CENTRE FOR VISION, SPEECH & SIGNAL PROCESSING (CVSSP)
UNIVERSITY OF SURREY

Transfer Report

from MPhil to PhD

Acoustic object modeling for content based retrieval from multimedia databases

Ioannis Paraskevas

September 2001

ISBN-13: 978-0-907311-40-9

List of figures

Session 1

1.1) Introduction

Audio is a significant type of media and forms an important part of audiovisual databases. The increasing number of digital databases has enforced the necessity of their effective management, based on the analysis of their audio content. The audio content analysis describes the computerized understanding of the semantic meanings of an audio document. Methods that materialize such an automated analysis are indispensable for the efficient access, digest and information retrieval. In a multimedia document, its semantics are embedded in multiple forms that are usually complimentary of each other. There is therefore a need to analyse all the data types: sound tracks, image frames, spoken words or even text that can be extracted from the image frames.

The automated analysis briefly described, requires the segmentation of the document into semantically meaningful units, and the classification of each unit into a predefined category. The majority of the current research approaches are focused on the visual information of the databases, which leads to a far too fine segmentation of the audiovisual sequence with respect to the semantic meaning of the data. The integration of all the multimedia components (audio, visual and textual information) will lead to a fully functional system that will achieve effective information retrieval from the databases. While much effort has been made in the area of visual recognition, general audio recognition has had less attention and it is considered that focussing on this would render some benefit to the general problem of archival retrieval. Some of the applications of audio segmentation and classification are the following: audio archive management, professional media production, commercial music usage, database surveillance, video annotation and so on.

The current research approaches to this area, on feature extraction and classification of audio sounds, is relatively new compared to speech recognition. However, the following is an overview of the most recent and relevant literature on audio classification. Rice [12] proposed a method in which a prototype audio clip is created, and then the aim is to find clips that sound like the original one. He used various sounds, such as: human sounds, animals, machinery, musical instruments, electronic tones and environmental sounds. Zhang [17] [18] proposed a method based on audio content analysis. According to this method, the audio signal was segmented and then classified into the following categories: speech, music, environmental sound and silence. Liu [14] proposed another method in which the feature extraction was based on the volume distribution, pitch contour and frequency related features. He categorized the audio signals in the following groups: news, weather reports, advertisements, football and basketball games. Foote [13] proposed a technique to retrieve audio clips by acoustic similarity. The feature extraction method was based on mel-scale cepstral coefficients while Li[19] proposed a method for content-based audio classification using a combination of perceptual and cepstral features. Pandit [21] also extended this work on audio recognition by using combinations of features. These included mel-scale cepstral coefficients, pitch value and zero crossing rate. As an alternative to mel-scale coefficients, LPC was used in other examples. Improvements in overall recognition rates were reported using these combinations.

In the work reported here, an alternative approach is described to the extraction of features for audio classification. There are two reservations to existing feature extraction methods that may limit the effectiveness of these audio recognition techniques. The first is that features such as LPC or cepstral coefficients were developed for speech recognition systems and were based on a model of the human speech production system. Such a model may not be appropriate for generalized audio classification problems. The second, is that such features tend to represent the magnitude spectrum of the acoustic unit and phase is under-represented though is partially included via time domain features such as zero-crossings and volume distribution. Here, an alternative approach is proposed where features are to be extracted from a spectral (frequency domain) representation but which include both magnitude and phase information. The approach firstly uses a suitable transform (such as the Fourier or Hartley) to generate a frequency-time surface or surfaces. A second transform is then applied to reduce the information present, so that it may be compactly presented to the chosen classifier. Since the surface is now in the form of an acoustic image, this second transform may be drawn from the image processing field and the Hough transform or DCT have been applied and results obtained. These are shown in section 3. The issues that are of concern are the types of transforms used and the manipulation of these to present the appropriate information to the classifier in a compact form.

The report is divided into three further sessions. The second one states the objective of the research and includes a review of the audio representation and analysis techniques. In the third an analytical description of the proposed transform-based method of representation and analysis is described. Finally, the fourth session consists of the future plans.

Session 2

Audio representation and analysis techniques

2.1) Overview of the system

Our system model can be divided, as for any other pattern recognition system model, into three stages: sensor, feature extractor and classifier. A schematic (Fig.1) that includes all the three stages and a brief theoretical description of each of them follows:

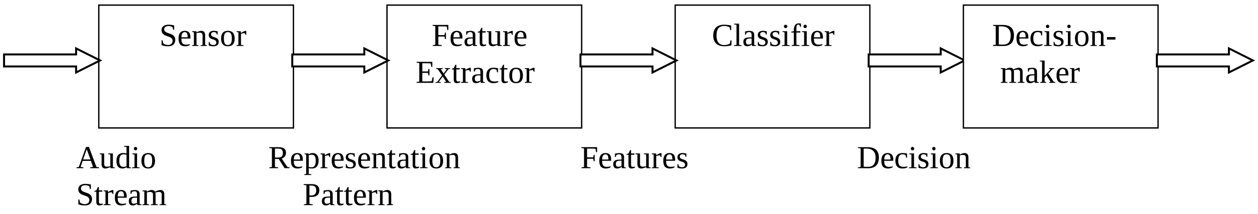


Fig.1 Pattern Recognition System Model

Sensor: its aim is to provide accurate representation of the audio stream to be classified. The performance limits of the system depend on it.

Feature extractor: it extracts the appropriate information from the representation pattern in order to reduce the dimensionality of the pattern recognition problem.

Classifier & Decision-maker: the last stage of the system, forms its decision making part. The classifier assigns patterns of unknown class membership to their appropriate categories [4].

2.2) Feature extraction of audio signals

Generally, the audio signal features used can be divided into three categories. The first one consists of the features that are related to the time domain of the signal and the second one includes features that are related to the signal's frequency domain. Features that belong to the first category (a) are: i) Short-time energy function, ii) Short-time average zero-crossing rate iii) Volume iv) Linear Prediction Coefficients to name but four.

Among the methods that belong to the second category (b) are: i) Pitch ii) Spectrogram iii) Frequency centroid iv) Bandwidth and v) Cepstral coefficients and Mel-frequency cepstral coefficients.

2.2) a) i) Short-time energy function

Definition/Mathematical expression:

The short-time energy function of a signal is defined as:

$E_n = \frac{1}{N} \sum_{m} [x(m)w(n-m)]^2$ where,

x(m): discrete time audio signal

n: time index of the short-time energy

w(m): rectangle window

$$\text{i.e. } w(m) = \begin{cases} 1, \text{ for } 0 \leq n \leq N-1 \\ 0, \text{ otherwise} \end{cases} \quad [17]$$

Comments:

The main reasons for using the short-time energy function are:

i) it is a convenient representation of the amplitude variation over the time

ii) for the special case in which the audio signal is speech, the values of E_n are in general, much smaller for the unvoiced components compared to the voiced ones.

iii) it can also be used in order to distinguish audible sounds from silence when the value of the SNR is high.

iv) the way the function varies over the time, may underline the rhythm and the periodicity of the sound.

2.2) a) ii) Short-time Average Zero-Crossing Rate

Definition/Mathematical expression:

In the context of discrete-time signals, a zero-crossing is said to occur if successive samples have different signs. The rate at which zero-crossings occur is a simple measure of the frequency content of a signal [17]. The short-time average zero-crossing rate is defined as:

$Z_n = \frac{1}{2} (\sum_{i=1}^{N-1} |sgn[s_n(i)]-sgn[s_n(i-1)]|) \frac{fs}{N} w(n-m)$, where, *fs*: sampling rate

w(n): rectangle window of length N

i.e. $$w(n) = \begin{cases} 1, \text{ for } 0 \leq n \leq N-1 \\ 0, \text{ otherwise} \end{cases}$$

and

i.e. $$sgn[x(n)] = \begin{cases} 1, & x(n) \geq 0, \\ -1, & x(n) < 0. \end{cases}$$

Figure 2 (fig.2) shows the time waveforms of commercial, news and sport clips. In the first one male speech over music background is recorded. The second one contains clean male speech, while the last one includes live broadcast from a basketball match. Figure 3 (fig.3) shows the corresponding curves of the average zero-crossing rate for commercial, news and sports clips respectively.

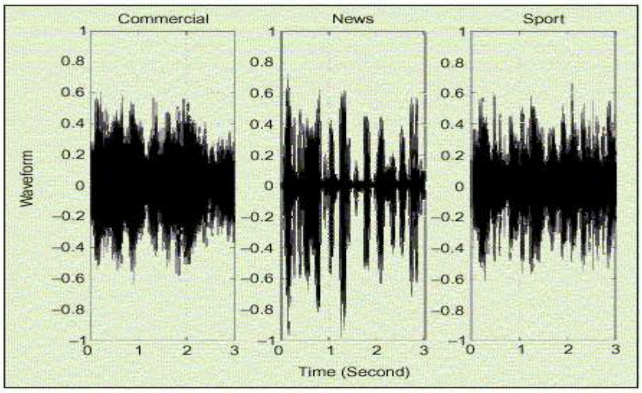


Fig.2 Time waveforms of commercial, news and sports clips

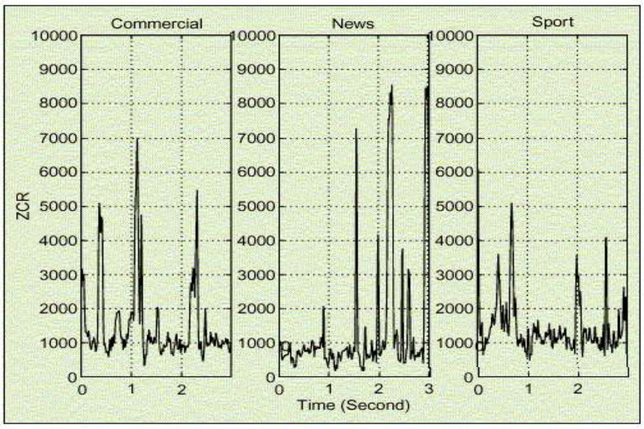


Fig.3 Average zero-crossing rate for commercial, news and sports clips

Comments:

i) The zero-crossing rate expression is used as a measure of discrimination between voiced and unvoiced speech. Generally, the unvoiced speech components have much higher zero-crossing rate values compared to the voiced ones. For example, the zero-crossing rate curve of the news clip consists of peaks and troughs because of the unvoiced and voiced components respectively. Therefore, the curve presents large variance and a wide range of amplitudes. Finally, in general terms, the zero-crossing rate curves are characterized by a relatively low and stable baseline with high peaks above it.

ii) The zero-crossing rate curves of the commercial and sport clips have a much lower variance and average amplitude compared to the curve of the news clip. The commercial clip has a relatively smooth curve since it has a music background, whereas the one of the sports clip is even smoother due to its noisy background. Generally, the zero-crossing rate curve of music clips is characterized by an irregular waveform with a changing baseline and a relatively small range of amplitudes.

iii) In order to distinguish other kind of audible sounds, based on the zero-crossing rate criterion, it is possible to use other characteristics of its curve such as regularity, periodicity, stability and amplitude range.

2.2) a) iii) Volume

Definition/Mathematical expression:

The volume (also referred as loudness) of a signal frame n is defined as:

$$v(n) = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} s_n^2(i)}$$

where,

s(i): discrete time audio signal

Figure 4 (fig.4) shows curves of volume for commercial, news and sports clips.

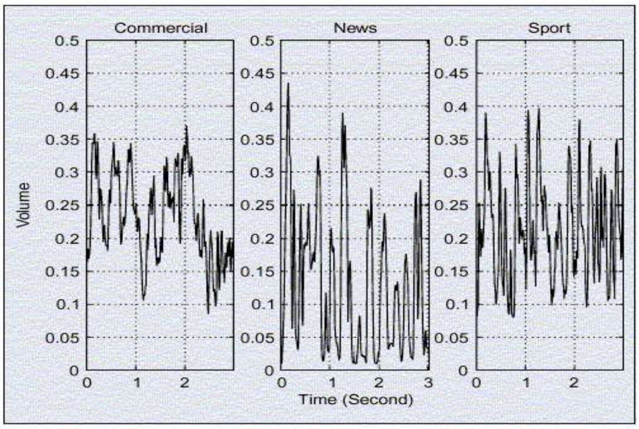


Fig.4 Volume or commercial, news and sports clips

<u>Comments</u>:

i) The mean and standard deviation of the volume of an audio signal can be used as descriptors [14].

ii) An audio frame can be characterized as silent or not, based on the comparison of its volume with a threshold determined by the volume distribution of the entire audio clip. After the silence is detected, it is possible to calculate the silence ratio, which is defined as the ratio of the silence interval to the entire period. The silence ratio varies according to the content of the audio clip. In news reports the silence ratio is higher compared to the commercial clips because in the first case there are regular pauses of the reporter's speech, whereas in the second case there is always some kind of background music.

iii) Volume can also be a helpful tool in the discrimination between voiced and unvoiced speech. Usually, unvoiced speech is characterized by low volume and high zero-crossing rate. So, by using both volume and zero-crossing rate, low energy unvoiced speech frames will not be misclassified as silence.

iv) The VDR (Volume Dynamic Range) is defined as: $\frac{\max(v)-\min(v)}{\max(v)}$ where max(*v*) and min(*v*) are the maximum and minimum volumes within an audio clip respectively. VDR, which is a measure of the variation of an audio clip's volume, does not change a lot in sports programs compared to news reports. This is explained, since in the first case there is usually a constant level of background sound so the volume does not change considerably, whereas in the second case, there are silent periods between speech, which result in a much higher VDR.

2.2) a) iv)Linear Prediction Coefficients

The linear predictive coefficients method has already being applied in speech analysis and coding. It is based on the digital model of a person’s vocal tract, which can be represented by the following transfer function: $H(z)=\frac{A}{1-\sum_{k=1}^{p}a_k z^{-k}}$

This system is excited, for voiced speech, by an impulse train, whereas for unvoiced speech it is excited by random white noise. From the last equation it is deduced that the relation between the audio samples x(n) and the excitation $\delta(n)$ is given by the following difference equation: $x(n)=\sum_{k=1}^{p}a_k x(n-k)+A\delta(n)$. Assuming that the signal is processed by a linear predictor: $x_p(n)=\sum_{k=1}^{p}a_k x(n-k)$.So, the predictor error can be estimated using the following equation: $e(n)=x(n)-x_p(n)=x(n)-\sum_{k=1}^{p}a_k x(n-k)$

The predictor error will be minimum, if $\sum_{k=1}^{p}a_k x(n-k)$of *x(n)* and $x_p(n)$ are equal i.e. when the a_k coefficients are equal. In this case, *e(n)* = $\delta(n)$ and so the predictor polynomial $P(z)=1-\sum_{k=1}^{p}a_k z^{-k}$is a good approximation of the denominator of the initial transfer function. The coefficients mentioned are computed for the Auto-regressive model using the Levinson-Durbin algorithm.[21]

2.2) b) i) Pitch

Definition/Mathematical expression:

Pitch is the fundamental frequency of an audio waveform and is an important parameter in the analysis and synthesis of speech and music [20]. Pitch information can be extracted by using either temporal or frequency analysis. The temporal analysis method is based on the computation of the short-time autocorrelation function $R_n(l)$ or AMDF (Average Magnitude Difference Function) $A_n(l)$, where

$$R_n(l) = \sum_{i=0}^{N-l-1} s_n(i)s_n(i+l) \quad \text{and} \quad A_n(l) = \sum_{i=0}^{N-l-1} | s_n(i+l) - s_n(i)|.$$

where $s(i)$: discrete time audio signal

Based on frequency analysis methods, pitch can be determined from the periodic structure in the magnitude of the Fourier transform or cepstral coefficients of an audio frame.

Figure 5 (fig.5) shows the curves of the autocorrelation function and AMDF of a typical male voice segment.

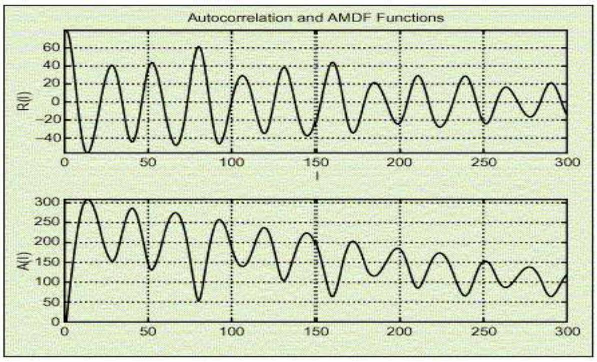


Fig.5 Autocorrelation function and AMDF of a typical male voice segment

Comments:

i) Generally, well-defined pitch characterizes only speech and harmonic music but it can still be used as a characteristic feature of the fundamental frequency of other audible waveforms.

ii) *Pitch determination based on the temporal analysis method*

From the autocorrelation and AMDF curves (fig.5) periodic peaks and valleys can be observed respectively. Peaks and valleys represent local maximum and minimum points. Using the local maximum/minimum points as well as the global maximum/minimum points of the curve, it is possible to specify the value of the pitch frequency. For example, from the AMDF curve of Fig.5 pitch frequency can be calculated by taking the reciprocal of the time period between the origin and the first valley. Generally, such valleys can be observed in voice and music audio clips but not in the noisy or the unvoiced ones.

iii) *Pitch determination based on the frequency analysis method*

When the frequency analysis method is chosen to determine the pitch, one way to do so is by calculating the maximum common divider for all the local peaks in the magnitude spectrum.

Figure 6 (fig.6) shows pitch curves for commercial, news and sports clips. The pitch contours are obtained using the autocorrelation function method.

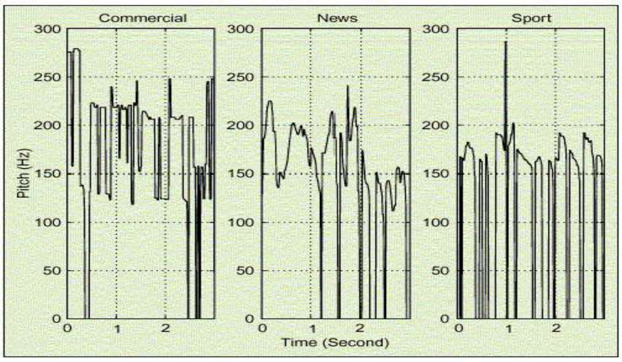


Fig.6 Pitch curves for commercial, news and sports clips

Comments:

i) The frames that are silent or where no pitch has been detected, are assigned a zero pitch frequency. Specifically, for the news clip first, by using the information obtained by the zero crossing rate and the volume curves, it is concluded that the segments with zero pitch correspond either to silence or unvoiced speech. Then for the sports clip, the zero pitch segments, which occur more often, correspond to periods with only background sounds but not silence as in the case of the news clip. Finally, discontinuous pitch segments in which the pitch has almost constant value, characterize the commercial clip. The music background of this kind of audio clips causes this characteristic.

ii) In speech, the pitch frequency basically depends on the speaker (male/female). In music signals it depends on the strongest note being played.

iii) There are three different features that are used in order to estimate the pitch variation: the standard deviation of the pitch, the smooth pitch ratio and the non-pitch ratio. The smooth pitch ratio feature estimates the percentage of music or voiced speech that exists within a clip, since only music and voiced speech are characterized by smooth pitch. On the contrary, the non-pitch ratio estimates the percentage of unvoiced speech or noise within an audio clip, since these two have no pitch.

2.2) b) ii) Spectrogram

Definition/Mathematical expression

Spectrogram is the 3-D plot that presents the magnitude spectrum (magnitude of the Fourier Transform) of a signal across time. Figure 7 (fig.7) shows the spectrograms of the commercial, news and sports clips.

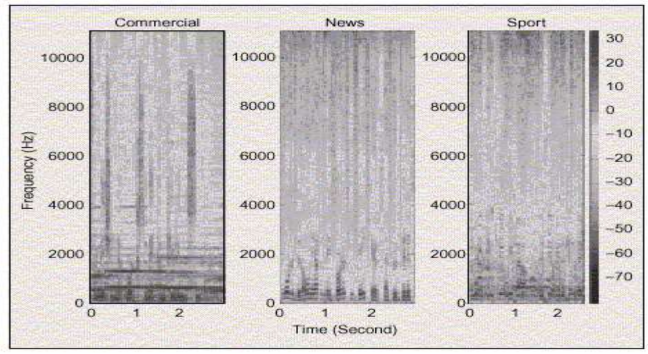


Fig.7 Spectrograms of the commercial, news and sports clips

Comments:

i) The distinct features between the three clips can be spotted more easily in the spectrogram compared to their time domain representations (Fig. 7 and Fig.2 respectively). Thus, the frequency domain representation of a signal may help more in the feature extraction process.

ii) The peak track in the spectrogram of an audio signal often reveals important characteristics of the sound [18]. For example, the spectral peak tracks of musical instruments, in their spectrograms, remain at the same frequency level and last for a certain period of time. Music may be classified into three subcategories, based again on the spectral peak track method: a) song, b) speech with music, and c) environmental sound with music background.

   a) Song audio signals are characterized by one of the following three features: ripple-shaped harmonic peak tracks due to voice sound, tracks with longer duration than speech or tracks with fundamental frequency higher than 300Hz.

   b) Speech with music background has its spectral peak tracks, concentrating in the lower to middle frequency bands and has lengths within a certain range.

   c) Finally, the environmental sound with music background does not have any certain characteristics.

iii) The major disadvantage of the use of the spectrum as a feature extraction tool is complexity. For audio retrieval applications, it is necessary to use a more compact way of signal representation in any domain to be used.

2.2) b) iii) iv) Frequency centroid, Bandwidth

Definition/Mathematical expression:

The frequency centroid (FC) and bandwidth (BW^2) of a signal are defined as:

$$FC(n) = \frac{\int_0^\infty \omega Sn(\omega)d\omega}{\int_0^\infty Sn(\omega)d\omega} \quad \text{and} \quad BW^2(n) = \frac{\int_0^\infty (\omega - FC(n))^2 Sn(\omega)d\omega}{\int_0^\infty Sn(\omega)d\omega}$$ where,

Sn(ω): power spectrum (magnitude square of the spectrum) of the audio signal

n: time index

Comments:

i) Let *ω* be a random variable and *Sn(ω)*, normalized by the total power, be the probability density function of *ω*. The mean and standard deviation value of *ω* correspond to the formulas of the frequency centroid (FC(n)) and bandwidth ($BW^2(n)$), respectively.

Note: It has been found that FC is related to the human sensation of the brightness of a sound [20].

2.2) b) v) Cepstral coefficients and Mel-frequency cepstral coefficients

Cepstral coefficients and Mel-frequency cepstral coefficients (MFCC) are mainly applied to speech and speaker recognition. Although, both techniques provide a smoothed representation of the original spectrum the MFCC technique takes into account the nonlinear property of the human hearing system with respect to different frequencies [20]. Generally, cepstral analysis attempts to deconvolve the excitation from the transfer function (mentioned in the [1.3) a)] paragraph), without making the assumptions that were necessary for linear prediction [8].

One way to generate the cepstral coefficients is from the linear prediction coefficients:

$$c_1 = \alpha_1$$

$$c_n = \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) \alpha_k c_{n-k} + \alpha_n$$ for $1 < n \leq p$

where c_i and a_i are the *i*th-order cepstral and linear predictor coefficients respectively.

Another way to generate cepstral coefficients is based on mel-scale.

Note: The mel-scale is defined in the following equation:

$$M=2595\log_{10}(1+\frac{f}{700})$$

These coefficients can be obtained by simulating critical-band filtering with a set of triangular band-pass filters, see Figure 8 (fig.8).

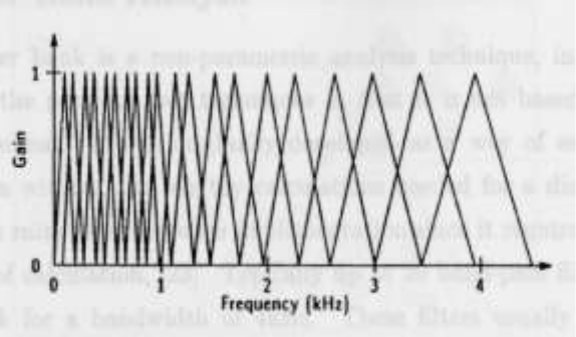


Fig.8 Triangular band-pass filter set

The filters are spaced linearly in the range 0 to 1000Hz. Also, each filter's center frequency is given by:

$f_{i+1} = 1.148 f_i$ where the initial frequency (f_1), is 1kHz.

So, the cepstral coefficients are obtained from the following formula:

$C_n = \sum_{k=1}^{L} X_k \cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{L}\right]$ for *n=1,2,…,M*

where X_k is the log-energy output of the kth filter, *L* is the number of filters in the desired bandwidth and *M* is the total number of coefficients required. [8].

Session 3

Transform-based signal representation and analysis

3.1) Overview of the method

In this paragraph a brief explanation of the proposed method will be given. The procedure can be divided into two main parts. The aim of the first part is to represent the signal in an efficient way so as to extract the information that will be useful for its correct classification. The second part has two aims. The first one is to keep the least possible necessary information extracted from the first part in order not to overload the classifier, and the second is to maximize the features that make one signal distinct from the others.

So, for the first part, two transforms were applied. The first one is the Fourier transform that provides a visual representation of the signal magnitude as well as the signal phase. The second one is the Hartley transform which gives a real frequency domain function for a real time signal. The disadvantage of the Fourier transform when compared to the Hartley is that information may be lost if the magnitude spectrum is retained but the phase is neglected. Using the Fourier transform, in order to preserve all the information, it is necessary to use both the magnitude and phase spectra. On the other hand, when using the Harley transform, only one time-frequency surface need be used, because both magnitude and phase information are present in the one surface. Later in this report, an alternative method is described where the signal is analyzed from its energy distribution point of view using the Wigner-Ville distribution. Although some phase information is lost, enhanced resolution is obtained which may provide improved performance. For the second part, in order to fulfill the compression aim the 2-D cosine transform is used. This transform is applied over both the Fourier and Hartley. Then, using the Hough transform, which is applied over the Wigner-Ville distribution, the maximization of the distinct features of each signal is achieved.
All the transforms and approaches that described before have been applied to signals and the results are presented.

3.2) The Fourier transform & the Fourier cosine transform

The Fourier transform and the Fourier cosine transform are not analysed and discussed because the reader is already familiar with both of them. Instead, in the next two pages are included the periodogram, phasegram and 2-D Cosine transform of two different audio signals. The first sound is a F1 racing car and the second is the sound of countryside atmosphere.
In the next pages are presented the graphs:
Fig.9 Spectrogram of an audio sample of Formula1 car
Fig.10 DCT-2 applied on the spectrogram of the Formula1 car
Fig.11 DCT-2 (focus) applied on the spectrogram of the Formula1 car
Fig.12 Phasegram of an audio sample of a Formula1 car
Fig.13 DCT-2 applied on the phasegram of the Formula1 car
Fig.14 DCT-2 (focus) applied on the phasegram of the Formula1 car
Fig.15 Spectrogram of an audio sample of country atmosphere

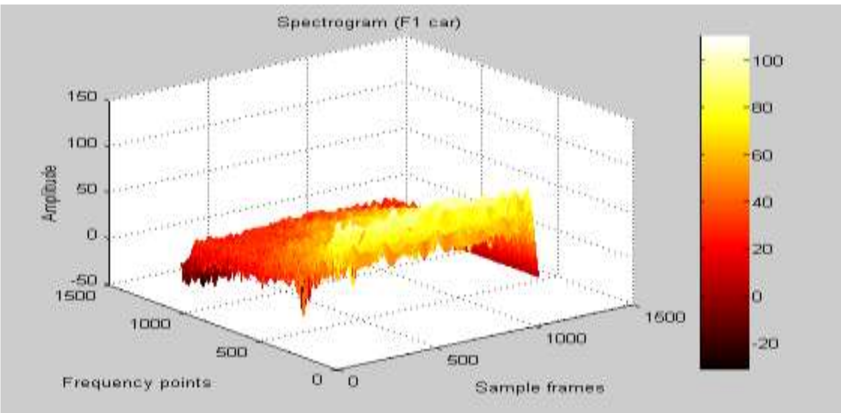


Fig.9 Spectrogram of a Formula1 car

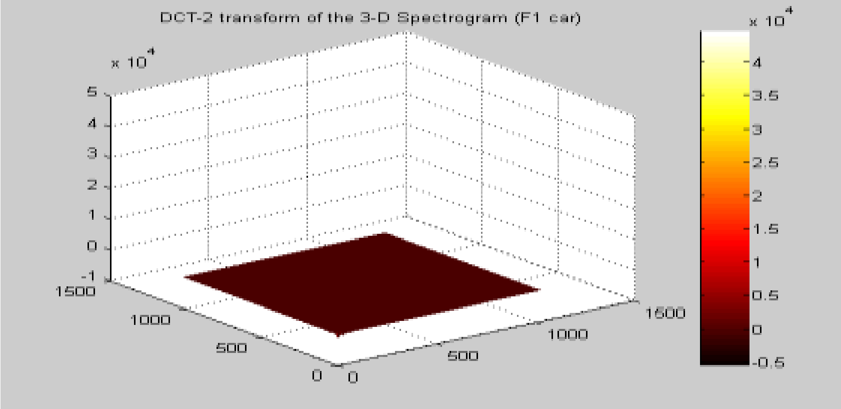


Fig.10 DCT-2 applied on the spectrogram of the Formula1 car

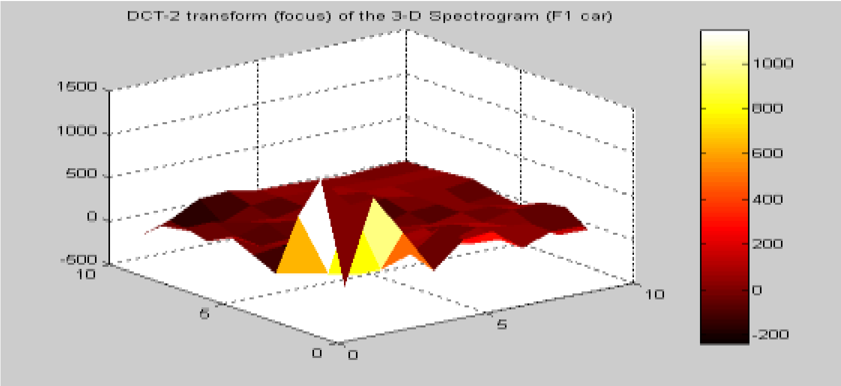


Fig.11 DCT-2 (focus) applied on the spectrogram of the Formula1 car

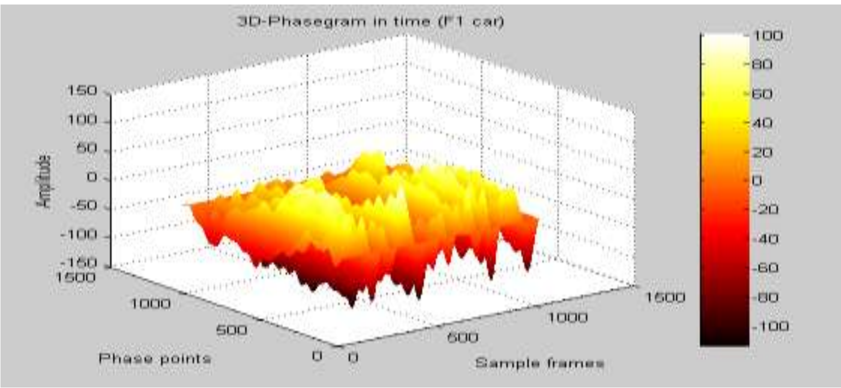


Fig.12 Phasegram of an audio sample of a Formula1 car

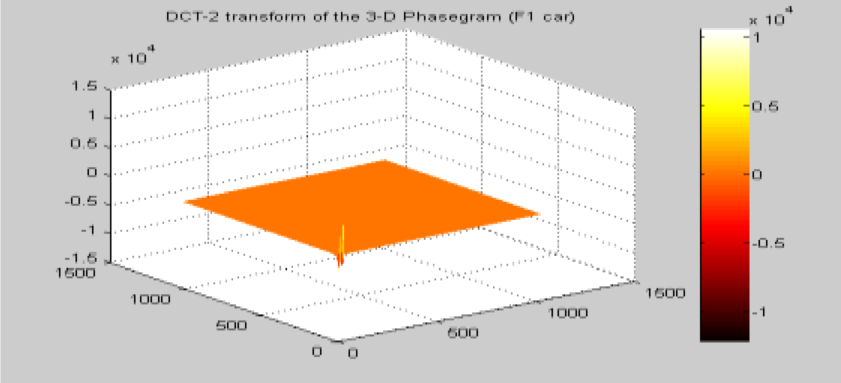


Fig.13 DCT-2 applied on the phasegram of the Formula1 car

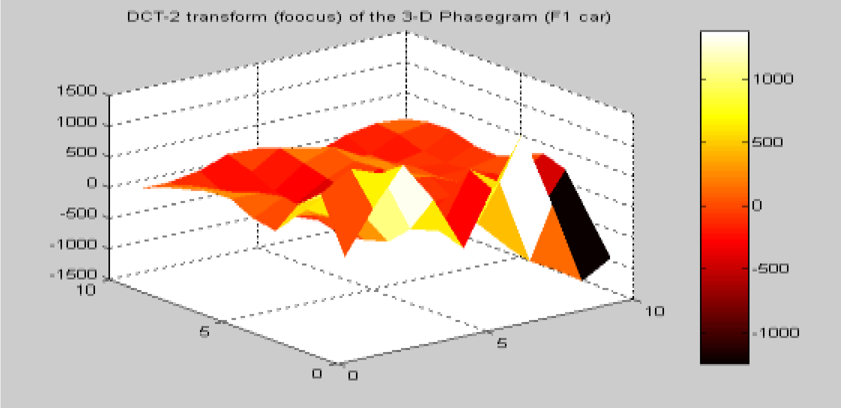


Fig.14 DCT-2 (focus) applied on the phasegram of the Formula1 car

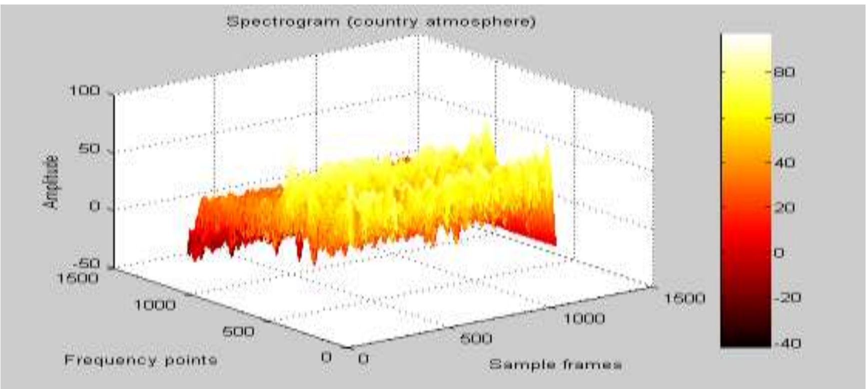


Fig.15 Spectrogram of an audio sample of country atmosphere

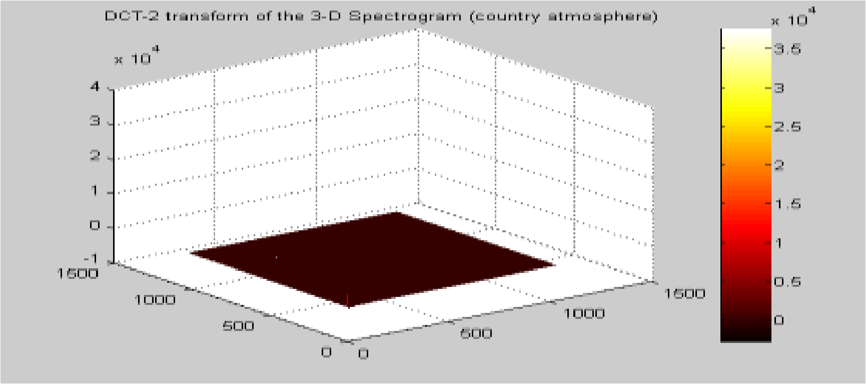


Fig.16 DCT-2 applied on the spectrogram of the country atmosphere

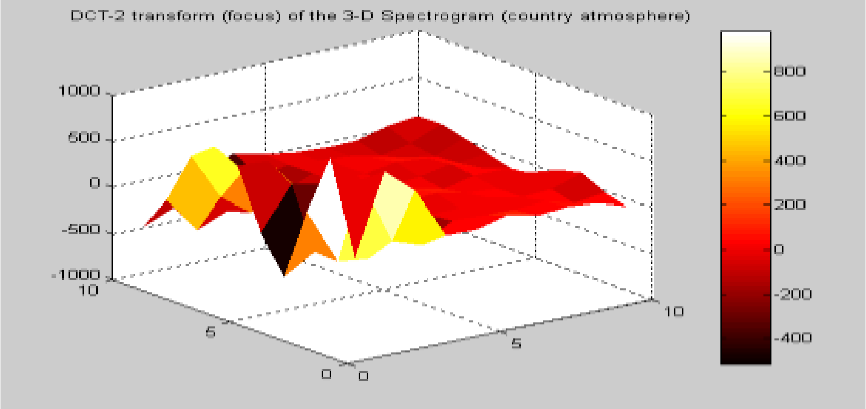


Fig.17 DCT-2 (focus) applied on the spectrogram of the country atmosphere

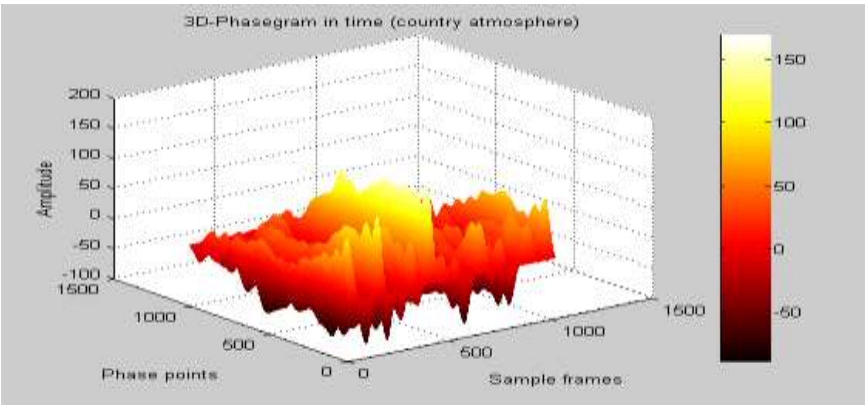


Fig.18 Phasegram of an audio sample of country atmosphere

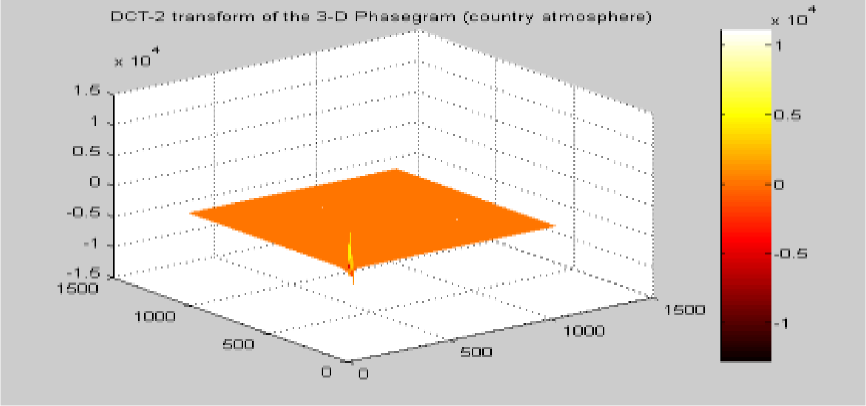


Fig.19 DCT-2 applied on the phasegram of country atmosphere

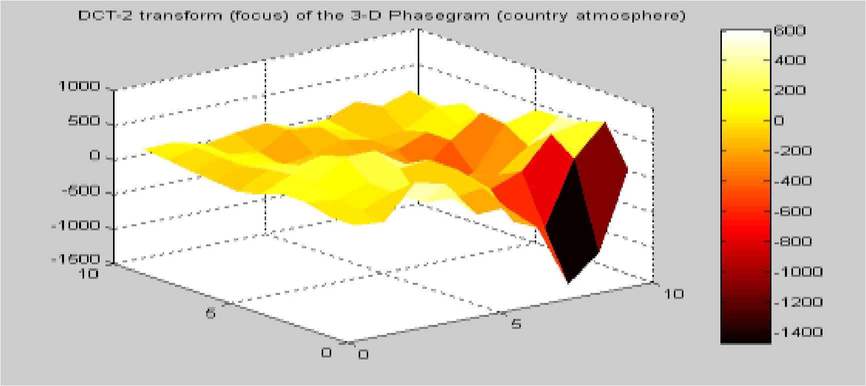


Fig.20 DCT-2 (focus) applied on the phasegram of country atmosphere

Comments-Discussion

The first graph (spectrogram) represents the time-frequency energy distribution of the signal. Phase information has been lost due to the use of the magnitude squared terms. So, in order to have a visual representation of the phase evolution across time, the phase-gram surface was also implemented. Both surfaces, generated via the Fourier transform, contain complex structures spread across the time-frequency plain. In order to present this information to a classifier, the information needs to be compressed. The 2-D Cosine transform (commonly used in image compression) was applied so as to compress all this information. Figures 13 and 19 show the result of applying the DCT to the phase-grams for two acoustic events. Figures 14 and 20 focus on the low order areas of interest of the 2-D Cosine transforms for each result, in order to visualise the different surface structure between the two audio signals. It is proposed that this might form the basis of a feature extraction process that would be applied to a suitable classifier.

3.3) The Hartley transform

Definition

The Hartley transform is one of a set of orthogonal transforms, which gives a real frequency domain function from a real time signal [6].

The mathematical expression of the Hartley transform and its inverse are:

H(ω) = $\int f(t)\,(\cos(\omega t)+\sin(\omega t))\,dt$ or

H(v) = $\frac{1}{N}\sum_{\tau=0}^{N-1} f(\tau)\,cas\left(\frac{2\pi v\tau}{N}\right)$ (discrete version)

where $\tau = t$, $\omega = 2\pi v$ cas $(\theta) = \cos(\theta) + \sin(\theta)$ and s(t) = $\int H(\omega)(\cos(\omega t)+\sin(\omega t))\,d\omega$ or

s($t)=\sum_{\nu=0}^{N-1} H(\nu)\,\mathrm{cas}\left(\frac{2\pi\nu t}{N}\right)$ (discrete version)

respectively.

Relationship between Hartley and Fourier transforms

The definition of the Fourier transform is:

S($\omega)=\int s(t)e^{-j\omega t}dt$, S($\omega)=\int s(t)(\cos(\omega t)-j\sin(\omega t))dt$ and S($\omega)=\int s(t)\cos(\omega t)dt-j\int s(t)\sin(\omega t)dt$

Let,

$S_R(\omega) = \int s(t)\cos(\omega t)dt$ and

$S_I(\omega) = \int s(t)\sin(\omega t)dt$

So, $S(\omega) = S_R(\omega) - j\,S_I(\omega)$

Now, the same procedure will be applied to the Hartley transform:

H(ω) = $\int s(t)(\cos(\omega t)+\sin(\omega t))dt$

H($\omega)=\int s(t)\cos(\omega t)dt+\int s(t)\sin(\omega t)dt$

Let, $S_R(\omega) = \int s(t)\cos(\omega t)dt$ and $S_I(\omega) = \int s(t)\sin(\omega t)dt$ So,

$H(\omega) = S_R(\omega) + S_I(\omega)$

Finally,

$S(\omega) = S_R(\omega) - j\,S_I(\omega)$ for the Fourier transform and

$H(\omega) = S_R(\omega) + S_I(\omega)$ for the Hartley transform.

The last two mathematical expressions that relate the two transforms, are considered as a very important tool for the derivation of the Hartley transform properties through the equivalent ones of the Fourier transform set.

A summary of the basic properties of the Hartley transform is the following:

Theorem	**f(x)**	**H(f)**
Linearity	$f_1(x) + f_2(x)$	$H_1(f) + H_2(f)$
Scaling/Similarity	f(kx)	$\lvert\frac{1}{k}\rvert\,\mathrm{H}(\frac{f}{k})$
Reversal	f(-x)	H(-f)

Shift	f(x-T)	H(f)=cos(2πfT)H(f)+sin(2πfT)H(-f)
Modulation	f(x) cos(2πf$_0$T)	H(f) = $\frac{1}{2}$ H(f - f$_0$) + $\frac{1}{2}$ H(f + f$_0$)
Convolution	f$_1$(x) * f$_2$(x)	$\frac{1}{2}$ [H$_1$(f)H$_2$(f)+H$_1$(-f) H$_2$(f) +H$_1$(f)H$_2$(-f)-H$_1$(-f) H$_2$(-f)]
Autocorrelation	f$_1$(x).f$_1$(x)	$\frac{1}{2}$ [H$_1$(f)^2+H$_1$(-f)^2]
Product	f$_1$(x)f$_2$(x)	$\frac{1}{2}$ [H$_1$(f)*H$_2$(f)+H$_1$(-f)*H$_2$(f) +H$_1$(f)*H$_2$(-f)-H$_1$(-f)*H$_2$(-f)]

The Hartley magnitude and phase

The definitions of the Fourier magnitude and phase are:

M(ω) = $\sqrt{S_R^2(\omega)+S_I^2(\omega)}$ and $\varphi(\omega)=\arctan\frac{S_I(\omega)}{S_R(\omega)}$ respectively.

Also, $S(\omega) = S_R(\omega) + S_I(\omega)$ (Hartley transform)

and according to complex number theory,

$S(\omega) =$ M(ω)(cos($\varphi(\omega))+j\sin(\varphi(\omega))$

= M(ω)cos $(\varphi(\omega))+jM(\omega)\sin(\varphi(\omega))$

So, $S_R(\omega)$ = M(ω)cos($\varphi(\omega)$) and

$S_I(\omega)$ = M(ω)sin($\varphi(\omega)$)

thus $H(\omega) = S_R(\omega) + S_I(\omega)$

= M(ω)cos($\varphi(\omega)$) + M(ω)sin($\varphi(\omega)$)

= M(ω)(cos($\varphi(\omega)$) + sin($\varphi(\omega)$))

Concluding, the Hartley transform magnitude is the same as the Fourier transform one, but the Hartley phase is defined as:

Y(ω) = $\frac{H(\omega)}{M(\omega)}$ = $\frac{M(\omega)(\cos(\varphi(\omega))+\sin(\varphi(\omega)))}{M(\omega)}$ = cos($\varphi(\omega)$) + sin($\varphi(\omega)$).

The last equation is a function of the Fourier phase only and has been called the “whitened Hartley spectrum” or the “Hartley phase spectrum” [6].

The Hartley phase spectrum properties are listed as follows:

i) it has zero mean

ii) it has a standard deviation of unity

iii) it is a continuous function of frequency

iv) it has upper and lower bounds of $\pm\sqrt{2}$

All four properties are independent of the signal statistics.

In the next pages are presented the graphs:

Fig.21 Hartley spectrum of an audio sample of Formula1 car (H.T)

Fig.22 Phasegram of an audio sample of a Formula1 car (H.T)

Fig.23 DCT-2 applied on the phasegram of the Formula1 car (H.T)

Fig.24 DCT-2 (focus) applied on the phasegram of the Formula1 car (H.T)

Fig.25 Hartley spectrum of an audio sample of country atmosphere (H.T)

Fig.26 Phasegram of an audio sample of country atmosphere (H.T)

Fig.27 DCT-2 applied on the phasegram of country atmosphere (H.T)

Fig.28 DCT-2 (focus) applied on the phasegram of country atmosphere (H.T)

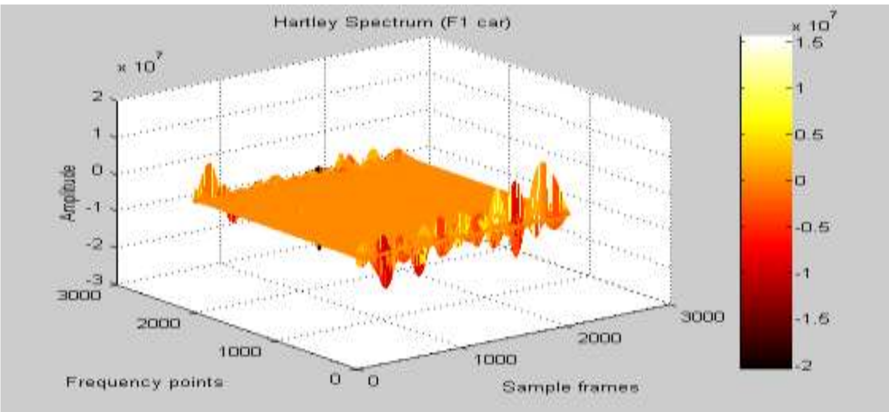


Fig.21 Hartley spectrum of an audio sample of Formula1 car (Hartley Transform)

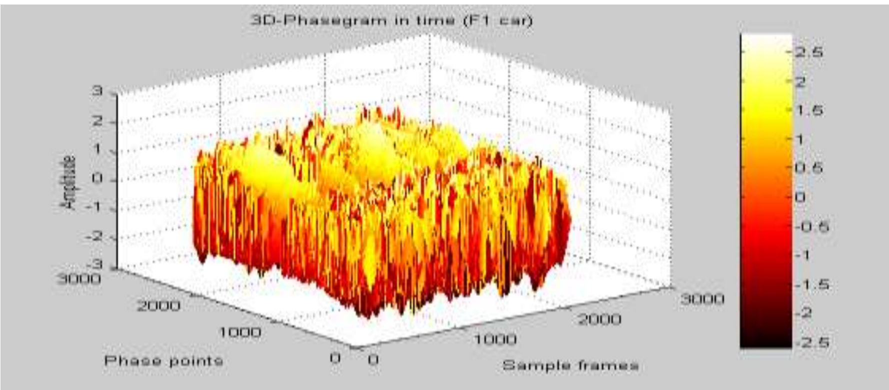


Fig.22 Phasegram of an audio sample of a Formula1 car (Hartley Transform)

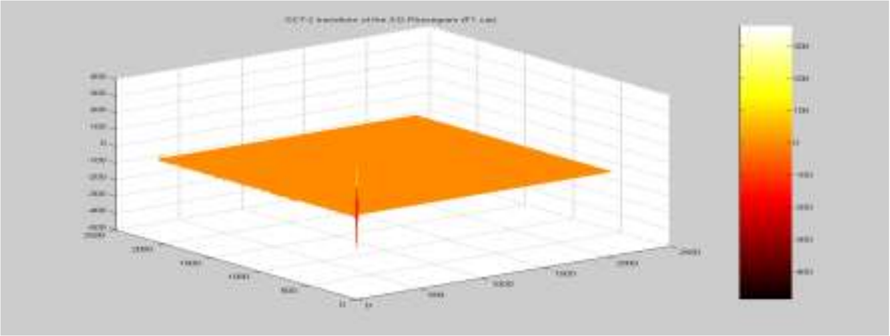


Fig.23 DCT-2 applied on the phasegram of the Formula1 car (Hartley Transform)

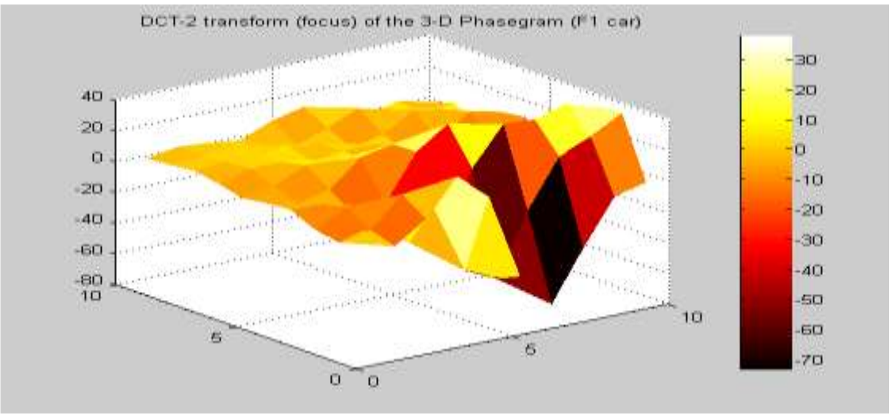


Fig.24 DCT-2 (focus) applied on the phasegram of the Formula1 car (H.T)

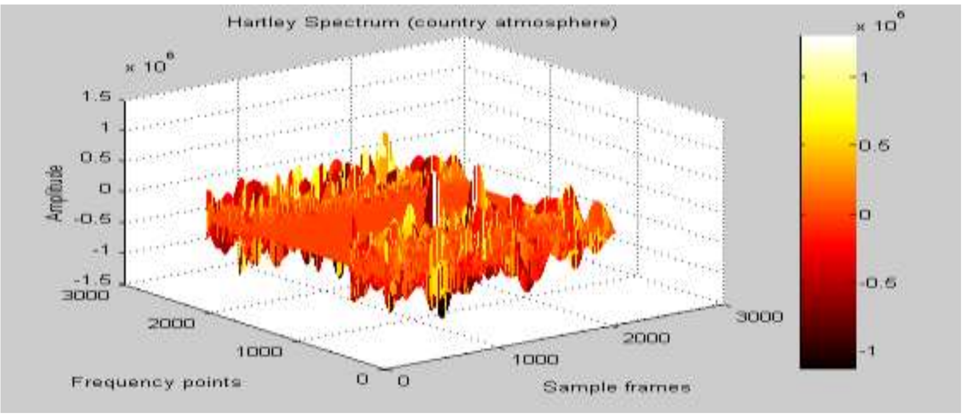


Fig.25 Hartley spectrum of an audio sample of country atmosphere (H.T)

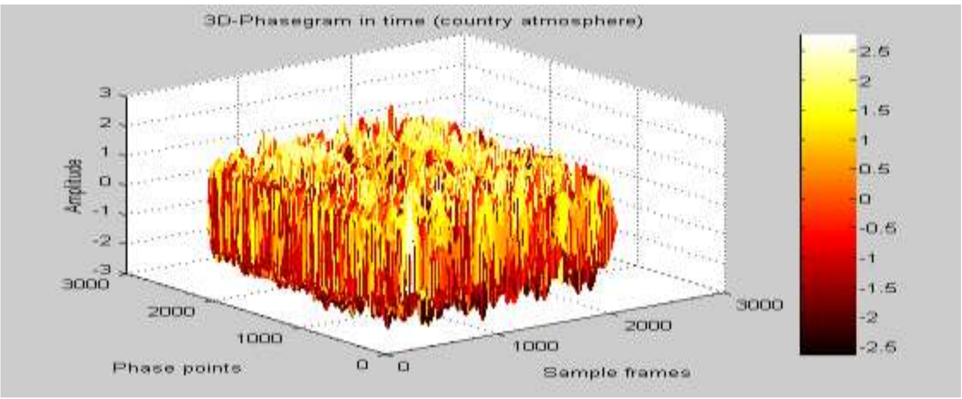


Fig.26 Phasegram of an audio sample of country atmosphere (H.T)

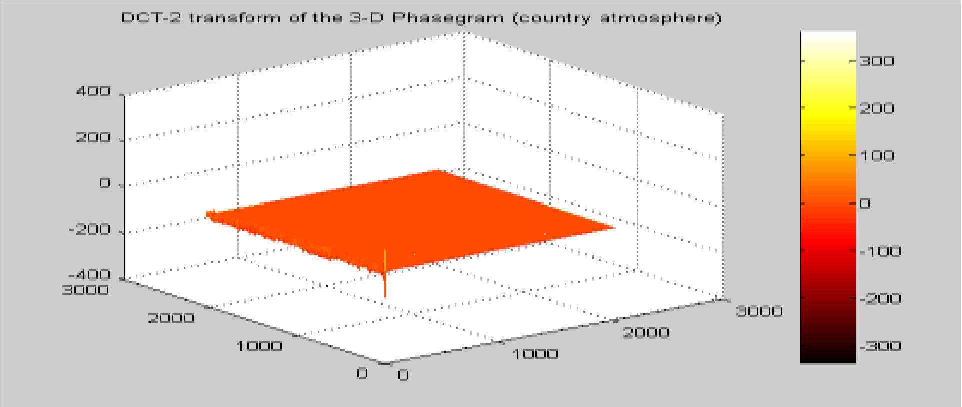


Fig.27 DCT-2 applied on the phasegram of country atmosphere (H.T)

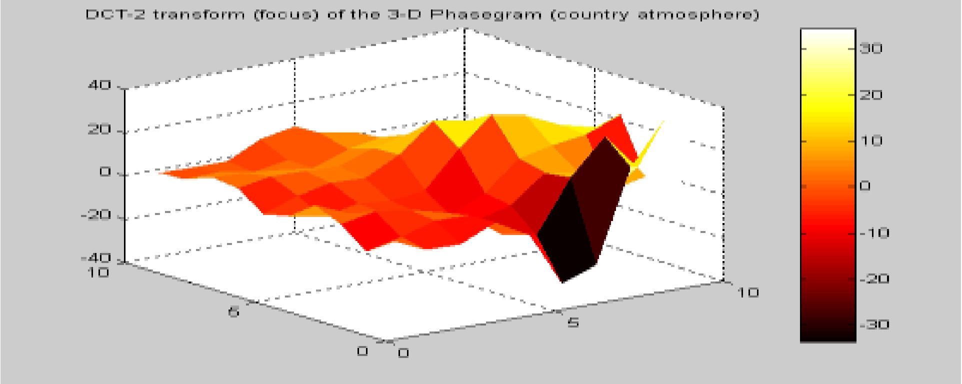


Fig.28 DCT-2 (focus) applied on the phasegram of country atmosphere (H.T)

Comments-Discussion

In order to overcome the information (phase) loss observed in the Fourier transform case, the Hartley transform was applied to the audio signal. Here, only one time-frequency surface is required to represent both magnitude and phase information. However, the phase information may be separated out and examined independently if needed. Both the magnitude and phase evolution of the signal across time were implemented but the 2-D Cosine transform was only applied to the phase surface, as an example of the compression properties of the DCT. Again, the last graph, focuses on the area of interest of the 2-D Cosine transform in order to visualise the different surface characteristics between the two audio signals. The use of these techniques with the Hartley transform are in the very early stages and these examples are presented here only as an indication of the possible lines of investigation that may be followed in the future.

3.4) The Wigner-Ville distribution

One very important property of the linear time-frequency representation using the Fourier Transform is the property of linearity. But, because of the complex nature of the spectrum, the time-frequency representation of a signal cannot be easily visualized. For that reason it is usual, to display the representation as time-frequency graphs of magnitude (squared) and phase separately. [6]

So, Linear Time-Frequency Representation (LTFR) is the complex time-frequency representation using the Fourier Transform and Quadratic Time-Frequency Representation (QTFR) is the spectrogram due the presence of the squared terms.

The QTFRs have important characteristics that are derived from the fact that the time-frequency function is real and relates to the distribution of signal energy (power). Let, T(f, τ) be the QTFR of a signal s(t). Two major properties that the QTFRs should have are: $\int_f T(f,\tau)df =$ |s(t)|^2 (instantaneous signal power) and $\int_\tau T(f,\tau)d\tau =$ |S(f)|^2 (power spectrum of the signal).

These properties are called marginal properties of the QTFR. Because the spectrogram does not fulfill the marginal properties the Wigner-Ville Distribution (WD) is applied instead.

The Wigner-Ville distribution is defined as:

$$W_x(t,f) = \int x\left(t+\frac{\tau}{2}\right)x^*\left(t-\frac{\tau}{2}\right)e^{-j2\pi f\tau}d\tau$$ and

$$W_X(t,f) = \int X\left(f+\frac{v}{2}\right)X^*\left(f-\frac{v}{2}\right)e^{j2\pi vt}dv$$

Because the WD is a quadratic function, its sampling has to be done in a careful way So, let us write it as follows:

$$W_x(t,f) = \int x\left(t+\frac{\tau}{2}\right)x^*\left(t-\frac{\tau}{2}\right)e^{-j2\pi f\tau}d\tau \qquad \int_{-\infty}^{\infty} x\left(t+\frac{\tau}{2}\right)x^*\left(t-\frac{\tau}{2}\right)e^{-j2\pi f\tau}d\tau$$

$$2\int_{-\infty}^{\infty} x(t+\tau)x^*(t-\tau)e^{-j4\pi f\tau}d\tau$$

If x is sampled with a period T_e, then $x[n] = x(nT_e)$ and the WD is evaluated at the sampling points nT_e in time. So, a discrete-time expression of the WD is obtained:

$$W_x(n,f) = 2T_e\sum_k x[n+k]x^*[n-k]e^{-j4\pi fk}$$

Except of the marginal properties some other important properties that characterize the WD are:

i) real valued

ii) time shift: if s'(t) = s(t-t') then W_D(t-t',f)

iii) frequency shift: if s'(t) = s(t)$e^{-j2\pi f't}$ then W_D(t,f-f')

iv) Convolution: s(t) = x(t)*y(t) (*: denotes convolution)

$W_{D_s}(t,f) = W_{D_x}(t,f) * W_{D_y}(t,f)$ (*: denotes convolution in time)

v) Multiplication: s(t) = x(t)y(t)

$W_{D_s}(t,f) = W_{D_x}(t,f) * W_{D_y}(t,f)$ (*: denotes convolution in frequency)

3.5) The Hough transform (applied to line detection)

Definition

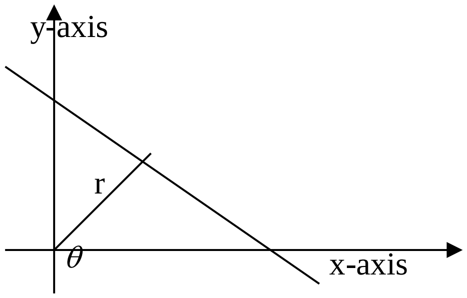
The Hough transform is a method that, in theory, can be used to find features of any shape in an image. In practice, it is generally used for finding straight lines or circles. The computational complexity of the method grows rapidly with more complex shapes.

Consider a point (x_i, y_i) on an image. The number of lines that can pass through this point are infinite. However, they all can be represented by the following equation:

$x\cos(\theta) + y\sin(\theta) = r$ where

r : represents the distance (perpendicular) of the line from the origin and

θ : represents the angle between this perpendicular and the x-axis



Note: $0 \leq \theta \leq 2\pi$

Assuming, in the last equation varying r and θ, and fixed x and y (i.e. X_i and Y_i respectively) then for each of the possible lines that pass through point (X_i, Y_i) the equation has coordinates r and θ in (r,θ) space. In other words, each of the lines that pass through point (X_i, Y_i) have a unique value of r and θ. Based on the previous assumptions, two spaces are defined i.e. the first one is the image (xy) space whereas the second one is the parameter (r, θ) space. Now, every line that passes through a point in image space is mapped to a line in parameter space.

So, the Hough transform is a mapping from image space to parameter space.

N points that belong to the same line, i.e. (x_i, y_i) for i = 1,2…N, are transformed into N sinusoidal curves $x\cos(\theta)+y\sin(\theta)$ in the parameter (r, θ) space, which intersect in the point (R, Θ).

<u>Comments/Conclusions</u>

i) A point in the image space corresponds to a sinusoidal curve in the parameter space.

ii) A point in the parameter space corresponds to a straight line in the image space.

iii) Points that belong to the same straight line in the image space correspond to curves through a common point in the parameter space.

iv) Points that belong to the same curve in the parameter space correspond to lines through a common point in the image space.

In the next pages are presented the graphs:

Fig.29 Chirp Signal

Fig.30 Wigner-Ville distribution of the chirp signal

Fig.31 Hough Transform (Detection of straight lines) of WD

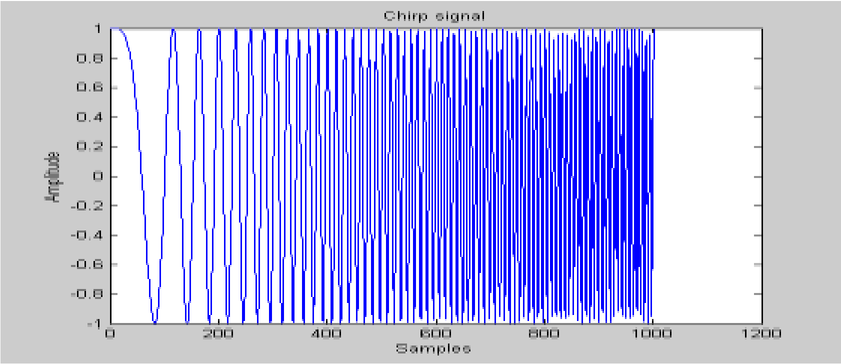


Fig.29 Chirp Signal

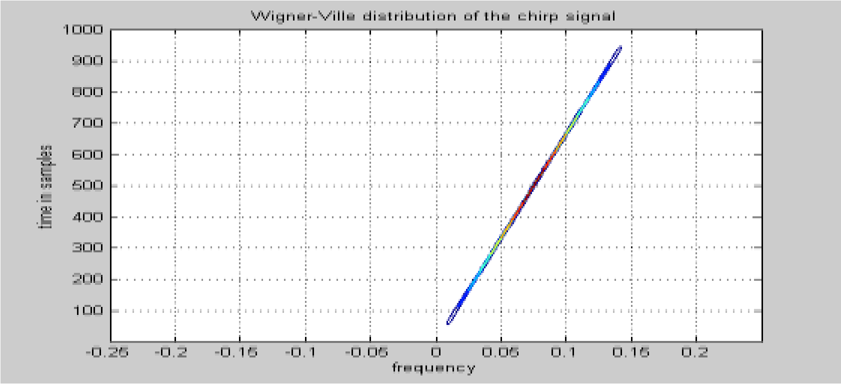


Fig.30 Wigner-Ville distribution of the chirp signal

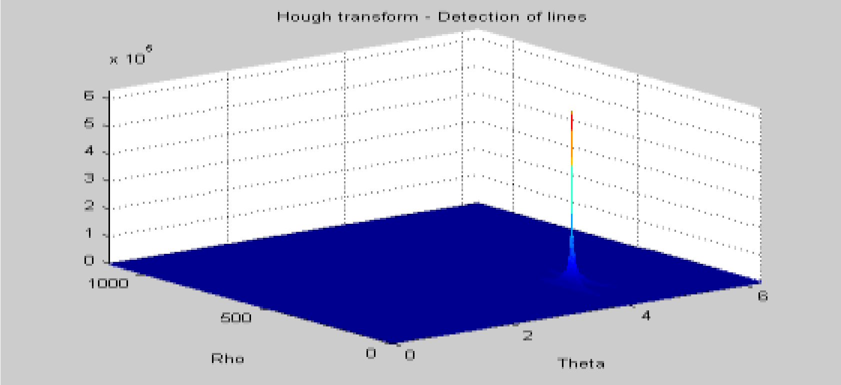


Fig.31 Hough Transform (Detection of straight lines) of WD

Comments-Discussion

The two-level transform method that is described and implemented in this paragraph, aims to visualise the difference amongst the energy distributions of various signals. The Wigner-Ville distribution provides the energy distribution of a particular signal, whereas the Hough transform detects the straight lines of the energy distribution surfaces and provides a way to visualise the difference surface characteristics between different kinds of signals. Although, the Wigner-Ville representation and is quite similar to the spectrogram technique analysed in the literature review this method has the advantage of preserving the marginal properties of the signal.

General conclusions and discussion

The first group of feature extraction methods i.e. Spectrogram and Phasegram were used as an introduction in order to understand the disadvantages of the Fourier-based methods. Applying the spectrogram to a signal there is information (phase) loss. So, the Hartley Transform is used so no phase loss is experienced, whereas the Wigner-Ville distribution is a time-frequency graph that preserves the marginal properties of the signal. The second level of transforms i.e. 2-D Cosine and Hough are used for compression and maximization of the distinct features respectively.

Session 4

4.1) Future plans

In the next time period, the project will be focused on the extraction of common characteristics between audio signals based on the feature extraction techniques described in the previous Session. The first (Fourier, Hartley and Wigner-Ville) processing level and the second (Hough and 2-D DCT) one will be combined in different ways in order to detect if there are other more effective ways of feature extraction. Also, it is planned to implement the wavelet transform so as to apply its compression properties to the signal.

The Wigner-Ville energy distribution belongs to the Cohen's class of energy distributions. There are other members of the same category of distributions that represent the signal in a way that helps more to the extraction of distinct features between the various clips. So, another future plan is to apply the appropriate energy distribution to the signal for better feature extraction.

After the feature extraction part has finished, the next step is the signal classification. At the beginning a simple Euclidean classifier will be implemented so as to decide which combination of the feature extraction methods provides the best results. The next step will involve the implementation of a more sophisticated, self-organized classifier that will provide more accurate estimations. In the classification part it may be used a network of two simple classifiers that will be the input to a super- classifier in order to reduce the dimensionality of the input vector to each one of them.

The two-level feature extraction technique has only been applied to artificially constructed signals. Artificially constructed signals can easier be classified because there is apriory information. On the other hand, for sounds of nature and for other kinds of sounds apriory information does not exist, so the feature extraction part has to be more intensive.

References

Books, Ph.D. theses & Lecture Notes

[1] R. Bracewell, “The Fourier Transform and Its Applications”, Second Edition, McGraw-Hill Book Company,1986

[2] Matlab, “Signal Processing Toolbox, User’s Guide”, The MathWorks, December 1996

[3] [2] Matlab, “Signal Processing Toolbox, User’s Guide”, The MathWorks, December 1996

[4] J. Kittler, Pattern Recognition Lecture Notes, University of Surrey, 2001

[5] EHS Chilton, Digital Speech Processing Lecture Notes, University of Surrey, 2001

[6] EHS Chilton, A Continuing Education Course on Advanced Digital Signal Processing Lecture Notes, University of Surrey, 2001

[7] Peter R. Green, Digital Signal Processing Lecture Notes, UMIST, 2000

[8] H.Kelleher, Continuous, Speaker Independent, Speech Recognition for a Speech to Viseme Translator, PhD thesis, University of Surrey, 1999

[9] Alexander D. Poularikas,”The Transforms and Applications Handbook”,CRC Press&IEEE Press, 1996

[10] Alan V. Oppenheim, Ronald W. Schafer, John R. Buck,”Discrete-Time Signal Processing”,Second Edition, Prentice Hall Sinal Processing Series, 1999

[11] Petre Stoica, Randolph Moses, “Introduction to Spectral Analysis”, Prentice Hall, 1997

Publications

Literature Review

[12] Stephen Rice, Find audio and video! See the audio! Comparisonics Corporation, Grass Valley, CA 95945 USA, www.comparisonics.com

[13] Jonathan Foote, Content-based retrieval of music and audio, Multimedia Storage and Archiving Systems (II), Procee. Of SPIE, vol.3229, pages 138-147, 1997

[14] Zhu Liu, J. Huang, Y. Wang, and T.Chen, Audio feature extraction and analysis for scene classification, Workshop on Multimedia Signal Processing (Electronic Proceedings), June 23-25 1997

[15] M.J. Ross, H.L. Schafer, Andrew Cohen, R.Freudberg, and H. Manley. Average magnitude difference function pitch extractor. IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-22(No.5):353-361, 1974

[16] R.Schafer and L.Rabiner, Digital representation of speech signals, Proceedings of IEEE, vol.63, No.4, pages 662-677, 1975

[17] T. Zhang and C.C. Jay Kuo, Hierarchical classification of audio data for archiving and retrieving, International Conference on Acoustic, Speech and Signal Processing 1999, volume 6, pages 3001-3004, 1999

[18] Tong Zhang and C.-C. Jay Kuo, Audio-guided audiovisual data segmentation, indexing,, and retrieval, SPIE Conference on Storage and Retrieval for Image and Video Databases (VII), pages 316-327, 1999

[19] Stan Z. Li, Content-based classification and retrieval of audio using the nearest feature line method, http://citeseer.nj.nec.com/202406.html

[20] Yao Wang, Zhu Liu, and Jin-Cheng Huang, Multimedia Content Analysis using both Audio and Visual Clues, IEEE Signal Processing Magazine, Pages 12-36, Noveber 2000

[21] M. Pandit, J. Kittler, W. J. Christmas, Audio Classification, ASSAVID project, May 2001

Wigner-Ville & Wigner-Hough Transform

[22] F. Auger and P. Flandrin, Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method, IEEE Transactions on Signal Processing, 43(5):1068-89, 1995

[23] S. Barbarossa Analysis of Multicomponent LFM Signals by a Combined Wigner-Hough Transform, IEEE Transactions on Signal Processing, 43(6), June 1995

[24] J. Bertrand and P. Bertrand, A class affline wigner functions with extended covariance properties, J.Math.Phys.,33(7), 1992

[25] L. Cohen, Time-Frequency Distributions-A Review, Proceedings of the IEEE, 77(7):941-980, 1989

[26] F. Hlawatsch and F. Boudreaux-Bartels, Linear and Quadratic Time-Frequency Signal Representations, IEEE SP Magazine, pages 21-67, 1992

Hough

[27] P.V.C. Hough, “Method and means for recognizing complex patterns”, U.S. Patent 3 069 654, Dec. 18, 1962

[28] R.O. Duda and P.E. Hart, “Use of Hough Transformation to detect lines and curves in pictures”, Commun. Ass. Comput. Mach.,vol.15, Jan. 1972

[29] Philip M. Merlin and David J. Farber, “A Parallel Mechanism for Detecting Curves in Pictures”, IEEE Transactions on Computers, January 1975