In compliance with the
Canadian Privacy Legislation
some supporting forms
may have been removed from
this dissertation.

While these forms may be included
in the document page count,
their removal does not represent
any loss of content from the dissertation.

Feature Extraction for Audio Classification

Submitted by

Ahmad Rami Abu-El-Quran, B.Sc.

A thesis submitted to the Faculty of Graduate Studies and
Research in partial fulfillment of the requirements for the
degree of

Master of Applied Science (M.A.Sc.)

Ottawa-Carleton Institute For Electrical and Computer Engineering
Department of Systems and Computer Engineering
Faculty of Engineering
Carleton University
Ottawa, Ontario
August, 2003

© Copyright 2003, Ahmad Rami Abu-El-Quran

National Library
of Canada

Bibliothèque nationale
du Canada

Acquisitions and
Bibliographic Services

Acquisitions et
services bibliographiques

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-88866-5

Our file Notre référence

ISBN: 0-612-88866-5

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

Canada

Abstract

This thesis proposes a new algorithm to discriminate between speech and non-speech audio segments. It is intended for security applications as well as talker location identification in audio conferencing systems, equipped with microphone arrays.

The proposed method is based on splitting the audio segment into small frames and detecting the presence of pitch in each one of them. The ratio of frames with pitch detected to the total number of frames is defined as the pitch ratio and is used as the main feature to classify speech and non-speech segments. The performance of the proposed method is evaluated using a library of audio segments containing female and male speech, and non-speech segments such as cocktail noise, footsteps, and traffic noise.

Two major contributions are proposed for the speech/non-speech discrimination. The pitch ratio algorithm is proposed to give robust decision for each class. Limiting the search to the human pitch range is also proposed to enhance the performance of the pitch ratio algorithm. Also, non-speech audio type classification using neural networks is proposed.

It is shown that the proposed algorithm can achieve a percentage of correct decision of 97% for the speech and 98% for non-speech segments, 0.5-seconds long.

Acknowledgements

I would like to express my appreciation for the valuable efforts of *Dr. Rafik Goubran* in supervising this research. I would also like to sincerely thank *Dr. Goubran* for his ongoing support and encouragement.

I would like to acknowledge the funding from Mitel, OGS, and NSERC. I would also like to thank *Darlene Hébert*, and *Christine McGregor* for there helps with the English review.

I am especially grateful to my professors at Ain Shams University, Cairo, Egypt, Carleton University, Ottawa, Canada and University of Ottawa, Ottawa, Canada, for the strong technical background that I have learned from my undergraduate and graduate courses. Also I would like to thank my colleagues who have helped me during my ongoing studies.

Finally, I would like to thank my family, especially my father *Mahmoud* and my mother *Salwa*, for their support and understanding and the sacrifices they have made through the years of this research. I would like to dedicate this thesis to my wife *Ruba* and my sons *Mahmoud* and *Zeyad*.

Table of Contents

ABSTRACT.....	III
ACKNOWLEDGEMENTS.....	IV
INTRODUCTION.....	1
1.1 Thesis Motivation.....	1
1.2 Problem Definition.....	4
1.3 Thesis Contributions.....	8
1.4 Thesis Outline.....	11
LITERATURE REVIEW	13
2.1 Audio Signal Classification Definition.....	13
2.2 History of the ASC	14
2.3 Main Features of the Audio Signal.....	15
2.3.1 The Energy	16
2.3.2 Sub-band Energy	17

2.3.3	Zero Crossing Rate	19
2.3.4	Periodicity (Pitch Detection)	20
2.4	Main Features of the Speech Signal.....	23
2.5	Decision Rule Using Threshold Boundaries.....	23
2.6	Speech/Non-Speech Classification	24
2.6.1	Voice Activity Detection (VAD).....	24
2.6.2	Audio Retrieval Using Speech/Non-Speech Decision	26
2.6.3	Conferencing and Security Applications Using Speech/Non-Speech Determination ..	27
<u>EXPERIMENTAL SETUP AND SIMULATION ENVIRONMENT</u>		<u>32</u>
3.1	Audio Library.....	32
3.2	Simulation Environment	36
3.3	Evaluation Criteria	36
<u>SPEECH/NON-SPEECH CLASSIFICATION USING THE PITCH RATIO</u>		<u>39</u>
4.1	Pitch Ratio Definition	39
4.2	Simulations and Results.....	42
4.3	Discussion.....	53
<u>SPEECH/NON-SPEECH CLASSIFICATION USING THE HUMAN PITCH RANGE.....</u>		<u>55</u>
5.1	Human Pitch Range.....	55
5.2	Simulations and Results.....	56
5.3	Discussion.....	64
<u>AUDIO CLASSIFICATION.....</u>		<u>66</u>
6.1	Audio Type Classification.....	66

6.2	Feature Vector.....	67
6.3	Neural Network.....	68
6.4	Simulations and Results.....	69
6.5	Discussion.....	74
<u>CONCLUSIONS AND FUTURE WORK.....</u>		75
<u>APPENDIX A: ACRONYMS.....</u>		78
<u>REFERENCES.....</u>		80

List of Tables

Table 2.3.1 Main Features of Audio Signal	15
Table 3.1.1 Audio Library Distribution	33
Table 5.2.1 Comparison Between Speech and Non-Speech Correct Decision Using Different Algorithms	61
Table 5.2.2 Comparison Between Different Audio Types Correct Decision Using Different Algorithms	62
Table 6.4.1 Training and Testing Data for Audio Classification.....	69
Table 6.4.2 Simulation Results for Two Classes	70
Table 6.4.3 Simulation Results for Four Classes Using 100 Hidden Nodes	71
Table 6.4.4 Simulation Results for Four Classes Using 50 Hidden Nodes	71
Table 6.4.5 Simulation Results for Four Classes Using 25 Hidden Nodes	72
Table 6.4.6 Simulation Results for Four Classes Using 10 Hidden Nodes	72
Table 6.4.7 Simulation Results for Four Classes Using 5 Hidden Nodes	73

Table 6.4.8 Simulation Results for Four Classes Using 2 Hidden Nodes	73
--	----

List of Figures

Figure 1.1.1 Microphone Array (Mitel™ Product)	3
Figure 1.1.2 ASC Algorithm to Be Used for Security Applications.	4
Figure 1.2.1 Main Steps of ASC for C Classes.....	5
Figure 1.2.2 Main Audio Classes.....	7
Figure 1.3.1 The Proposed Algorithm	9
Figure 1.3.2 Audio Library Classes	11
Figure 2.3.1 The Energy of the Audio Signal.....	17
Figure 2.3.2 Zero Crossing Rate of Audio Signal	20
Figure 2.3.3 The Pitch of the Audio Signal	21
Figure 2.6.1 VAD Using the Pitch Detection	25
Figure 2.6.2 Microphone Array Used in Focusing the Beam on the Speaker (Mitel™)..	28
Figure 2.6.3 (a) Example of Audio Detection of Non-Speech Audio Segment (Door Opening) (b) Zooming on the Part with Pitch Detected.....	30

Figure 2.6.4 Segmentation of the Audio Segment.....	31
Figure 3.1.1 The Directory of the Audio Library CD.....	35
Figure 3.3.1 Classification System	37
Figure 3.3.2 Algorithm Evaluation Methods	38
Figure 4.1.1 Example of Speech Audio Segment.....	40
Figure 4.1.2 The Relation between the Frame Numbers and the Total Duration of the Audio Segment.....	42
Figure 4.2.1 Percentage of Correct Decision vs. Pitch Ratio Variation (Frame Size 20 ms, Overlap Ratio 0.5, and Pitch Range 70 – 280 Hz)	43
Figure 4.2.2 Percentage of Average Correct Decision for different segment duration vs. Frame Size (ms) (Overlap 0.5, and Pitch Range 70 – 280 Hz).....	44
Figure 4.2.3 Percentage of Average Correct Decision vs. the Segment Size (ms) (Overlap 0.5, Pitch Range 70 – 280 Hz, and Frame Size 100 ms).....	45
Figure 4.2.4 Speech/Non-Speech Classification Regions (Frame Size 40 ms, Overlap Ratio 0.5, and Pitch Range 70 – 280 Hz)	46
Figure 4.2.5 Percentage of Average Correct Decision Variation vs. Frame Size (ms) (Overlap = 0.875, 0.75, and 0.5, and Pitch Range 70 – 280 Hz)	48
Figure 4.2.6 Percentage of Correct Decision for the Speech and the Non-Speech vs. Frame Size (ms) (Overlap Ratio 0.75, and Pitch Range 70 – 280)	49

Figure 4.2.7 Comparison of the Percentage of Correct Decision vs. Frame Size (ms)	
Using the Pitch Ratio and without the Pitch Ratio (Overlap Ratio 0.75, and Pitch Range 70 – 280 Hz).....	50
Figure 4.2.8 Comparison of Speech and Non-Speech Percentage of Correct Decision vs. Frame Size (ms) Using the Whole Audio Segment and without the Pitch Ratio (Overlap 0.75, and Pitch Range 70 – 280 Hz).....	51
Figure 4.2.9 Comparison Between the Percentage of Correct Decision vs. Frame Size (ms) for the Proposed Method and the Previous Methods (Overlap Ratio 0.75, and Pitch Range 70 – 280 Hz)	53
Figure 5.2.1 Comparison between the Pitch Ranges Percentage of Correct Decision vs. Frame Size (ms) (Overlap Ratio 0.6).....	56
Figure 5.2.2 Percentage of Correct Decision Variation vs. Pitch Range (Hz) (Lower Limit Fixed to 70 Hz and the Upper Limit is the Variable, Frame Size 70 ms, and Overlap Ratio 0.6).....	57
Figure 5.2.3 Histogram of the Number of Speech Audio Segments vs. Pitch Ratio (Frame Size 70 ms, Overlap Ratio 0.6, and Pitch Range 70 – 280 Hz).	59
Figure 5.2.4 Histogram of the Number of Non-Speech Audio Segments vs. Pitch Ratio (Frame Size 70 ms, Overlap Ratio 0.6, and Pitch Range 70 – 280 Hz).....	59

Figure 5.2.5 Histogram of the Number of Speech Audio Segments vs. Pitch Ratio (Frame Size 70 ms, Overlap Ratio 0.6, and Pitch Range 70 – 2000 Hz)....	60
Figure 5.2.6 Histogram of the Number of Non-Speech Audio Segments vs. Pitch Ratio (Frame Size 70 ms, Overlap Ratio 0.6 and Pitch Range 70 – 2000 Hz).....	60
Figure 5.2.7 Histogram for the Different Algorithms Correct Decision.....	64
Figure 6.1.1 Non-Speech Audio Type Classification	67

List of Symbols

α	Audio segment duration
β	The number of the frames that have pitch detected
γ	Frame size
η	The total number of frames
e_r	Ratio of signal energy in different frequency bands
e_p	Product of signal energy in different frequency bands
e_{low}	Fraction of total energy in the low frequency band
E_n	Short time average energy
E_{rm}	Root mean square energy
E_s	Log energy

n	Time index
N	Window length
$R(l)$	Autocorrelation function
$w(n)$	Window function
χ	Overlap ratio size
$x(m)$	Discrete time audio signal
Z_n	Zero crossing rate.

Chapter One

Introduction

1.1 Thesis Motivation

One of the most important human senses is hearing. The revolution of information technology has lead to attempts to simulate the human senses, as well as automate the detection of any event, in other words to create the ability to distinguish between different types of audio segments, essentially performing audio signal classification (ASC).

Over the last few years, ASC has piqued the interest of researchers [1]. ASC has become more important, responding to the need for more robust and effective ASC in multimedia and security applications.

We must first break the issues into smaller segments. By solving these smaller problems first, it will lead to a multi-dimensional system for the larger issues in future work.

Pitch detection algorithms occupy a wide area in speech processing research [1], [2], [3], [4]. Voice pitch is the most significant characterization of human speech.

Pitch detection can be utilized in the ASC application to classify the audio signal as speech or non-speech segments. However, the pitch detection algorithm is not sufficient as a stand alone algorithm to make the speech/non-speech decision. Further processing is needed to achieve the desired functionality.

There have been a wide variety of pitch detection algorithms proposed since the early 50's; these range from the time domain to the frequency domain as well as the cepstral domain [4].

A possible application of ASC is modern video conferencing using a microphone array and a steerable camera [5]. In this case, the microphone array detects the location of the sound source [6], passes the information to the steerable camera, which tracks the talker [7]. In that case, it is important to determine whether the sound source is a talker or background noise. This is achieved by analyzing the audio segment and determining

whether the segment is speech or non-speech. The performance of these systems critically depends upon the accuracy of this classifier.

The security applications can be achieved by classifying the audio segment to certain classes such as breaking glass, footsteps, and door opening. Before classifying the audio segment to a certain class, speech/non-speech discrimination is needed. Once this has been done, the non-speech segment can be classified to certain subclasses.

The audio source tracking capability [9], [10], [11], [12] of a microphone array, as shown in figure 1.1.1, can be utilized for security applications. Combining the tracking capability with the ASC can achieve the security applications [8].

Figure 1.1.1 Microphone Array (Mitel™ Product)

As a result, the new algorithm using the pitch detection algorithm can be implemented using the microphone array to classify the audio segment as speech or non-speech and then classifying the non-speech to a certain subclass, as shown in figure 1.1.2.

Figure 1.1.2 ASC Algorithm to Be Used for Security Applications.

1.2 Problem Definition

The objective of this research is to introduce an ASC scheme that achieves both better performance and lower complexity in speech/non-speech classification, to be used in the security area using a microphone array. In addition, it is favorable that the algorithm will be deployed in more than one environment. In order to achieve this objective, this research studied the dynamics of different ASC algorithms and the ways of proficiently using these dynamics in speech/non-speech determinations.

Figure 1.2.1 Main Steps of ASC for C Classes

Fundamentally, all ASC algorithms have the same main stages, as shown in figure 1.2.1. The first stage is the input of the audio segment, then a certain duration from the input audio segment is selected by the windowing. Feature extraction should be done for each window. Finally, according to the extracted feature value, the classification can be done.

The audio segments library is characterized by certain parameters. These parameters are described as follows:

- Duration: The time duration of the input audio segment.
- Sampling rate: Number of samples per second to the original analog signal.
- Resolution: The number of bits that represent each sample.
- Recording type: The number of channels, engaged in the recording process, for example, mono or stereo.

The windowing step divides the input audio segment into a smaller frame size. The windowing parameters can affect the values and the quality of the feature extraction step.

These parameters can be described as follows:

- Window length: The time duration of the window.
- Window overlapping: The percentage of the previous window incorporated into the new window.
- Window type: The function used in creating the window, for example hamming or rectangular window.

Feature extraction will be applied on each frame. The features can be calculated using mathematical operators such as multiplication, addition, or the gradient. They also can be calculated using statistical operators such as standard deviation, mean, or correlation. The feature may require both the mathematical and the statistical operators.

Another important step in the feature extraction is the normalization to some numerical scale, which is useful for the classification stage. More than one feature are calculated and assembled in what we call a feature vector.

Classification is the decision making process; it determines which set of the input feature vectors most likely belong to a certain class. The goal of the classification process is to find a certain classification threshold to make the right decision of different audio classes. Comparing the feature vector to the predefined threshold can give us the most likely decision for a certain class.

There are many types of classifiers, such as fixed threshold [1], [13], neural networks [14], [15], [16], [17], [18], [19], Gaussian Mixed Model (GMM) [20], [21], Hidden Markov Model (HMM) [22], [23], and "adaboost" [24].

Audio classes can be described in different ways; some of these ways overlap in some subclasses. The main audio classes are shown in figure 1.2.2.

Figure 1.2.2 Main Audio Classes

Speech is produced by the human vocal tract vibrations. Speech can be categorized into languages, gender or to certain speakers.

The definition of music in this section incorporates only music that is produced from man-made instruments. Birdsong is considered to be environmental sound, not music.

Noise is random or pseudorandom signals that can be classified using the distribution of the energy in the spectrum of the signal. White noise contains a uniform energy distribution, while colored noise contains a non-uniform energy distribution. Famous examples of colored noise are the pink and brown noises [25].

Cocktail noise is defined as a number of sound sources with almost the same energy level produced at the same time. These sources may be a speech or environmental. An example of cocktail noise is the sounds that can be found in the classroom before the professor enters.

In this research, environmental sounds are defined as sounds that come from nature, such as rain, wind and the bird sounds. Also, sounds such as traffic and a door closing are assumed to be environmental sounds.

1.3 Thesis Contributions

The major contribution of the research presented in this thesis is to introduce a novel speech/non-speech algorithm that achieves better performance in the security applications using a microphone array. Furthermore, the architecture of this algorithm, as shown in figure 1.3.1, incorporates in its design two major contributions and one minor contribution. The following is a summary of the contributions of the research reported in this thesis dissertation:

Figure 1.3.1 The Proposed Algorithm

◦ Pitch Ratio Calculation

Pitch ratio calculation, as proposed and studied in this thesis, introduces and is incorporated in the speech/non-speech classification. Calculations are initiated by splitting the audio segment into small frames and detecting the presence of pitch in each one of them. The ratio of frames that have pitch detected to the total number of frames is defined as the pitch ratio.

$$\text{Pitch Ratio} = \frac{\text{Number of the frames that have pitch}}{\text{Total number of frames in the audio segment}} \quad (1.3.1)$$

Unlike other speech/non-speech classification [5], which tracks the pitch or calculates the speech signal duration, this algorithm calculates the ratio of pitched part duration in

the audio segment to the total duration of the audio segment. The speech/non-speech decision can be achieved with better performance using the pitch ratio algorithm.

- *The Human Pitch Detection*

The second major contribution is human pitch detection. This is done by limiting the search to the human pitch range. Using this method search, the speech/non-speech classification can be significantly improved. This method overcomes the misclassification problems of some non-speech audio segments, which contain pitch value out of the human pitch range value.

- *The Audio Library*

The audio library is the minor contribution. The significance of the audio library comes from its wide range of audio segment classes.

The audio library contains speech and non-speech audio segments as outlined in figure 1.3.2. The non-speech samples cover a wide variety of environmental and noise audio classes, which relate to the security field. Also, it is important to note that music is out of the scope of this research.

Figure 1.3.2 Audio Library Classes

1.4 Thesis Outline

The contents of this thesis are organized as follows:

Chapter 2 contains a literature review of generic audio signal classification algorithms and existing speech/non-speech discrimination algorithms.

Chapter 3 contains the experimental setups of the research. Chapter 4 and chapter 5 contain the pitch ratio algorithm and human pitch range algorithm, respectively. The

implementation of these algorithms along with simulation results are also presented in these chapters, and followed by a discussion.

In chapter 6, audio classification using the neural networks for the non-speech audio segments is explained. Simulation results on the audio classification algorithm are given.

Finally, chapter 7 is devoted to concluding remarks and suggestions for future research topics.

Chapter Two

Literature Review

In this chapter, the generic models of the audio signal classification algorithms are presented. Also, focus on the work done in the area of speech non-speech classification is given. This study depends on the common features and classification principles, found in the literature.

2.1 Audio Signal Classification Definition

Audio signal classification is defined as a process, which discriminates between a number of audio types such as environmental sounds, music, or speech.

2.2 History of the ASC

Speech has always been the most attractive way of communicating between humans. If we go back in history, we will find that the first electric communication device used Morse signals. Later, Alexander Graham Bell invented the microphone and, in 1876, his "electrical speech machine," known now as the telephone. By 1878, Bell had finished the first telephone exchange in New Haven. From that time, audio and more particularly speech, became important research areas [26].

Bell Laboratories introduced automatic speech recognition (ASR) in 1952. Various researches have investigated ASR methods. Since that time, an algorithm to distinguish between speech and non-speech was observed.

In 1959, Bell Laboratories introduced Time Assignment Speech Interpolation (TASI). The TASI system used listening time and other natural pauses in a conversation to transmit portions of other conversations, thus increasing the telephone circuit capacity of communication channels. Since the introduction of the TASI system, the need for a robust Voice Activity Detection (VAD) has been increasing [26].

VAD algorithms are engaged in many speech-processing applications, such as speech coding, noise estimation and cancellation, audio conferencing, and automatic speech recognition [27].

ASC has caught the attention of researchers in the past few years [1], [13], [28]. As the demand grows for multimedia and security applications, the need for robust and effective ASC becomes more apparent.

Research on ASC lags behind that of speech recognition, due to the difficulty in classifying the non-speech audio[1].

2.3 Main Features of the Audio Signal

The audio signal is characterized by certain features. The main features are summarized in table 2.3.1.

Table 2.3.1 Main Features of Audio Signal

Energy Function	Zero Crossing Rate (ZCR)	Fundamental Frequency (Pitch)
<ul style="list-style-type: none">• Used to distinguish between voiced speech and unvoiced speech, which has lower energy levels.• Distinguish silence when SNR is high.	<ul style="list-style-type: none">• Voiced period has much smaller ZCR than unvoiced period.• ZCR of music has lower variance and average amplitude than speech.	<ul style="list-style-type: none">• Speech components contain harmonics in general while non-speech components do not contain harmonics.• Most environmental sounds are non-harmonics.

2.3.1 The Energy

Short time average energy E_n provides us with the representation of the amplitude variation over a period of time [29]. The energy-based algorithms can provide a certain threshold for speech/non-speech classification [30].

$$E_n = \frac{1}{N} \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \quad (2.3.1)$$

$x(m)$ is the discrete time audio signal, n is the time index and $w(n)$ is the window function with length N [31].

Figure 2.3.1 The Energy of the Audio Signal.

The difficulties of speech detection in the low SNR can be overcome by using the sub-band energies [3].

2.3.2 Sub-band Energy

This method depends on the calculation of the energy in different frequency bands. This is done by using different types of filters (LPF, HPF, and BPF).

The main issue with using this method is that the energy distribution of the speech spectrum is concentrated in the low frequency band, while the non-speech is distributed on the whole frequency band [32],[33].

As proposed in [14], the ratio of signal energy e_r is defined as the ratio of signal energy above frequency $f1$ (defined as $e_+(f1)$) to signal energy below frequency $f2$ (defined as $e_-(f2)$). e_r is defined as follows:

$$e_r = \frac{e_+(f1)}{e_-(f2)} \quad (2.3.2)$$

The product of signal energy e_p is defined as follows:

$$e_p = e_+(f1) * e_-(f2) \quad (2.3.3)$$

Also, a fraction of total energy of the low frequency band E_{low} is defined as follows:

$$E_{low} = e_-(f2) / E_n \quad (2.3.4)$$

where E_n is the total energy [15].

The filter bank can be utilized to calculate the sub-band energy. The filter bank is also utilized in designing the wavelets which can be used in many audio processing areas [2], [34], [35].

2.3.3 Zero Crossing Rate

Zero crossing rate (ZCR), as shown in figure 2.3.2, is said to occur if successive samples with different signs are observed [36]. The unvoiced part has a high ZCR, while the voiced part has a lower ZCR [1], [37], [38], [39]. Short time average ZCR is defined as follows:

$$Z_n = \frac{1}{2} \sum_{m=-\infty}^{\infty} [\text{sgn}[x(m)] - \text{sgn}[x(m-1)]] w(n-m) \quad (2.3.5)$$

where

$$\text{sgn}[x(i)] = \begin{cases} 1 & x(i) \geq 0 \\ -1 & x(i) \leq 0 \end{cases} \quad (2.3.6)$$

$w(n)$ is the window function with length N. [36]

The major drawback of the ZCR is the sensitivity to the background noise.

Figure 2.3.2 Zero Crossing Rate of Audio Signal

2.3.4 Periodicity (Pitch Detection)

The fundamental frequency (pitch), as shown in figure 2.3.3, is an important parameter for the speech [3]. The periodicity can be determined by detecting similarities in the signal waveform [3], [40], [41], [42], [43].

Figure 2.3.3 The Pitch of the Audio Signal

Pitch detection is done by comparing the similarity of the waveform and the shifted version of the same waveform. The maximum similarity happens if the shifted distance is equal to the pitch period.

The pitch determination can be carried out in both the time domain and the frequency domain. A comparison between the different methods can be found in [4].

Autocorrelation method is one of the most effective methods for pitch detection. This method detects the maxima of the autocorrelation function. The autocorrelation is described as follows:

$$R(l) = \sum_{m=1}^{N-|l|} x(m)x(m+|l|), \quad |l| = [0, 1, \dots, N-1] \quad (2.3.7)$$

$R(l)$ is the autocorrelation function and $x(m)$ is discrete time audio signal with length N [3].

Autocorrelation-based pitch detectors are noted to perform well in the presence of white noise; in [3] a simulation was carried out for pitch detection over a range of SNR $[\infty, -18 \text{ dB}]$.

Before applying the autocorrelation on the audio signal, a central clipping is used, which is described in [4]. The center clipping is defined as follows:

$$s(m) = \begin{cases} x(m) - c_l & \text{if } x(m) \geq c_l \\ x(m) + c_l & \text{if } x(m) \leq -c_l \\ 0 & \text{otherwise} \end{cases} \quad (2.3.8)$$

where

$$c_l = k \min[c_{l1}, c_{l2}] \quad (2.3.9)$$

k is a constant, c_{i1} and c_{i2} are the maximum amplitudes in the first and last third of the frames respectively [25].

The autocorrelation method works perfectly in a white noise environment. Also, it is designed to detect speech in the frequency range [0, 900 Hz] [4]. The main problem in designing ASC, depending on periodicity algorithms, is the sensitivity to any periodic signal.

2.4 Main Features of the Speech Signal

Speech is produced by the movement of the vocal tract as a response to an excitation signal [36]. The signal may be periodic at the glottis or noise-like due to the deformation in the vocal tract [44].

The voiced part has periodicity characteristics and the unvoiced part has noise-like behavior [33], as shown in figure 2.6.1.

2.5 Decision Rule Using Threshold Boundaries

This discipline is concerned with classifying the objects (in this thesis the object is the audio segment) into a number of classes [45], [46].

Each class must be known a priori. The features, needed to perform the classification, may be mean, standard deviation, covariance, etc.; each object has l features $x_i, i=1..l$. These features are assembled in a feature vector [47]; the feature vector is described as follows:

$$x = [x_1, x_2, \dots, x_l] \quad (2.5.1)$$

The decision threshold between classes is known as classifier.

2.6 Speech/Non-Speech Classification

Speech/non-speech classification can be categorized into three major classes. The key issue in this categorization is the time duration needed to make the decision.

2.6.1 Voice Activity Detection (VAD)

The VAD period is usually in the order of 20 to 40 milliseconds [41], [48], [49]. The VAD is a process that separates the speech from the noise[50], otherwise known as distinguishing the voiced from the unvoiced periods [33], as shown in figure 2.6.1.

Figure 2.6.1 VAD Using the Pitch Detection

VAD is utilized in speech processing applications. The major examples of these applications are speech enhancement [3], [29], [30], speech recognition [39], [51], and speech coding [32], [36], [50], [52].

Many types of the VAD have been proposed to achieve better delay, accuracy and sensitivity performance. The earlier algorithms of the VAD are based on energy level and

pitch detection [3], and cepstral features [32], [53]. These algorithms have been implemented in particular applications [33], [50].

The major problem in VAD algorithms is the noisy environment; many algorithms have been developed to overcome this problem. One of the solutions is the high order statistics [39].

2.6.2 Audio Retrieval Using Speech/Non-Speech Decision

The multimedia database, stored as digital audio, is growing every day. This has increased the need for an efficient technique of audio retrieval, based on the audio content.

Speech/non-speech algorithms can help classify the audio into two classes. After the initial categorization, the audio inputs are further sorted into subclasses [1],[13]. Using the information provided in the previous step, each audio type will be processed according to its class.

The audio retrieval does not need to be a real time process, therefore the audio segment duration may be in the order of 15 seconds [54]. The audio retrieval can be done using hierarchical classification [55]. Using hierarchical classification, the input audio

segment is first classified into a certain subclass; the search will then be more effective due to the reduction of the search space [1].

The silence ratio (SR), proposed in [1] and [13], is used as the main feature to distinguish between speech and non-speech. Also, the ZCR is used in combination with the SR to get better results.

To be considered as a silence period, a sample must satisfy three conditions. These three conditions are as follows:

- Amplitude threshold: the threshold of the maximum sample level to be considered as inaudible.
- Time threshold: the time duration that the signal must remain under the amplitude threshold before it is considered to be a silence period.
- Silence ratio threshold: the ratio of the sum of the durations of the silence periods to the duration of the audio segment. The audio segment is considered to be speech if the silence ratio is larger than the silence ratio threshold [13].

2.6.3 Conferencing and Security Applications Using Speech/Non-Speech Determination

Audio teleconferencing combined with video tracking is one of the growing research areas [10] ,[11] ,[12], [56], [57]. The microphone array can process a high quality audio

signal from certain locations by focusing a beam on the desired location and attenuating the undesired noise from other locations, as shown in figure 2.6.2 [5].

Figure 2.6.2 Microphone Array Used in Focusing the Beam on the Speaker (Mitel™)

Cameras, used in conference rooms, can be used to track the movement of the speaker, with the help of speech source localization [10]. This functionality can be interrupted by noise interference such as footsteps, drawer sounds or door sounds.

Speech/non-speech determination can solve the problem of noise interference. Functionality can be improved by prohibiting the camera from moving toward the audio source if it is not a speech source. The time period required for the decision making is in the order of 125 milliseconds [5].

Speech is assumed to be a stationary process in the time period 20-40 milliseconds [36]. This time duration is not long enough to make the speech/non-speech decision because the periodicity which can be found in short time duration of the non-speech can cause misclassification of the non-speech, as shown in figure 2.6.3.

Reliable speech/non-speech determination can be achieved by splitting the audio segment into smaller frames then getting the decision of the segment according to the heuristic decision, as shown in figure 2.6.4.

The work done in [5] calculates three major features in the audio segment to make the speech/non-speech decision. The algorithm proposed in [5] is summarized in the following points:

- Energy of the speech signal should change smoothly, without abrupt changes. Segments with large energy variation are assumed to be a non-speech.
 - The speech duration of the English word is usually in the order of 70 milliseconds [5].
 - Pitch tracking of the speech signal is the main feature in this algorithm. The audio segment is classified as a speech signal if the change in pitch is not fast.
-

Figure 2.6.3 (a) Example of Audio Detection of Non-Speech Audio Segment (Door Opening)
(b) Zooming on the Part with Pitch Detected.

Figure 2.6.4 Segmentation of the Audio Segment

Security applications can be achieved by combining the tracking capability of the microphone array with capability of audio classification. The first stage of audio classification is the speech/non-speech classification. The decision of the audio type can be made after the audio is classified as non-speech. The non-speech class must include the audio segments for the security applications, as shown in figure 1.3.2.

Chapter Three

Experimental Setup and Simulation Environment

In this chapter, the environmental setup and simulation are presented. The audio library and its classes are discussed in the following sections. The evaluation criteria for the proposed algorithm are also discussed.

3.1 Audio Library

The audio library, as shown in figure 1.3.2, contains two main classes, the speech class and the non-speech class. Also, the speech class contains two subclasses male and female.

The non-speech audio subclasses, such as breaking glass and drawer opening sounds are to be classified as non-speech in the first stage of the classification. Once this has been done, the non-speech classification of each audio sample will be farther classified into its own audio type.

Table 3.1.1 shows the number of audio segments for each audio type in the database.

Table 3.1.1 Audio Library Distribution

Audio Type	Number of audio segments
Male	275
Female	275
Total of speech files	550
Chair	20
Door	40
Drawer	20
Fan noise	30

Footsteps	41
Breaking glass	40
Rain	50
Cocktail noise	20
Thunder	40
Traffic	20
Water	20
Wind	20
Total of non-speech audio files	381
Total audio library Files	911

A database containing a total of 911 speech and non-speech audio segments has been collected. The specifications of the database are shown in the following points:

- Audio segment duration: 0.5 seconds.
 - Sampling rate: 16000 Hz.
-

- Resolution: 16 bits.
- Recording type: Mono (one channel).

The speech class contains clean speech of relatively high signal to noise ratio. The security applications are not required to work under noisy conditions in our proposal.

The audio database is recorded on compact disc (CD) with the standard specifications. The directories of the CD are created as shown in figure 3.1.1.

Figure 3.1.1 The Directory of the Audio Library CD.

3.2 Simulation Environment

The simulation of the proposed algorithm is done using Matlab™ software, the operating system is Microsoft Windows™, and hardware is the Intel™ P4 processor with speed of 1.5 GHz.

The audio files are saved as wav files (Windows PCM). Each audio file has the duration of 0.5 seconds. The wav files are introduced to the Matlab™ software as a batch file. The pitch ratio of each file is saved in an output text file.

3.3 Evaluation Criteria

Each audio segment in the audio library is tested in the simulation and the pitch ratio of it is stored in output text file. To determine an output decision, a threshold must be established. This threshold is the pitch ratio which maximizes both the speech and the non-speech percentage of correct decision. This is done by maximizing the average of the percentage of correct decision for speech and non-speech classes.

The input file type is already known. A comparison between the decision of the output file type and the input file type will provide the percentage of correct decision for each file, as shown in figure 3.3.1

Figure 3.3.1 Classification System

The proposed algorithm performance will be evaluated. A comparison of the performance of the algorithm with and without the thesis contributions is done. A comparison of the performance with the previous algorithms under the same conditions is also done, as shown in figure 3.3.2.

Figure 3.3.2 Algorithm Evaluation Methods

The performance evaluation is done by applying a number of known audio segments to the system and calculating the percentage of correct decision. A comparison between the known audio segment and the system decision is made; if the system output is the same as the known audio segment, then the decision will be correct.

Chapter Four

Speech/Non-Speech Classification Using the Pitch Ratio

Speech/non-speech classification using pitch ratio is introduced in this chapter. In the following sections, pitch ratio algorithms are developed; simulations and results are presented. A discussion and concluding remarks are given at the end of the chapter.

4.1 Pitch Ratio Definition

Pitch is one of the important features of human speech. It must be noted that there is a part of human speech that doesn't have pitch. In this research thesis, the aim is not to differentiate the voiced parts from the unvoiced parts. This research is concerned with speech/non-speech classification [56].

The speech input segment contains voiced parts and unvoiced parts, as shown in figure 4.1.1. As suggested by the aforementioned fact, not all the durations of the input segments contain pitches. The proposed method is based on splitting the audio segment into small frames and detecting the presence of pitch in each one of them. The ratio of frames with pitch detected to the total number of frames is defined as the pitch ratio and is used to classify speech and non-speech segments.

Figure 4.1.1 Example of Speech Audio Segment

The non-speech audio segment may contain pitch (music is not considered in this research), as shown in figure 2.6.3. So we can not make our decision on the pitch detection as a stand alone feature, because some of the audio segments will be classified as a speech.

The proposed pitch ratio algorithm can overcome the misclassification due to the pitch found in small time duration in non-speech segments. The pitch ratio classification takes into consideration the pitch found in the non-speech audio segments. The pitch ratio algorithm assumes that the speech audio segment has pitch intervals of longer duration than the non-speech audio segments.

Pitch ratio can be defined as the ratio of frames that have pitch detected to the total number of frames in the same segment.

$$\text{Pitch Ratio} = \frac{\beta}{\eta} \quad 0 \leq \text{Pitch Ratio} \leq 1 \quad (4.1.1)$$

where β is the number of frames that have pitch detected and η is the total number of the frames in the same segment.

The total duration of the frames should be equal to or greater than the audio segment duration; this is due to the overlap between the frames, as shown in figure 4.1.2.

Figure 4.1.2 The Relation between the Frame Numbers and the Total Duration of the Audio Segment

The total number of frames is calculated as follows:

$$\eta = \left\lfloor \frac{\alpha - \gamma}{1 - \chi} \right\rfloor \quad (4.1.2)$$

where α is the audio segment duration, γ is the frame size and χ is the overlap size. $\lfloor \bullet \rfloor$ is the round down operator.

4.2 Simulations and Results

To study the performance of the pitch ratio algorithm, a simulation was performed and simulation results of the speech/non-speech classification are presented. The simulation covers the performance of the classification for both speech and non-speech. Finally, a comparison of the performance with and without pitch ratio will be discussed.

For the speech/non-speech classification, we need certain thresholds. The upper threshold is fixed to 1 and the lower threshold splits the audio space into two regions. In this thesis, the maximization of the average decision of the speech and non-speech will be considered as a threshold condition.

Figure 4.2.1 Percentage of Correct Decision vs. Pitch Ratio Variation (Frame Size 20 ms, Overlap Ratio 0.5, and Pitch Range 70 – 280 Hz)

The percentage of correct decisions versus the pitch ratio variation is shown in figure 4.2.1. The average percentage of correct decision has a maximum value of 95.6% at pitch ratio value 0.31; this value is considered as the threshold.

The length of the audio segment affects the percentage of average correct decision as shown in figure 4.2.2.

Figure 4.2.2 Percentage of Average Correct Decision for different segment duration vs. Frame Size (ms) (Overlap 0.5, and Pitch Range 70 – 280 Hz)

In the figure 4.2.2 it is shown that, the longer segment size, the higher the percentage of correct decision, this also can be shown in figure 4.2.3.

Figure 4.2.3 Percentage of Average Correct Decision vs. the Segment Size (ms) (Overlap 0.5, Pitch Range 70 – 280 Hz, and Frame Size 100 ms)

Maximizing the average percentage of correct decision can split the audio space into two regions, as shown in figure 4.2.4.

Figure 4.2.4 Speech/Non-Speech Classification Regions (Frame Size 40 ms, Overlap Ratio 0.5, and Pitch Range 70 – 280 Hz)

The percentage of correct decision varies with the frame length. The reason that the percentage of correct decision for non-speech varies with the frame length is the periodicity in a short time frame, as shown in figure 2.6.3 [5]. Also, it is well known that the speech process is assumed to be stationary in the range of 20-40 milliseconds [27]. The previous two factors affect the classification decision. If the frame is small the misclassification possibility for the non-speech will be high due to the periodicity which

can be found in the small time duration frame; also, longer frame size is worse for the stationary condition of the speech process.

Simulation is done to study the effect of the frame size on the percentage of correct decision using different overlap ratios. As shown in figure 4.2.5, the average percentage of correct decision does not always have a value higher than 96% in the frame size 20-100 milliseconds. Also, this figure shows the effect of the overlap ratio on the average percentage of correct decision, the higher the overlap ratio, the higher the percentage of correct decision. These results are the output of the simulation using the 16 KHz audio segments.

Figure 4.2.5 Percentage of Average Correct Decision Variation vs. Frame Size (ms) (Overlap = 0.875, 0.75, and 0.5, and Pitch Range 70 – 280 Hz)

The speech percentage of correct decision variation with the frame size is shown in figure 4.2.6. The behavior of the curve is not smooth but we can see that the percentage of correct decision is maximized in the region of 20-130 milliseconds frame size and it starts to decay as the frame size approaches the input audio segment size.

**Figure 4.2.6 Percentage of Correct Decision for the Speech and the Non-Speech vs. Frame Size (ms)
(Overlap Ratio 0.75, and Pitch Range 70 – 280)**

The non-speech behavior, figure 4.2.6, is maximized in the region of 20-130 milliseconds frame size, then it starts to decay as the frame size approaches the size of the input audio segment.

Figure 4.2.7 Comparison of the Percentage of Correct Decision vs. Frame Size (ms) Using the Pitch Ratio and without the Pitch Ratio (Overlap Ratio 0.75, and Pitch Range 70 – 280 Hz)

The average percentage of correct decision of the speech/non-speech classification using the pitch ratio is compared to the behavior of the speech/non-speech classification without the pitch ratio. The classification without the pitch ratio is calculated by testing each frame in the audio segment and if there is pitch detected in any frame in the audio segment it is assumed to be speech as shown in figure 4.2.7. In this figure the curve begins to descend after the frame size reaches 70 ms in the pitch ratio case this is due to

the fact that the number of frames are decreased, thus reducing quantity of data to be processed.

Figure 4.2.8 Comparison of Speech and Non-Speech Percentage of Correct Decision vs. Frame Size (ms) Using the Whole Audio Segment and without the Pitch Ratio (Overlap 0.75, and Pitch Range 70 – 280 Hz)

In the previous work [1], the speech decision is made if pitch is detected in the whole audio segment (figure 4.2.8). A comparison between the speech/non-speech discrimination using the whole audio segment and the speech/non-speech discrimination using the frame pitch detection without the pitch ratio is presented in figure 4.2.8.

Speech percentage of correct decision using the frame without pitch ratio has a high value in the range 20 – 250 milliseconds. This case happens because the speech audio segments must have a pitch in one of the frames, in the case of the small frames. But when the frame duration exceeds 250 milliseconds, the frame starts to lose the stationary characteristics, as shown in figure 4.2.8.

The non-speech percentage of correct decision using the frame without pitch ratio has a low value at the small frame size. This happens because the non-speech audio segment has a periodicity in the short duration. So the non-speech audio segments will be classified as speech. When the frame starts to be larger, the periodicity starts to disappear; this will enhance the percentage of correct decision.

Speech/non-speech discrimination using the whole audio segment is not a function of the frame size, as shown in figure 4.2.8.

The proposed method achieves a high percentage of correct decision concurrently for both the speech and the non-speech decisions. The results of the percentage of correct decision for the speech and non-speech is higher than 97% in the frame range 20 - 100 milliseconds, as shown in figure 4.2.9.

Figure 4.2.9 Comparison Between the Percentage of Correct Decision vs. Frame Size (ms) for the Proposed Method and the Previous Methods (Overlap Ratio 0.75, and Pitch Range 70 – 280 Hz)

4.3 Discussion

The pitch ratio algorithm is introduced. The pitch ratio algorithm show better performance for the speech/non-speech discrimination than the speech/non-speech discrimination without the pitch ratio. Also, the algorithm shows a better performance than using the whole segment to make the decision for the speech/non-speech discrimination.

The pitch ratio algorithm can be utilized in the security application as well as audio conferencing room, because of the discrimination capability between speech and non-speech for the audio types which can be found in these areas. These results are the output of the simulation using the 16 KHz audio segments from the database.

Chapter Five

Speech/Non-Speech Classification Using the Human Pitch Range

The effect of the human pitch on the speech/non-speech classification is introduced. In the following sections, a human pitch range algorithm is developed, and simulations and results are presented. A discussion and concluding remarks are given at the end of the chapter.

5.1 Human Pitch Range

Human pitch range has different values for every human being. The pitch range is governed by the larynx. The pitch range for a male is 50 – 250 Hz and for a female 120 – 500 Hz [36]. The pitch is changed according to stress, emotions and intonation.

Limiting the search for the pitch to the human pitch range can overcome the misclassification of some audio types of non-speech. So not only the pitch detection but also the value of the pitch found will be used in the speech/non-speech classification.

5.2 Simulations and Results

Limiting the pitch to the human pitch range affects the percentage of correct decision. The highest correct decision is achieved by limiting the search to the human pitch range.

Figure 5.2.1 Comparison between the Pitch Ranges Percentage of Correct Decision vs. Frame Size (ms) (Overlap Ratio 0.6)

As shown in figure 5.2.1, the lower limit of the pitch range is fixed to 70 Hz and the upper limit is variable. The highest percentage of correct decision can be achieved with in the upper pitch limit value of 200-340 Hz. The value of the pitch range 70-340 Hz agrees with the human pitch range. This result agrees with our expectation that limiting the pitch search to the human pitch range will improve the percentage of correct decision.

Figure 5.2.2 Percentage of Correct Decision Variation vs. Pitch Range (Hz) (Lower Limit Fixed to 70 Hz and the Upper Limit is the Variable, Frame Size 70 ms, and Overlap Ratio 0.6)

The non-speech percentage of correct decision starts at a high value in the human pitch range. As the pitch range increases, the percentage of correct decision of the non-speech decreases. The non-speech percentage of correct decision reaches zero when the value of the upper pitch limit reaches 8000 Hz (half the sampling rate of 16 KHz). The behavior of the non-speech percentage of correct decision is shown in figure 5.2.2.

If the value of the pitch range is chosen to be in the human pitch range, the value of the speech pitch ratio will be concentrated in the range of $[0.5 - 1]$, as shown in figure 5.2.3; also, the value of the non-speech pitch ratio will be concentrated in the range of $[0 - 0.5]$, as shown in figure 5.2.4.

The pitch range drastically affects the non-speech percentage of correct decision if the value exceeds the human pitch range. In figure 5.2.6, the pitch range is $[70 - 2000 \text{ Hz}]$ and as is shown, the non-speech audio segments are spread over all the pitch ratio range. This spreading will have an impact on the pitch ratio threshold which will affect the percentage of correct decision. For the speech cases in the pitch range $[70 - 2000 \text{ Hz}]$, the speech pitch ratio is concentrated in the range of $[0.8 - 1]$. This value is good for the speech classification but it is bad for the speech/non-speech discrimination.

Figure 5.2.3 Histogram of the Number of Speech Audio Segments vs. Pitch Ratio (Frame Size 70 ms, Overlap Ratio 0.6, and Pitch Range 70 – 280 Hz).

Figure 5.2.4 Histogram of the Number of Non-Speech Audio Segments vs. Pitch Ratio (Frame Size 70 ms, Overlap Ratio 0.6, and Pitch Range 70 – 280 Hz).

Figure 5.2.5 Histogram of the Number of Speech Audio Segments vs. Pitch Ratio (Frame Size 70 ms, Overlap Ratio 0.6, and Pitch Range 70 – 2000 Hz).

Figure 5.2.6 Histogram of the Number of Non-Speech Audio Segments vs. Pitch Ratio (Frame Size 70 ms, Overlap Ratio 0.6 and Pitch Range 70 – 2000 Hz).

The pitch ratio algorithm using the human pitch range achieves a higher performance and lowew complexity. The pitch ratio algorithm achieves a high percentage of correct decision in both the speech and the non-speech under the same conditions.

Table 5.2.1 Comparison Between Speech and Non-Speech Correct Decision Using Different Algorithms

Algorithm	Speech	Non-speech
Whole audio segment	88.8 %	93%
Without pitch ratio (frame size 70 ms, overlap 0.6)	100%	52.9%
LPC algorithm	75%	70%
Pitch ratio without human pitch range (frame size 70 ms, overlap 0.6)	99.8%	2.7%
Pitch ratio with human pitch range (frame size 70 ms, overlap 0.6, pitch range 70 - 280)	97%	98.1%

A comparison between different algorithms is shown in table 5.2.1. This table shows the performance of the pitch ratio algorithm with the human pitch range. The LPC

algorithm for speech/non-speech discrimination is the algorithm proposed in [5]. As is shown in the previous table, the pitch ratio algorithm performance without the human pitch range is very poor. To enhance the performance, the human pitch range is introduced.

Different audio types have been used in testing the algorithms. Table 5.2.2 and Figure 5.2.7 show the percentage of correct decision for different audio types. The proposed algorithm shows a high performance in all types. The LPC algorithm shows a high performance for the female audio type and very poor performance for the male and cocktail noise audio types.

Table 5.2.2 Comparison Between Different Audio Types Correct Decision Using Different Algorithms

Audio type	Whole audio segment	LPC algorithm	Pitch ratio with human pitch range
Male	86.2%	51%	96.8%
Female	90.2%	98.9%	97.1%
Chair	80%	60%	100%
Door	80%	66.5%	90%

Drawer	90%	70%	100%
Fan noise	100%	76.6%	100%
Footsteps	90.2%	82.7%	95.1%
Breaking Glass	97.5%	55%	100%
Rain	100%	82%	100%
Cocktail Noise	95%	35%	100%
Thunder	97.5%	97.5%	100%
Traffic	100%	50%	100%
Water	95%	85%	100%
Wind	100%	45%	95%

Figure 5.2.7 Histogram for the Different Algorithms Correct Decision

5.3 Discussion

The human pitch range algorithm is introduced. The human pitch range shows better performance for the speech/non-speech discrimination when it is combined with the pitch ratio algorithm.

A comparison between previous algorithms and the proposed algorithm is shown. It has been demonstrated that the proposed algorithm has a better performance. The audio

library used is oriented toward security and teleconferencing applications. We conclude that the proposed algorithm is suitable for these applications.

Chapter Six

Audio Classification

Audio classification results using neural networks are introduced. In the following sections, audio classification is developed, and simulations and results are presented. A discussion and concluding remarks are given.

6.1 Audio Type Classification

In the security field, it is helpful if the audio type is known. Our goal in this chapter is to classify the audio type for the non-speech, such as breaking glass and footsteps. This is shown in figure 6.1.1. The results will be simulated using the neural networks. A Feature vector will be created depending on the literature review of the audio retrieval.

Figure 6.1.1 Non-Speech Audio Type Classification

6.2 Feature Vector

The feature vector is created using five main audio features. The features are created for a frame length of 40 milliseconds. For each frame, the power spectrum (PS) [54], mel-filtered cepstral coefficients (MFCCs) [58], magnitude distribution (MD)[59], discrete Fourier transform (DFT), and the pitch value are extracted.

For the PS, the input frame is analyzed using a 7-channel filter bank. The square of the output of each channel is averaged over the number of samples of the same output. The histogram is then created using 3 bins for each frequency band in the filter bank.

The MFCC vector is computed by applying the triangle filters with the mel-spaced set to the DFT, then the discrete cosine transform is applied [36], [58].

The MD vector is calculated using the time domain. The histogram with 30 bins is created as a vector for the MD. Finally, the DFT vector and the pitch value for each frame is calculated.

For the vectors PS, MFCC, MD and DFT, the following steps are taken:

- 1- The difference between the successive elements in the same vector is calculated, which is defined as the first difference (Δ)[60].
- 2- The mean and the standard deviation of each vector of the following vectors are calculated (PS, MFCC, MD, DFT, Δ PS, Δ MFCC, Δ MD, Δ DFT)
- 3- The pitch value is calculated for each frame.
- 4- The 16 feature values from step 2 are appended to the pitch value from step 3 to form a 17 element feature vector.

6.3 Neural Network

The neural network used in this part of the thesis is the Feed Forward (FF) Back Propagation (BP) neural network (NN). Three layers are used [61] in designing the network; the hidden layer is constructed using a number of hidden nodes and the output

layer has the number of nodes equal to the number of the audio types. The neural network is designed using the Matlab™ toolbox. The neural network is trained to get an output decision of 1 at the output node corresponding to the audio type and get decision 0 at the other nodes.

6.4 Simulations and Results

The feature vector for each frame is created. In the first part of the simulation, we will use the neural network to classify two audio types (Glass Breaking and Drawer sounds). The second part of the simulation will classify four audio types (Fan, Wind, Glass Breaking and Door sounds).

Table 6.4.1 Training and Testing Data for Audio Classification

Audio type	Training	Testing
Door Sounds	790	632
Drawer Sounds	790	632
Fan	790	395
Glass Breaking	790	632
Wind	790	395

The training set is created using frames of 40 milliseconds duration. The feature vector for each frame is calculated. The training and testing frames number are shown in table 6.4.1.

The training and the testing data are totally uncorrelated (the training data set is different from the testing data set).

In the first part of the simulation the neural network is trained to discriminate between the breaking glass and the drawer sounds. The simulation results are shown in table 6.4.2.

Table 6.4.2 Simulation Results for Two Classes

Event	Glass Breaking	Drawer Sound	Correct Decision
Glass Breaking	599	33	94.7%
Drawer sound	0	632	100%

In the second part of the simulation, the neural network is trained to discriminate between four audio types (Fan, Wind, Glass Breaking and Door sounds). The simulation results are shown in table 6.4.3, table 6.4.4, and table 6.4.5 for 100, 50, 25, 10, 5 and 2 hidden nodes respectively.

Table 6.4.3 Simulation Results for Four Classes Using 100 Hidden Nodes

Classifier Decision Classifier	Door Sounds	Fan	Glass Breaking	Wind	Correct Decision
Door Sounds	539	13	66	14	85.9%
Fan	3	388	3	1	98.2%
Glass Breaking	14	36	561	21	88.7%
Wind	8	0	7	380	96.2%

Table 6.4.4 Simulation Results for Four Classes Using 50 Hidden Nodes

Classifier Decision Classifier	Door Sounds	Fan	Glass Breaking	Wind	Correct Decision
Door Sounds	526	44	53	9	83.2%
Fan	0	393	1	1	99.5%
Glass Breaking	21	73	523	6	84.2%
Wind	14	3	3	375	95%

Table 6.4.5 Simulation Results for Four Classes Using 25 Hidden Nodes

Classifier Decision Classifier	Door Sounds	Fan	Glass Breaking	Wind	Correct Decision
Door Sounds	516	46	52	18	81.6%
Fan	2	390	2	1	98.7%
Glass Breaking	34	49	542	7	85.7%
Wind	13	4	4	374	94.7%

Table 6.4.6 Simulation Results for Four Classes Using 10 Hidden Nodes

Classifier Decision Classifier	Door Sounds	Fan	Glass Breaking	Wind	Correct Decision
Door Sounds	458	59	70	18	76.7%
Fan	1	392	1	1	99%
Glass Breaking	53	75	500	4	79.1%
Wind	15	3	0	377	95.4%

Table 6.4.7 Simulation Results for Four Classes Using 5 Hidden Nodes

Classifier Decision Classifier	Door Sounds	Fan	Glass Breaking	Wind	Correct Decision
Door Sounds	478	57	67	30	75.6%
Fan	3	390	1	1	98.7%
Glass Breaking	64	71	495	2	78.3%
Wind	12	0	1	382	96.7%

Table 6.4.8 Simulation Results for Four Classes Using 2 Hidden Nodes

Classifier Decision Classifier	Door Sounds	Fan	Glass Breaking	Wind	Correct Decision
Door Sounds	397	108	83	44	62.8%
Fan	7	380	4	4	96.2%
Glass Breaking	144	124	358	6	56.6%
Wind	7	4	6	378	95.6%

6.5 Discussion

Audio classification for different non-speech audio types is introduced. The neural network is used. The simulation shows promising results for non-speech audio classification.

The work done in this chapter depends on the literature review of audio retrieval algorithms. The security requirements will be achieved by combining the pitch ratio algorithm and the audio classification results. A hierarchical system using the pitch ratio and the audio classification will be built in future work.

Chapter Seven

Conclusions and Future Work

This thesis has made contributions by offering a new feature extraction for speech/non-speech classification to be used in security and teleconferencing applications. The proposed algorithm achieved better performance with respect to the previous algorithms. It has also contributed an audio type classification for non-speech audio using neural networks. Some supporting algorithms were developed.

The proposed method was based on splitting the audio segment into small frames and detecting the presence of pitch in each one of them. The ratio of frames with pitch detected to the total number of frames was defined as the pitch ratio and was used as the main feature to classify speech and non-speech segments. The performance of the proposed method was evaluated using a library of audio segments containing female and

male speech, and non-speech segments such as computer fan noise, cocktail noise, footsteps, and traffic noise.

Two major contributions were proposed for the speech/non-speech discrimination. The pitch ratio algorithm was proposed to give robust decision making for each class. Limiting the search to the human pitch range was also proposed to enhance the performance of the pitch ratio algorithm. Also, non-speech audio type classification using the neural networks was proposed. This approach shows promising results.

This thesis proposed the speech/non-speech algorithm for the security area and evaluated its performance. Also, a comparison between the proposed algorithm and the previous work was shown in this thesis.

It is shown that the proposed algorithm can achieve a percentage of correct decision of 97% for the speech and 98% for non-speech segments, 0.5-seconds long.

This result can also be extended to the audio-visual area. Controlling the motion of the camera by limiting the camera's sensitivity to a speech source thus avoids the camera interruption that otherwise would have been caused by non-speech such as footsteps and door.

Extending the algorithm to cover the music area and including music as a non-speech class is a challenge for future research, especially as music has pitch in the human pitch

range. Classification of music as a non-speech can be done in the future using the energy sub-band where the energy of the music is distributed in a wider frequency range than speech.

The audio classification results show a promising algorithm for audio type classification. In future work, an investigation of the feature vector and the classification method will be done. The classification methods that can be used in audio classification are the Gaussian Mixed Model (GMM), Hidden Markov Model (HMM) and the histogram.

Appendix A: Acronyms

ASC	Audio Signal Classification
ASR	Automatic Speech Recognition
BP	Back Propagation
FF	Feed Forward
GMM	Gaussian Mixed Model
HOS	High Order Statistics
SNR	Signal to Noise Ratio
VAD	Voice Activity Detector
TASI	Time Assignment Speech Interpolation

RMSE Root Mean Square Energy

LPF Low Pass Filter

NN Neural Network

HMM Hidden Markov Model

HPF High Pass Filter

BPF Band Pass Filter

References

- [1] G. Lu, and T. Hankinson, "A technique towards automatic audio classification and retrieval," in *Proc. Int. Conf. Signal Process.*, vol. 2 , pp. 1142–1145, Oct. 1998.
- [2] L. Janer, J.J. Bonet, and E. Lleida-Solano, "Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms," in *Proc. Int. Conf. Spoken Language*, vol. 2, pp.1209–1212, Oct. 1996.
- [3] D.A. Krubsack, and R.J Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech," *IEEE Trans. Signal Process.*, vol. 39, issue 2, pp. 319–329, Feb. 1991.
- [4] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comprehensive performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, pp. 399–417, Oct. 1976.

-
- [5] Qiyue Zou, Xiaoxin Zou, Ming Zhang, and Zhiping Lin, "A robust speech detection algorithm in a microphone array teleconferencing system," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, pp. 3025–3028, May 2001.
 - [6] Jie Huang, N Ohnishi, and N.Sugie, "Spatial localization of sound sources: azimuth and elevation estimation," in *Proc. Instrumentation and Measurement Technology Conf.*, vol. 1, pp. 330–333, May 1998.
 - [7] D. Wojtaszek, and R. Laganieri, "Tracking and recognizing people in colour using the earth mover's distance," in *Proc. IEEE Int. Workshop on Haptic Virtual Environments and Their Applications*, pp. 91–96, Nov. 2002.
 - [8] C.C. Chibelushi, F. Deravi, and J.S. Mason, "Audio-visual person recognition: an evaluation of data fusion strategies," in *Proc. Security and Detection, European Conf.*, pp. 26–30, Apr. 1997.
 - [9] M.S Brandstein, J.E. Adcock, and H.F. Silverman, "A closed-form method for finding source locations from microphone-array time-decay estimates," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, pp. 3019–3022, May 1995.
 - [10] Jian-Feng Chen, Lijun Jiang, and Wee Ser, "A framework for speaker tracking using microphone array and camera," in *Proc. Int. Conf. Signal Process.*, vol. 2, pp. 1384–1387, Aug. 2000.
-

-
- [11] D.E Sturim, M.S Brandstein, and H.F Silverman, "Tracking multiple talkers using microphone-array measurements," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol.1, pp. 21–24, Apr. 1997.
 - [12] Ming Zhang, and M.H. Er, "Tracking direction of speaker for microphone array in the far field," in *Proc. Int. Conf. Information Engineering*, pp. 541–544, Jul. 1995.
 - [13] G. Lu, and T. Hankinson, "An investigation of Automatic Audio Classification and segmentation," in *Proc. Int. Conf. Signal Process.*, vol. 2, pp. 776–781, Oct. 2000.
 - [14] A. Bendiksen, and K. Steiglitz, "Neural networks for voiced/unvoiced speech classification," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol.1, pp. 521–524, Apr. 1990.
 - [15] R.P. Cohn, "Robust voiced/unvoiced speech classification using a neural net," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol.1, pp. 437–440, Apr. 1991.
 - [16] D.K. Emge, T. Adali, and M. Kemal Sonmez, "Least relative entropy for voiced/unvoiced speech classification," in *Proc. Int. Joint Conf. Neural Networks*, vol. 5, pp. 2976–2980, Jul. 1999.
-

-
- [17] T. Ghiselli-Crippa, and A. El-Jaroudi, "A fast neural net training algorithm and its application to voiced-unvoiced-silence classification of speech," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol.1, pp.441–444, Apr. 1991.
- [18] T. Ghiselli-Crippa, and A. El-Jaroudi, "Voiced-unvoiced-silence classification of speech using neural nets," in *Proc. Int. Joint Conf. Neural Networks*, vol. 2, pp. 851–856, Jul. 1991.
- [19] S.J. Kia, and G.G. Coghill, "A mapping neural network and its application to voiced-unvoiced-silence classification," in *Proc. Int. Conf. Artificial Neural Networks and Expert Systems, First New Zealand International, Two-Stream Conference*, pp. 104–108, 1993.
- [20] M.F BenZeghiba, and H Boulard, "Hybrid HMM/ANN and GMM combination for user-customized password speaker verification," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, pp. 255–288, Apr. 2003.
- [21] G. Singh, A. Panda, S. Bhattacharyya, and T. Srikanthan, "Vector quantization techniques for GMM based speaker verification," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, pp. 65–68, Apr. 2003.
- [22] J. Kleban, and Y. Gong, "HMM adaptation and microphone array processing for distant speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, pp. 1411–1414, Jun. 2003.
-

-
- [23] A.V.Nefian, Liang Luhong, Pi. Xiaobo, Xiaoxiang Liu, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, pp. 2013–2016, 2002.
- [24] Oh-Wook-Kwon, and Te-Won Lee, "Optimizing speech/non-speech classifier design using adaboost," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. I_436–I_439, Apr. 2003.
- [25] A. Alkulaibi, J.J. Soraghan, and T.S. Durrani, "Fast HOS based simultaneous voiced/unvoiced detection and pitch estimation," in *Proc. Signal Process. Workshop, Statistical Signal and Array Process.*, pp. 194–197, Jun. 1996.
- [26] <http://www.bellsystemmemorial.com/belllabs.html>.
- [27] K. El-Maleh, and P. Kabal, "Comparison of voice activity detection algorithms for wireless personal communications systems," in *Proc. Canadian Conf. Electrical and Comp. Eng., Engineering Innovation: Voyage of Discovery*, vol. 2, pp.470–473, May 1997.
- [28] B. Clarkson, and A. Pentland, "Extracting context from environmental audio," in *Proc. Int. Sym. Wearable Computers*, pp. 154–155, Oct. 1998.
- [29] Kyoung-Ho Woo, Tae-Young Yang, Kun-Jung Park, and Chungyong Lee "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, issue 2, pp.180–181, Jan. 2000.
-

-
- [30] Zenton Goh, Kah-Chye, Tan Tan, and B.T.G., "Speech enhancement based on a voiced-unvoiced speech model," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 401–404, May 1998.
- [31] J. Proakis, "*Digital Signal Processing*", McGraw Hill, 1996.
- [32] D.K. Freeman, C.B. Southcott, and I. Boyd, "A voice activity detector for the Pan-European digital cellular mobile telephone service," *IEE Colloquium, Digitized Speech Communication via Mobile Radio*, pp.6/1–6/5, Dec. 1988.
- [33] K. Srinivasan, and A. Gersho, "Voice activity detection for cellular networks," in *Proc. Speech Coding for Telecom. Workshop*, pp.85–86, Oct. 1993.
- [34] J. Stegmann, and G. Schroder, "Robust voice-activity detection based on the wavelet transform," in *Proc. Speech Coding for Telecom. Workshop*, pp.99–100, Sep. 1997.
- [35] N. Erdol, and R. Schultz, "Voice activity detection over multiresolution subspaces," in *Proc. Sensor Array and Multichannel Signal Process. Workshop*, pp.217–220, Mar 2000.
- [36] John R. Deller, John Hansen, and John Proakis, "*Discrete-time Processing of Speech Signals*," New York, Institute of Electrical and Electronics Engineers., 2000.
-

-
- [37] R. Ahn, and W.H. Holmes, "Harmonic-plus-noise decomposition and its application in voiced/unvoiced classification," in *Proc. Conf. Computer and Telecommunication., Control and Power Engineering, Region 10*, vol. 2 ,pp. 587–590, Dec. 1997.
- [38] R. Ahn, and W.H. Holmes, "Voiced/unvoiced/silence classification of speech using 2-stage neural networks with delayed decision input," in *Proc. Int. Sym. Signal Process. and Its Applications*, vol. 1, pp. 389–390, Aug. 1996.
- [39] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech and Audio Process.*, pp.217–231, Mar. 2001.
- [40] H. Kobatake, "Optimization of voiced/unvoiced decisions in nonstationary noise environments," *IEEE Trans., Acoust., Speech Signal Process.*, vol. 35, pp.9–18, Jan. 1987.
- [41] S.G. Tanyer, and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech and Audio Process.*, vol.8, issue 4, pp.478–482, Jul. 2000.
- [42] S.G. Tanyer, and H. Ozer, "Voice activity detection in nonstationary Gaussian noise," in *Proc. Int. Conf. Signal Process.*, vol. 2 , pp.1620–1623, Oct. 1998.
- [43] R. Tucker, "Voice activity detection using a periodicity measure," in *Proc. IEE Communications, Speech and Vision*, vol. 139, issue 4, pp. 377–380, Aug. 1992.
-

-
- [44] Thimas F. Quatieri, “*Discrete-time speech signal processing principles and practice*,” Prentice Hall signal processing series, 2002.
 - [45] B. S. Atal, and L. R. Rabiner, “A pattern recognition approach to speech to voiced-unvoiced-silence classification with application to speech recognition,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, pp.201–212, Jun. 1976.
 - [46] L. J. Siegel, “A procedure for using pattern classification techniques to obtain a voiced/unvoiced classifier,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 83–88, Feb. 1979.
 - [47] Sergios Theodoridis, and Konstantinos Koutoumbas, “*Pattern Recognition*,” Academic Press 1999.
 - [48] F. Basbug, S. Nandkumar, and K. Swaminathan, “Robust voice activity detection for DTX operation of speech coders,” in *Proc. Workshop Speech Coding*, pp. 58–60, Jun. 1999.
 - [49] F. Beritelli, S. Casale, and A. Cavallaero, “A robust voice activity detector for wireless communications using soft computing,” *IEEE Journal, Selected Areas in Communications*, vol. 16, issue 9 , pp. 1818–1829, Dec. 1998.
 - [50] D.K. Freeman, G. Cosier, C.B. Southcott, and I. Boyd, “The voice activity detector for the Pan-European digital cellular mobile telephone service,” in
-

- Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol.1, pp. 369–372, May 1989.
- [51] D. O'Shaughnessy, H. Tolba, "Towards a robust/fast continuous speech recognition system using a voiced-unvoiced decision," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 413–416, Mar. 1999.
- [52] A. Benyassine, E. Shlomot, Su. Huan-Yu, and E. Yuen, "A robust low complexity voice activity detection algorithm for speech communication systems," in *Proc. Speech Coding for Telecommunication Workshop*, pp.97–98, Sep. 1997.
- [53] J.A. Haigh, and J.S. Mason, "Robust voice activity detection using cepstral features," in *Proc. Conf. Computer, Communication, Control and Power Engineering, Region 10*, pp. 321–324, Oct. 1993.
- [54] Kunio Kashino, Gavin Smith, and Hiroshi Murase, "Quick Audio Retrieval Based on Histogram Feature Sequences," in *Journal of Acoust. Society of Japan (E)*, vol.21, pp.217–219, 2000.
- [55] T. Zhong, and C. C. Jay Kuo, "Hierarchical classification of audio data for archiving and retrieving," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 6 , pp. 3001–3004, Mar. 1999.
-

-
- [56] Ahmad R. Abu-El-Quran, and Rafik A. Goubran, "Pitch-based feature extraction for audio classification," in *Proc. Haptic, Audio, Visual Environments and their Applications*, Sept. 2003.
- [57] Y. Grenier, and S. Affes, "Microphone array response to speaker movements," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 247–250, Apr. 1997.
- [58] J. Foote, "A similarity measure for automatic audio classification," In *Proc. AAAI Sym. on Intelligent Integration and Use of Text, Image, Video*, 1997.
- [59] S. Gazor, and W. Zhang, "Speech probability distribution," in *Letters Signal Process.*, vol.10, pp. 204–207, Jul. 2003.
- [60] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search and retrieval of audio," *IEEE Multimedia Magazine*, vol. 3, pp. 27–36, July 1996.
- [61] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–366, 1989.
-