)).^2;    %% LOUDNESS (in the energy domain)
   LOUDNESS = LOUDNESS';                               %% corresponding to each tone

   LOUDNESS_MEAN = mean(LOUDNESS(:));

   LOUDNESS_STABILITY = LOUDNESS - LOUDNESS_MEAN;
   figure;
   plot(t,LOUDNESS_STABILITY)
   axis([0 DT -0.02 0.02]);
   title(['LOUDNESS STABILITY CORRESPONDING TO ', num2str(ROW), ' TONE']);
end
```

% TIMBRE STABILITY WITH THE TRISTIMULUS METHOD

```
first_group = input('Give the number of the first partials: ');
second_group = input('Give the number of the second partials: ');
terzo_group = input('Give the number of the terze partials: ');

for COLUMN = 1 : fix((LENGTH-OVERLAP)/(WINDOW_LENGTH-OVERLAP))

   group1 = abs(B(1 : first_group,COLUMN));
   Nmax = max(group1);
   Nequiv1 = 0.85 * Nmax + 0.15 * sum(group1);

   group2 = abs(B(first_group + 1 : first_group + second_group,COLUMN)) ;
   Nmax = max(group2);
   Nequiv2 = 0.85 * Nmax + 0.15 * sum(group2);

   group3 = abs(B(first_group + second_group + 1 : first_group + second_group + terzo_group,COLUMN));
   Nmax = max(group3);
   Nequiv3 = 0.85 * Nmax + 0.15 * sum(group3);

   N = Nequiv1 + Nequiv2 + Nequiv3;

   X(COLUMN) = Nequiv3 / N;
   Y(COLUMN) = Nequiv2 / N;
end

figure;
plot(X,Y);
axis([0 1 0 1]);
title('TRISTIMULUS METHOD');
```

% FREQUENCY TONE STABILITY

```
B_PHASE = angle(B);

for ROW = 1 : 6
   for COLUMN = 1 : fix((LENGTH-OVERLAP)/(WINDOW_LENGTH-OVERLAP))
      HARMONIC_DELTA_PHASE(COLUMN) = abs(B_PHASE(ROW + 1,COLUMN) - B_PHASE(ROW,COLUMN));
      DELTA_T(ROW) = t(ROW + 1) - t(ROW);
      FREQUENCY_DEVIATION(COLUMN) = HARMONIC_DELTA_PHASE(COLUMN)/(2*pi*DELTA_T(ROW));
   end
      FREQUENCY_DEVIATION_MEAN = mean(FREQUENCY_DEVIATION);
      figure;
      plot(t,FREQUENCY_DEVIATION,'b-',t,FREQUENCY_DEVIATION_MEAN,'m*');
      axis([0 DT -200 200]);
      title(['FREQUENCY DEVIATION BETWEEN ', num2str(ROW),' E ', num2str(ROW + 1) ' TONE']);
      FREQUENCY_DEVIATION_STD = std(FREQUENCY_DEVIATION);
end
```

Bertini Graziano is senior researcher by the Istituto di Scienza e Tecnologie del CNR in Pisa. His research interests concern the audio signals numeric processing in the speech and musical domains with special attention to DSP based applications

Di Salvo Vincenzo Giovanni graduated in Telecommunications Engineering at the University of Pisa. He is currently active in the research group of the Signals and Images Labs of the Istituto di Scienza e Tecnologie della Informazione del CNR in Pisa. His interests are addressed to Methodologies of Analysis and Synthesis of Telecommunications Signals.