

# A Sound Source Classification System Based On Subband Processing

Oytun Türk, Helin Dutagacı, Ömer Şaylı and Levent M.Arslan

Boğaziçi University, Electrical and Electronics Eng. Dept., Bebek, 80815, Istanbul, Turkey

## Abstract

A classification system that aims to recognize the presence of sounds from different sources is described. The type of audio signals considered are speech, music, noise and silence. Appropriate subband processing is applied for the characterization of each sound source. The algorithm operates in four steps to classify the contents of a given audio signal. The acoustical parameters and statistical measures to be used in the classification process are obtained via an off-line training procedure. In the silence and onset detection stages, we aim to label the starting and finishing instants of the acoustical events present in the audio signal. Acoustical parameters of the given signal are extracted and classification is carried out using linear discrimination with common covariance matrix. Experimental work is carried out on a database that contains mixtures of human speech, musical instruments, background noise of different types and silence. Experimental results demonstrate that the system yields %XX.X classification success for speech/music mixtures, %XX.X for speech/noise mixtures, %XX.X for different musical instruments, %XX.X for mixtures containing speech, music and noise.

## 1.INTRODUCTION

### 1.1.Motivation and Applications

Human auditory system is subject to different types of sound mixtures in daily life. Although the characteristics of these mixtures vary greatly due to the sources producing the mixture and the acoustical environment, human ear is successful in identifying the types of sounds in a complex sound mixture in many cases. The main goal of Computational Auditory Scene Analysis (CASA) is to implement systems that can imitate this behavior.

The applications include:

- (i) *Transcription of audio containing sounds from different sources such as speech, music, and noise*. The sounds may both exist simultaneously (i.e. different sound sources contribute to the sound mixture at the same time) and non-simultaneously (different sound sources are active at the same time).
- (ii) *Speaker identification using acoustical features.*
- (iii) *Identification of musical instruments, monophonic or polyphonic musical transcription and musical rhythm (tempo) tracking.*
- (iv) *Acoustical model based front-end processing for speech recognition* [Spina and Zue, 1996b].

(v) *Content based multimedia parsing* [Naphade and Huang, ?].

In this study we aim to :

- (i) implement modules to extract different acoustical parameters from audio signals,
- (ii) analyze different audio signal databases using acoustical parameter extraction modules,
- (iii) use the results of the analysis in determining the best feature sets in transcribing sounds from different sources,
- (iv) implement and test the performance of a sound source classification system using acoustical features.

### 1.2. Review of ASA Literature

Auditory Scene Analysis(ASA) aims to explain the way that the human auditory system processes complex sound mixtures. Thus, the methods related to ASA usually employ information from the diverse fields of psycho-acoustics and signal processing. The work in this field was initiated with the work of Bregman in 1970s which is presented in [Bregman, 1990]. Many computational models emerged through 1980 to 1990 which aim to determine the pitch of complex sounds, transcribe complex

musical scenes, and classify audio signals [Scheirer, 1998].

General audio classification using acoustical features is addressed in many studies. In [Spina and Zue, 1996a], the authors focused on a system to transcribe broadcast news which contained clean speech, noise corrupted speech, background noise, music, and silence. Several other studies were carried out for broadcast news transcription including [Cook, et.al 1997] and [Gauvain, et.al 1997]. Video indexing using audio data is another practical application that has attracted the attention of researchers [Saraceno and Leonardi, 1997].

As musical signals exhibit a considerable amount of complexity in terms of acoustical events, several studies were aimed at automated music transcription [Martin 1996], instrument identification [Brown, 1998] and rhythm tracking[Goto and Muraoka, 1996].

### 1.3. New Methods Proposed

In this study, information from different frequency bands of the signal spectrum is used for accurate and robust onset detection. Psycho-acoustical evidence on time domain and frequency domain masking is employed.

The harmonic analysis stage also relies on appropriate subband processing. It is applied to decide on the harmonic components that may be present in the audio signal. Autocorrelation based pitch detection is carried out using a subband based scheme.

The filterbank used for subband analysis is a bank of bandpass filters that model human cochlea. Appropriate choice of center frequencies and bandwidths is made according to [Zwicker and Fastl, 1999] and [ISO/IEC, 1993].

### 1.4. Paper Outline

Next section (Section 2) starts with the presentation of the block diagram of the implemented sound source recognition system as well as a sample output produced by the system. Basic module outputs are also presented. Section 3 describes the methods used for feature extraction and implementation details of the main modules of the system. The modules carry out silence detection, onset detection, acoustical feature extraction and classification. Section 4 presents the audio database used and the experimental results obtained. Finally, the paper is concluded with a discussion of the results and further improvements in Section 5.

## 2. SYSTEM OVERVIEW

Following flowchart demonstrates the building blocks of the sound source classifier implemented. First, silence regions are detected using energy and average zero-crossing rate measurements. Next, Onset detection is carried out for determining the exact starting instants of the acoustical events present in the signal. Acoustical features for each acoustical event is extracted and the classification algorithm is executed to label the sound sources present in the input audio signal. Training data obtained using an off-line training procedure is used for classification purposes.

The output of the system are the sound source labels for the audio file. These labels show the sound sources present at different time intervals(Figure X). Silence detection is carried out for a coarse estimation of the silence regions in the signal as shown in Figure X. Onset detection carries out a much more detailed process for determining the starting instant of each acoustical event(Figure X).

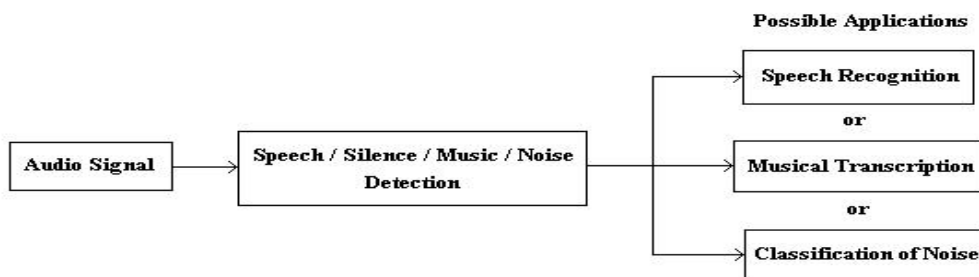


Figure X. Flowchart of the Sound Source Classification System

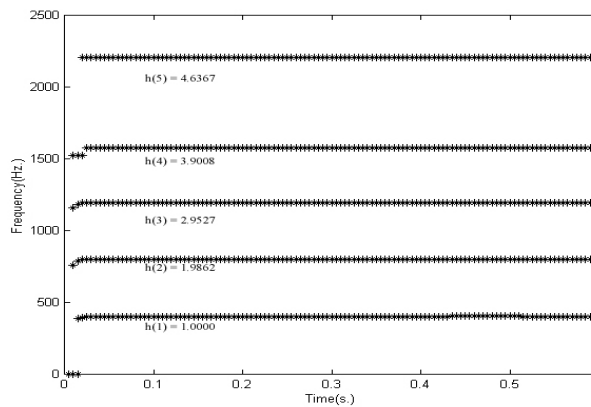
The harmonics detector is tested with a complex sinusoid containing pure sinusoidal components at 0.4, 0.8, 1.2, 1.6, and 2.0kHz respectively. Notice that the complex sinusoid has a pure harmonic structure in which first five harmonics of a 200Hz. sinusoid is present. Figure X illustrates the results of

f0 detection at each subband and the output harmonics vector obtained using the method described above. Note that  $h(n)$  are the average values obtained along the frames. As first four harmonics are detected successfully, these are used as the harmonic features for classification

**Figure X. Classification Output = Waveform + Spectrogram + Transcription**

**Figure X. Silence Detection = Energy + Zero-crossing Rate + Transcription**

**Figure X. Onset Detection = Waveform + Onset transcription by a human listener + Onset transcription by the system**



**Figure X. Harmonics Detection**

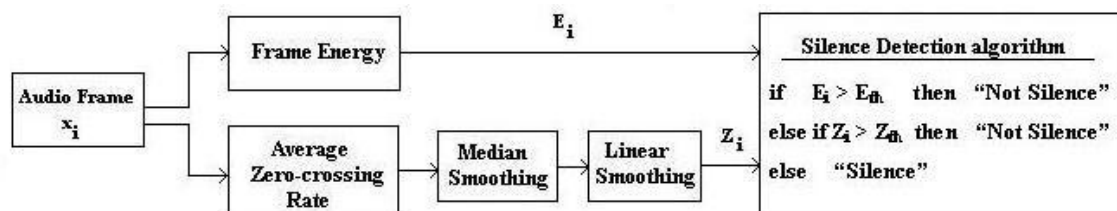
### 3. IMPLEMENTATION

#### 3.1 Silence Detection

Silence detection is carried out on a frame-by-frame basis by employing a hybrid silence detector that uses both energy and zero-crossing rate for detecting silent frames. Median filtering followed by linear filtering is used for smoothing purposes as

described in [Rabiner and Schafer, 1978]. The flowchart for silence detection is presented in Figure X.

$E_{th}$  and  $Z_{th}$  are determined considering the max and min values of the energy and zero-crossing rates respectively. After zero-crossing rate detection, 5 point running median filtering and linear filtering using a 3 point Hanning window as a filter is applied for smoothing.



**Figure X. Silence Detection Flowchart**

## 3.2. Onset Detection

### 3.2.1. Method

One of the very basic and crucial steps in Auditory Scene Analysis(ASA) is onset detection. It refers to the determination of discrete events, or groups in acoustic signals. Accuracy of onset detection is of crucial importance for any signal processing application that requires acoustical segmentation. Since the main objective in ASA is to segment the audio signal according to different acoustical events present in it, the accuracy of onset detection directly effects the overall success. Factors that seem to affect our perception of onsets are amplitude change(better if measured in terms of loudness), pitch change, timbre change or a change in frequency distribution of the audio signal. The challenges related to onset detection can be summarized as follows:

(i) The process should be able to detect the difference between gradual modulation changes and amplitude changes from *real* onsets. This problem leads to constraining the set of acoustic signals in some other algorithms.

(ii) Detecting all onsets one-by-one is difficult, whereas other closely related ‘high level’ properties may be easier to detect, i.e. beat of a sound. Long-term signal properties can be used to correct the errors that may be done in single detection for this purpose.

(iii) Incorrect onset times are likely to be detected with complex envelopes, such as slowly rising ones. Proposing a robust onset detection algorithm that can handle different types of acoustical signals is not easy. Criteria used for onset detection should be applicable to a wide range of sounds.

We follow the path of [Klapuri, 1999] and use knowledge from psychoacoustics to overcome these problems.

### 3.2.2 Psychoacoustics

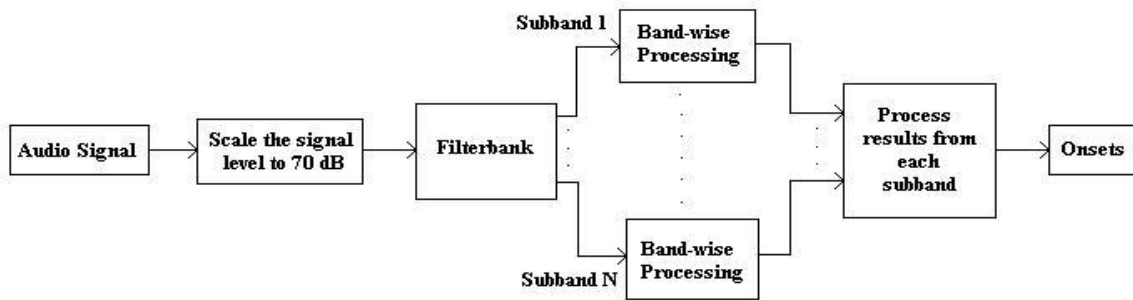
Before getting into the details of the onset detection algorithm, it will be useful to explain the psychoacoustical motivations behind it. The first motivation follows from the hypothesis that some

sort of cross-band integration (correlation) is carried out by our hearing system over audio signals, not simply summation across frequency bands as reported in [2]. Thus it is promising to consider subbands of the original signal trying to detect onsets in these subbands rather than processing the signal as a whole. In [2], a number of experiments were performed which show that if the acoustical signal is divided into a minimum number of four bands and the corresponding bands of a noise signal are modified using the subband signal amplitude envelopes, the resulting modified noise signal has a similar rhythmic perception as that of the original signal. Further support follows from the evidence that previous studies which tried to process signal amplitude as a whole (or less than 4 bands) did not yield satisfactory results.

Second point is the use of relative intensity increase. Psychoacoustically, not pure amplitude increases but increases in amplitude with respect to current signal amplitude level are important. It is easier to perceive the same amount of amplitude level change in a signal which is lower in amplitude level than a louder one. The third point stems from the fact that the time distance between each onset has a lowest bound. It is shown in several experiments that tone bursts which are closer in time less than some specific amount ( $\sim 5$  ms.) are not perceived as a discrete sequence of events. So onset candidates closer to another stronger onset candidate should be merged into the stronger one.

### 3.2.3. Algorithm Implementation

The algorithm shown below is suggested in [Klapuri, 1999]. It makes use of the psychoacoustical evidence discussed in section 2.2.2. First the signal level is scaled to 70 dB using the model of loudness described in [Moore et al, 1997]. Then, because of the reasons explained, the signal is decomposed into several subbands. Band-wise processing is carried out at each subband to detect onset candidates. The results from each subband are then combined to yield the final onset times. This is presented in Figure X.



**Figure X. Onset Detection Algorithm[Klapuri ,1999].**

In [Klapuri, 1999], 21 filters which cover the frequency range from 44 Hz to 18 kHz were used. Lowest three are one-octave and remaining are third octave band-pass filters. But from Scheirer's study, we know that the amplitude envelopes of the output of four filter banks are necessary and most of the time sufficient for pulse extraction. Using this fact, we have used five filters for subband decomposition in onset detection. The filters were sixth order elliptic filters, with 3 dB of ripple in the passband and 40 dB of rejection in the stopband. Center frequencies were 0.4, 0.8, 1.2, 1.6 and 3.2 kHz. The sampling frequency was 44.1 kHz. Bandwidths of these filters were chosen according to the human auditory system characteristics, i.e. for center frequencies less than 0.5 kHz the bandwidth was 0.2 kHz and for center frequencies greater than 0.5 kHz it was 0.2 times the center frequency.

Bandwise signal processing which is applied to each subband signal for onset detection can be described as follows:

- (i) The subband signal is half-wave rectified.
- (ii) The envelope of the amplitudes is calculated by convolving the half-wave rectified signal with a 200 ms half-Hanning window. This convolution is important because it does energy integration like our auditory system does. It emphasizes the most recent inputs and masks rapid modulations. This helps to solve problems created by rapid modulating signal effects.
- (iii) The signal is decimated to reduce the computational requirements. After this step, we still have the envelope information without unrequired high resolution.
- (iv) The first derivative of the decimated signal is calculated. This signal is used in many other algorithms to detect onsets. But as suggested in

[Klapuri, 1999], the relative difference function is calculated by dividing the derivative of the signal to itself. This method emphasizes the relative envelope increments, and it is a better property to detect onsets than pure derivative of the signal.

(v) Thresholding is applied at each band to drop out weaker onset candidates. For onset candidates closer than 50 ms, we choose the one with the higher intensity components. Intensity component of an onset candidate is the value of envelope between the onset and the point where the amplitude starts decreasing.

All the onsets from different bands are combined. For each onset candidate, the likelihood is calculated as follows: Within 50 ms of the candidate onset, all the intensities of the amplitude envelope of the acoustic signal are summed at the locations of onset candidates. Thereby calculating the possibility of the onset occurrence for each candidate, we eliminate those, which have less probability of occurrence within the 50 ms window. Thresholding is applied once more to get final onset times.

### 3.3 Acoustical Feature Extraction

The following features are used in order to classify audio signals. The number in each parenthesis shows the number of parameters extracted for each feature. Finally all parameters are collected to generate acoustical feature vectors for each segment of audio signals (a 24 x 1 vector for each frame). A training database is used to generate acoustical feature vectors for speech, musical instruments (guitar and piano), background noise, and mixtures of these. The training database is labelled manually to obtain accurate results. Off-line

training is carried out then. The acoustical features are as follows:

-*Harmonics(4)*. First 4 harmonics are used. Harmonic structure detection is carried out on a subband basis as explained in Section 3.3.1 below.

-*Energy in lower subbands(6)*. Total energy in the following subbands: [300 500], [500 800], [800 1600], [1600 2200], [2200 3200], [3200 6000] Hz

-*Energy envelope(2)*. Mean of energy envelope and variance of energy envelope.

-*Zero - crossing rate(2)*. Mean of zero-crossing rate and variance of zero-crossing rate.

-*Temporal features(4)*. Duration (rise time from 10 % of the maximum energy plus the decay time to the 10% of maximum energy), rise rate from 10 % of the maximum energy, decay rate to the 10 % of the maximum energy, decay rate to the 50 % of the maximum energy.

-*Spectral flux(6)*. Mean of the spectral flux of the low frequency half of the spectrogram, mean of the spectral flux of the high frequency half of the spectrogram, mean of the spectral flux of the whole spectrogram, variance of the spectral flux of the low frequency half of the spectrogram, variance of the spectral flux of the high frequency half of the spectrogram, variance of the spectral flux of the whole spectrogram.

### 3.3.1. Harmonics Detection

The simplest sound one can encounter in daily life is a pure tone. A pure tone can be defined as a signal that contains a single sinusoidal component. Most of the time the scene is much more complex: Many pure tones and noise can be present in a sound mixture. When a signal is composed of several pure tones it is regarded as a complex tone. *If the frequencies of the pure tones are integer multiples of a common basic or fundamental frequency, the resulting complex tone is called an harmonic complex tone* [Zwicker and Fastl, 1999]. The harmonic structure is an important feature of many audio signals. Music and speech signals may exhibit strong harmonic structures which can be used as a feature in automated recognition and transcription systems as well as in synthesis.

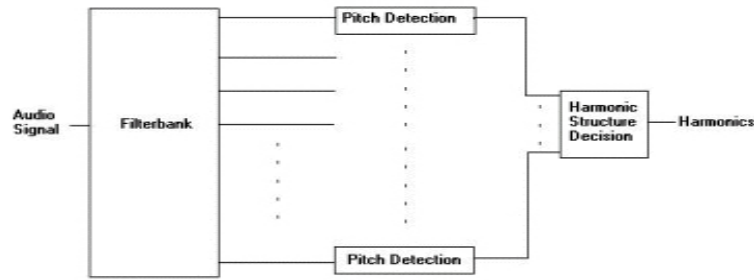
Harmonic structure detection is especially important for musical signal processing applications such as automated transcription and instrument identification. Most of the musical instruments (and

also human singing voice) exhibit a regular harmonic structure in the sounds they produce in which the peaks of the short term spectrum are related via an integer multiples relationship. Thus, correct detection of the harmonic structure is critical in the decision of the note(s) played by the instrument(s) and in the recognition of the instrument(s) that produced the sound.

In our implementation for harmonic structure detection, we have employed the psychoacoustic properties of the Human Auditory System(HAS) in a front end processing module. In basic terms, the human ear serves as a frequency analyzer. The cochlea which resides in the inner ear enables the frequency analysis procedure because different regions of the cochlea are sensitive to different frequencies. The overall auditory system can be modeled by a set of band-pass filters with center frequencies starting from 10 Hz to 20000 Hz. This range is the average frequency range that our auditory system is sensitive to. Any frequency outside this range can not be perceived by humans. A common used model for HAS is a set of band-pass filters. Each band is known as a critical band. The critical bands are defined around a center frequency in which the noise bandwidth is increased until there is a just noticeable difference in the tone generated at the center frequency [Swanson, et.al. 1998]. The critical bands are well studied in MPEG research and in the design of our filterbank we have used some of the central frequencies described in [ISO/MPEG]. The center frequencies of such a filter bank is also given in [Zwicker and Fastl, 1999].

Different filterbanks can be used for the harmonic structure detection process and we have tested two such filterbanks in this research. The output of each filterbank is processed separately to obtain harmonic candidates at each band. This is illustrated in Figure X.

An autocorrelation based approach is employed for pitch detection. This scheme is a variation of the log-lag correlogram method described in [Ellis, 1996]. The main idea behind the correlogram approach is to describe the signal in three domains : time, frequency, and autocorrelation lag. This gives a compact description of the signal using HAS characteristics. The peak values of the autocorrelation function of each bandpass filter output is used to estimate the pitch value for each critical band.



**Figure X. Flowchart for harmonics detection**

Harmonics decision is carried out as follows:

(i) The first subband  $f_0$  output that is nonzero is found. This value is denoted by  $f_0$ -i.e. it is the fundamental frequency of the first harmonic component.

(ii) The  $f_0$  outputs for the remaining subbands which correspond to higher frequencies are normalized with  $f_0$ . Let  $f(n)$  be the  $f_0$  value for the  $n$ th subband. Then after normalization the *normalized harmonic value for the  $n$ th subband* defined by the following equation is given by:

$$h_n = f(n) / f_0$$

We have 3 cases to consider for  $h_n$  :

(i)  $h_n = 0$  : This case corresponds to a non-harmonic (and also unvoiced) subband.

(ii)  $|h_n - K| \leq \epsilon$ , where  $K$  is the closest integer to  $h_n$  and  $\epsilon$  is a small value as compared to 1 ( $\epsilon = 0.05$  produced satisfactory results in our experiments) : In this case a harmonic is detected in the current subband (i.e.  $K$ th harmonic)

(iii)  $|h_n - K| > \epsilon$  : No harmonic detected. But the harmonicity value obtained in this subband is kept for classification.

We have also observed that when bandpass filters with sharp cut-offs are used, the subbands should be overlapped in the frequency domain by some amount. If strictly tuned bandpass filtered with no overlap is used, harmonic components located almost at the cut-off frequencies can be lost.

### 3.4. Classification

The aim of this part of the study is the classification of sounds generated by different sources using appropriate acoustical features. For

this purpose, a set of features describing both temporal and spectral characteristics of audio signals were determined. After the features are extracted from a set of samples as described in the previous section, the parametric classification scheme is trained with this training data. Then the classification algorithm is tested on a different group of sound samples (i.e. the test database).

The audio signal classes to be detected are as follows:

1. Glass sound, 2. Metal sound, 3. Plastic sound, 4. Wood sound, 5. Sound of object out on table, 6. Flute sound, 7. Piano sound, and 8. Guitar sound

The first four class of sounds are obtained from crushing two objects of the same kind. The fifth class of sound is obtained by putting a ceramic cup on a wooden table. The last three classes are musical instruments. Using these 24 features, both spectral and temporal characteristics are represented. The energy contour, zero-crossing rate and temporal features are estimated on a frame-by-frame basis. The frame size was 2.9ms with 50% overlap. (Sound data was sampled with 44100 Hz, and each frame contained 128 samples.)

#### 3.4.1 Method

The features extracted from the training data are first normalized to unity. Linear discrimination method with common covariance matrix is used as the classification scheme. 15 sample sounds per class was used for training and 5 samples per class was used for test for our case.

Following figures are scatter plots of some of the features:

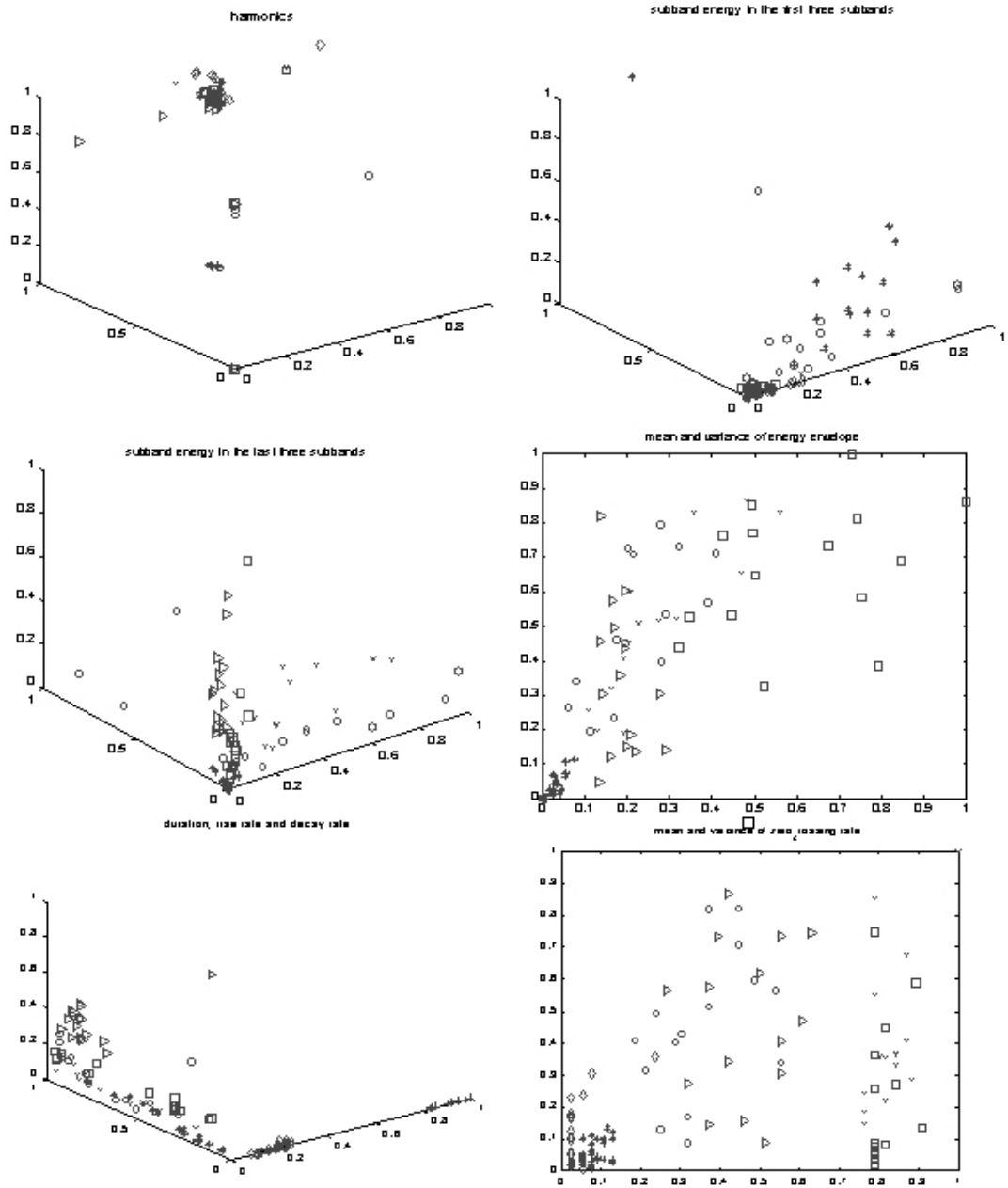


Figure. Scatter plots for some features

Harmonics			Energy Envelope			Spectral flux	
1	6.042		Mean	69.060		9.581	
2	6.888		Variance	46.700		16.520	
3	2.774					18.180	
4	4.890		Zero-crossing rate			36.660	
			Mean	377.500		44.820	
Lower Subband Energy			Variance	18.700		51.140	
1	38.080						
2	7.438		Temporal features				
3	12.610		Duration	1698.000			
4	20.230		Rise rate	46.070			
5	6.088		Decay rate(10%)	47.920			
6	27.260		Decay rate(50%)	17.600			

**Figure. F-ratios for the features**

#### 4. Experiments

Preliminary tests for the sound source classification system are carried out using a training and test database recorded for this study. In these tests, the main aim was to evaluate the performance of the sound source classifier in the case of different sounds from different objects. Tests on speech /silence /noise /music detection are being carried out and they have not been completed yet. 20 sound samples for each sound source is recorded using a microphone. The samples were 16 bit, 44100 kHz signals. 15 of the samples for each class is used in the off-line training process and 5 samples were used

in tests. As real flute and piano sounds were not available at the time of the tests, we used synthesized sounds generated by a MIDI program. However, as all the sounds produced by the program for a specific instrument are the same, different amounts of effects such as reverb or vibrato are added to the synthesized sounds to increase variability. Electric bass sounds were natural.

When all the 24 features were used in the classification stage, the following error rates given in Table X are obtained.

		Glass	Metal	Plastic	Wood	Object to table	Flute	Piano	E.Bass	Total
Train	Number of samples	15	15	15	15	15	15	15	15	120
	Error	1	0	0	0	0	0	1	2	4
	Percentage of success	93	100	100	100	100	100	93	87	96
Test	Number of samples	5	5	5	5	5	5	5	5	40
	Error	0	1	0	0	0	0	0	4	5
	Percentage of success	100	80	100	100	100	100	100	20	88

**Table X. Error rates**

From the error rates, we observe that the classification success for sounds from different

objects is almost perfect. The worst performance was obtained in the case of the bass sound. This was

expected since the sounds from other instruments were synthesized and naturally did not contain as much variation as the bass sounds did. In the on-going tests, more training data is being employed for bass sounds to increase the performance. The “object” sounds were simple as compared to instrument sounds, so the expected classification success for them was higher than the instruments.

Analysis of variance (ANOVA) is carried out first on the training data. The aim was to determine appropriate set of acoustical features for classifying different classes. For this purpose, the acoustical features are extracted frame by frame (window size = 25 ms, skip size = 10 ms using a Hamming window). Each combination of the vectors generated for each group is used in ANOVA with the hypothesis that the group means are equal. The f-ratio values obtained for each feature is given in the table below.

**TABLE. ANOVA f-ratio results**

An LBG vector quantizer is used to generate codebooks for each sound class to represent frame feature vectors. The size of the codebook is varied to obtain better classification results as demonstrated in the figure below.

**FIGURE. Codebook size vs. Classification success(%)**

## 5. CONCLUSION and FUTURE WORK

In this study, we have focused on the problem of general audio data transcription and implemented a system that can be used in the case of complex auditory scenes. Although the current system is a simple and insufficient one for the solution of such a complicated problem, basic building blocks can be improved in the future for better classification performance.

As stated in Section 4, further tests are being carried out on the system. For this purpose, new training data which contains many variants (i.e. same sound source in different recording environments) of each sound source should be collected and used.

New acoustical features can be incorporated into the current system. This is basically a design problem and appropriate acoustical features should be employed according to the case at hand. More

complex sound mixtures should be used in both training and classification stages to improve the system towards a more realistic classifier. This is strongly expected if one aims to transcribe “daily-life” audio data which may contain many sound sources such as speech, noise generated by different sources, music, machine sounds, etc.

Harmonics detector can be improved for automated musical transcription. For this purpose, the onset detector should also be employed. Harmonic structure can be extracted using musical information (i.e. chords) when the input audio signal contains music.

## References.

- [Bregman, 1990] A. Bregman. Auditory Scene Analysis. Cambridge MA : MIT Press. 1990.
- [Brown, 1998] J.C. Brown, “Computer Identification of Wind Instruments Using Cepstral Coefficients” in Proceedings of the 16<sup>th</sup> International Congress on Acoustics and 135<sup>th</sup> Meeting of the Acoustical Society of America, pp. 1889-1890, Seattle.
- [Cook, et.al 1997] C.D. Cook, D.J. Kershaw, J.D.M. Christie, C.W. Seymour, S.R. Waterhouse, “Transcription of Broadcast Television and Radio News: The 1996 Abbot System” ICASSP 1997, vol. 2, pp 723-726.
- [Gauvain, et.al 1997] J. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, “Transcribing Broadcast News Shows”, ” *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 1997, vol. 2, pp. 715-718.
- [Goto and Muraoka, 1996] M. Goto and Y. Muraoka, “Beat Tracking Based On Multiple-Agent Architecture- A Real-Time Beat Tracking System For Audio Signals”, ICMAS 1996, pp. 103-110.
- [ISO/IEC, 1993] ISO/IEC 11172-3:1993. Information technology -- Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s -- Part 3: Audio.
- [Jain et.al., 1996] Jain U. et.al, “Recognition of continuous broadcast news with multiple unknown speakers and environments” In *Proc. DARPA Speech Recognition Workshop*, Feb., 1996.
- [Klapuri, 1999] Klapuri, “Sound Onset Detection by Applying Psychoacoustic Knowledge”, *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 1999.
- [Martin, 1996] K.D. Martin, “A Blackboard System For Automatic Transcription of Simple Polyphonic Music”, MIT Media Laboratory Perceptual Computing Section Technical Report No.385, 1996.
- [Moore, et.al., 1997] B. Moore, B. Glasberg, T. Baer “A Model for the Prediction of Thresholds, Loudness, and Partial Loudness”, *J. Audio Eng. Soc.*, Vol. 45, No. 4, pp. 224-240, April 1997.
- [Naphade and Huang, ?] M.R. Naphade and T.S. Huang, . Stochastic Modeling of Soundtrack for Efficient Segmentation and Indexing of Video. Technical Report. Coordinated Sciences Laboratory and Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign. <milind@ifp.uiuc.edu, huang@ifp.uiuc.edu>
- [Rabiner and Schafer, 1978] L. Rabiner and R. Schafer. Digital Processing of Speech Signals. Prentice-Hall, Inc., Englewood Cliffs, New Jersey. 1978.

**[Saraceno and Leonardi, 1997]** C.Saraceno and R.Leonardi, "Audio as a Support to Scene Change Detection and Characterization of Video Sequences" *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 1997, vol 4, pp. 2597-2600.

**[Scheirer, 1998]** E.D.Scheirer. Music Perception Systems.Ph.D. Dissertation Proposal. Massachusetts Institute of Technology. 1998.

**[Spina and Zue, 1996a]** M.S. Spina and V.W.Zue.Automatic Transcription of General Audio Data : Preliminary Analyses.In *Proceedings of the International Conference on Spoken Language Processing*, pp. 594-597.

**[Spina and Zue, 1996b]** M.S. Spina and V.W.Zue.Automatic Transcription of General Audio Data. Spoken Language Systems Group. Laboratory For Computer Science. Massachusetts Institute of Technology.

**[Tsekeridou and Pitas, ?]** S.Tsekeridou and I.Pitas,"Speaker Identification for Audio Indexing Applications", Dept. of Informatics, Aristotle Univ. of Thessaloniki, Thessaloniki, Greece. <pitas@zeus.csd.auth.gr>

**[Zwicker and Fastl, 1999]** E.Zwicker and H.Fastl. Psychoacoustics.Springer-Verlag Berlin Heidelberg New York.1998