AUTOMATIC INFANT CRY ANALYSIS AND RECOGNITION

by

QIAOBERG XIE

B. A. Sc. (Computer Application), China Textile University, 1982
M. A. Sc. (Computer Science and Application), China Textile University, 1985

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

in
THE FACULTY OF GRADUATE STUDIES
THE DEPARTMENT OF ELECTRICAL ENGINEERING

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

June 1993

© Qiaobing Xie, 1993
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature)

Department of Electrical Engineering

The University of British Columbia
Vancouver, Canada

Date Sept. 20, 1993
Abstract

This dissertation is a report of my investigation on introducing modern speech processing/recognition techniques to the field of infant cry research, and on developing efficient and effective methodologies for automatic assessment of the physical/emotional situation of infants using the information derived from the cry signals.

I first identify some problems facing present infant cry research, especially those obstructing the practical applications of the results generated from basic research. By demonstrating the similarities between infant cry generation and adult speech generation, I establish the theoretical foundation for the development of my new automatic cry processing/analysis techniques.

In particular, I develop the new concept of cry phonemes as an effective method for representing cry signals for automatic cry analysis. Based on the cry phonemes, I further define a composite parameter, the H-value, which can be calculated from the cry signal, and is found to be a reliable indicator of the distress level of the infant.

Using these new concepts, I design two automatic infant cry analysis systems. One system is based on my newly developed nonparametric VQ-kernel classifier, and the other system is based on the Hidden Markov Model technique. Each of these systems estimates the H-value from the cry signal automatically. This, in turn, is utilized in the automatic assessment of the infant’s distress level.

The performance of these two systems was evaluated with cries uttered by 36 infants. I found that both systems give assessments of infants’ distress levels consistent with the perceptions of experienced parents who listened to the recording of the same cries. This demonstrates the effectiveness of my newly developed techniques.

In addition, the methodologies developed in this research can be easily generalized and applied to other problems of normal and abnormal infant cry analysis.
# Table of Contents

Abstract ii

Table of Contents iii

List of Tables viii

List of Figures ix

Acknowledgments xii

1 INTRODUCTION 1

1.1 The Overall Research Plan .............................................. 2

2 INFANT CRY STUDY: HISTORY, ADVANCES AND CHALLENGES 5

2.1 History ............................................................................. 5

2.2 Major Directions in the Study of Infant Cries ........................ 7

2.3 Techniques and Instruments .............................................. 10

2.3.1 Auditory Analysis ...................................................... 11

2.3.2 Non-acoustical Analysis .............................................. 11

2.3.3 Time Domain Acoustic Analysis .................................. 12

2.3.4 Sound Spectrographic Analysis .................................... 12

2.3.5 Computer-aided Analysis ............................................ 15

2.4 Remaining Problems ..................................................... 16

2.4.1 Controversy on How Much Information Infant Cry Conveys 16

2.4.2 Problems Facing Modern Computerized Infant Cry Analysis 18

2.5 Summary ........................................................................... 21
3 AN OVERVIEW OF APPLICABLE VOICE AND SPEECH PROCESSING

TECHNIQUES

3.1 Human Vocalization Organs and Speech Production .......................... 23

3.2 The Digital Model of Speech Production and Linear Prediction Coding (LPC) of
Speech Signals .............................................................................. 27

3.2.1 Time-varying Filter Model of Speech Generation .......................... 27

3.2.2 Linear Prediction of Speech Signal ............................................. 28

3.3 Pattern Recognition ..................................................................... 30

3.4 Automatic Speech/Speaker Recognition .......................................... 32

3.4.1 Automatic Speech Recognition .................................................. 32

3.4.2 Speaker Recognition ................................................................. 33

3.5 Vector Quantization ................................................................. 34

3.5.1 Some Details of Vector Quantization ......................................... 35

3.6 Hidden Markov Models ............................................................... 36

3.6.1 Description of HMMs ............................................................... 36

3.6.2 Three Problems ..................................................................... 37

3.6.3 Speech Recognition with HMMs .............................................. 38

4 FEASIBILITY STUDIES ON APPLYING SPEECH PROCESSING

TECHNIQUES TO CRY ANALYSIS

4.1 Investigating the Physiological Similarities between Infant Crying and Adult
Speech ......................................................................................... 41

4.1.1 A Survey: Physiology of Infant Crying and a Physioacoustic Model of Cry
Production .................................................................................. 41
4.2 Formant Frequency Estimate Using an LPC Model ........................................ 47

4.2.1 Theoretical Analysis .................................................................................... 47

4.2.2 Experiment .................................................................................................. 50

4.2.3 Results ........................................................................................................ 51

4.3 Some Technical Considerations of Automatic Cry Analysis ...................... 53

4.3.1 Training of the Automatic Cry Analyzer .................................................... 54

4.3.2 Feature Selection ....................................................................................... 55

5 THE H-VALUE: A MEASURE FOR THE AUTOMATIC ASSESSMENT OF
NORMAL INFANT DISTRESS-LEVELS FROM THE CRY SIGNAL ............................. 58

5.1 The Level of Distress (LOD) Ratings of Normal Infant Cries ...................... 59

5.2 Definitions of "Cry Phonemes" for Normal Infant Cries ............................... 61

5.3 The Relationship between "Cry Phonemes" and the Parents' LOD Ratings .... 65

5.3.1 Experiments .............................................................................................. 66

5.3.2 Results ....................................................................................................... 68

5.4 The Introduction of the H-value and Its Application in Quantifying LOD
Assessments ....................................................................................................... 74

5.5 Discussions and Remarks ............................................................................. 75

6 CRY ANALYZER I: A NONPARAMETRIC STATISTICAL
CLASSIFIER-BASED METHOD .............................................................................. 77

6.1 A Survey of Traditional Statistical Classification Methods and Their Limitations 77

6.2 The Development of the VQ-based Nonparametric Classification Approach .. 80

6.2.1 Development of the Algorithms ................................................................. 83

6.2.2 Experiments of the Application of Our VQ-based Classifiers ................. 88

6.2.3 Summary .................................................................................................... 94
6.3 Automatic Cry Analyzer Based on the VQ-kernel Methods

6.3.1 System Design

6.3.2 Determination of VQ-kernel Classifier Parameters

6.3.3 Training of VQ-kernel Classifier

6.4 Results and Discussions

7 CRY ANALYZER II: A HIDDEN MARKOV MODEL (HMM) -BASED METHOD

7.1 System Design

7.1.1 Data Preprocessing and Feature Extraction

7.1.2 Vector Quantization of the Feature Vectors

7.1.3 Segmentation of the Signal into Recognition Units

7.1.4 Configuration of the HMM-based Analyzer and the Calculation of the H-value

7.2 Training of the HMMs

7.3 Experiment Results and Discussion

8 CONCLUSIONS

8.1 Accomplishments and Contributions

8.2 Some Topics for Future Study on Automatic Infant Cry Analysis

8.2.1 Refining the Techniques and Methods We Have Developed

8.2.2 Extending Our New Methodology to Other Infant Cry Research Topics

8.2.3 Developing New Products Based on Our Computerized Infant Cry Analysis Techniques

Appendix A Linear Predictive Coding (LPC) Algorithms

Appendix B Linde-Buzo-Gray’s Algorithm of Vector Quantizer Design
List of Tables

5.1 Characteristics of the 36 infants. .............................................. 66

6.1 CPU time used for finding the reduced set in the speech data experiment. (* including CPU time used for classification) ........................................ 93

7.1 Distribution of the micro-segments in our training set. ..................... 116
### List of Figures

3.1 Schematic diagram of the human vocal system. (after Flanagan et al.) 24

3.2 A typical glottal pulse train. 25

3.3 Spectra in speech generation — (a) idealized spectrum of glottal pulse train, (b) frequency response of vocal tract, where the peaks correspond to formants, and (c) spectrum of the resultant speech. 26

3.4 A two-stage speech generation model. 26

3.5 The digital time-varying filter model of speech production. 27

3.6 A pattern recognition system. 30

3.7 A 3-state HMM with 4 output symbols. 37

4.1 A simplified view of the cry production model. 44

4.2 (A) Spectrum of an idealized periodic source; (B) spectrum of an turbulence source; (C) vocal tract transfer function; (D) spectrum of an idealized radiation characteristic; and (E) spectrum of an idealized output. (after Golub and Corwin) 45

4.3 LPC formant estimate experiment of infant cries. 50

4.4 (a) Estimated formant frequencies by using the LPC-based method on a pain cry. Estimations marked with circles indicate the presence of the hyperphonation. (b) Spectrogram of the same cry. In both graphs, the trajectories of the first formant, second formant, and third formant are labeled by A, B, and C, respectively. 52

5.1 The time-frequency patterns of the 10 “cry phonemes” shown in the computer-derived spectrograms. They are: (1) trailing, (2) flat, (3) falling, (4) double harmonic break, (5) dysphonation, (6) rising, (7) hyperphonation, (8) inhalation, (9) vibration, and (10) weak vibration, respectively. 64
5.2 An example of labeling a pain elicited cry into the “cry phonemes.” Numbers on the top indicate the mode or type of the “cry phoneme” of each segment of the cry signal. The curve below indicates the energy of the signal. 65

5.3 The relationship of parents’ LOD ratings to the percentage occurrences of different “cry phonemes.” 73

5.4 Relationship between the H-value and the parents’ LOD ratings. 75

6.1 The classification error rates of our VQ-based classifiers and other traditional reduced data classifiers. 90

6.2 The classification error rates of our VQ-based classifiers, the traditional reduced data NN classifiers, and Fukunaga’s reduced Parzen classifier for the speech data. 92

6.3 Automatic infant cry analysis system based on the VQ-kernel classifier. 96

6.4 The generation of the reduced training sets for the VQ-kernel classifier. 98

6.5 Determination of the decision threshold, \( t \). 101

6.6 The estimated classification error rate vs. the decision threshold \( t \). 101

6.7 The H-values which are estimated by our VQ-kernel classifier-based cry analyzer versus the actual measurements. The estimation errors appear as the vertical distances from the circles to the diagonal dashed line. 104

6.8 Relationship between the H-values estimated by the VQ-kernel classifier-based cry analyzer and the parents’ LOD ratings. 105

7.1 The block diagram of our HMM based cry distress level analysis system. 108

7.2 Vector quantization. 109

7.3 The segmentation of a cry signal. 111

7.4 Topology of the hidden Markov models for micro-segment identification. 113
7.5 The structure of the H-HMM (the same topology is used for the E-HMM) ..... 114
7.6 The HMMs-based classifier. ....................................................... 115
7.7 The computer estimations versus the actual measurements of the H-values of the 58 
cries in the testing set. The estimation errors appear as the vertical distances from 
the circles to the diagonal dashed line. ........................................ 118
7.8 Computer estimated H-value versus parents’ LOD rating. .................... 119
Acknowledgments

First, my greatest debt of gratitude goes to my supervisors, Professor Rabab K. Ward and Professor Charles A. Laszlo, who have introduced and guided me into this exciting field of research, and provided sound advice and invaluable critical feedback at every stage of this project. Without their support, this thesis could not have been written. I would also thank Dr. M. J. Yedlin, who served as an examining committee member in my proposal defence, departmental defence, and final oral examination, for his advice and suggestions.

I am very grateful to Dr. Ruth V. E. Grunau of BC Children’s Hospital for her allowing me to use the infant cry data she collected in her previous research.

Thank you also goes to the parents who participated in our experiment and provided the important LOD ratings.

I also wish to acknowledge Pingnan Shi and Hui Peng for their friendship and encouragement over these four long years.

Finally, I would like to acknowledge that this project is partly supported by NSERC grants to Dr. Rabab K. Ward and Dr. Charles A. Laszlo.

The collection of the infant cry data which Dr. Grunau supplied us was supported by a grant to Dr. K. D. Craig from NSERC.

I dedicate this work to Mrs. Wei-Lie Zeng and Mr. Shou Jiang Xie, my mother and father.
Chapter 1

INTRODUCTION

This work presents our research in applying modern digital computer-based speech processing/recognition techniques to the automatic analysis of infant vocalizations.

Crying and other forms of non-verbal vocalizations play a very important role in the communication of infants with their caretakers. It has also been speculated since the last century that such vocalizations also contain information about the infants' physical and emotional situations, and that the proper interpretation of that information can lead to effective and efficient methods of monitoring the infants’ well-being. In turn, this could lead to the diagnosis of some diseases and abnormalities.

During the past three decades significant progress has been made in the understanding of the infant cry generation mechanism and various relationships between infants’ physical situations and different cry attributes. Nevertheless, many important questions still remain unanswered, and no widely accepted applications of infant cry analysis are found in practice. It is particularly striking that many years of effort to convert the findings and results of this fundamental research into practical applications has brought few results. This, in our opinion, is due to the lack of utilizing highly effective and efficient methods of analysis in infant cry research. We believe that more engineering-oriented methodologies should be introduced into infant cry research so that the findings and results of past cry research can be transferred into a form suitable to practical applications.
In recent years, automatic speech processing/recognition technology has gone through rapid development. Many new techniques, such as vector quantization and hidden Markov modeling, have been introduced and successfully applied to automatic speech processing/recognition. If we consider the infant cry signals as signals closely related to speech, it is an excellent opportunity to investigate the introduction of modern speech processing/recognition techniques into infant cry research. As a result of introducing speech processing techniques into infant cry research, we expect to obtain new analysis methods which are effective and efficient from the viewpoint of engineering-oriented application. Our plan of research is detailed in the following.

1.1 The Overall Research Plan

We plan to achieve the following objectives in our investigation.

1. To identify the problems facing infant cry application research — We will review the various theories, methodologies, techniques, and instrumentation developed for infant cry research in the past three decades. We will also examine the large amount of findings and results on infant crying in the literature. In particular, we will try to identify the most difficult problems and the obstacles that hinder the improvement of the effectiveness and efficiency of the techniques employed in the study of infant cries. This will allow us to evaluate the potential of modern signal and speech processing methods in infant cry research and clinical applications.

2. To investigate the theoretical feasibility of applying speech processing/recognition technology to infant cry research — It is important to note that the theory and methodology of modern speech processing is heavily based on specific physioacoustic models of speech
Chapter 1: INTRODUCTION

generation. The success of our introduction of modern speech processing technology into infant cry research is, therefore, fundamentally related on the physiologic and acoustic similarities between the mechanisms of speech generation and cry generation. This phase of our work is designed to compare speech and cry generation mechanism, and to determine the applicability of existing physioacoustic models to the cry sound.

3. To address the special problem of assessing normal infant distress levels from cry signals — Provided that we can show that speech processing technology and methodology is applicable to cry sounds, we will then concentrate our investigation on one special problem. This problem is the automatic assessment of the physical and/or emotional situation of normal (clinically healthy) infants from the cry sounds emitted by them. In particular, we plan to develop effective and efficient methods to characterize and represent normal infant cry sounds in a way suitable to automated analysis. Based on this signal representation technique, we will then try to find parameters to measure and quantify the distress level of normal infants. These parameters should be able to effectively indicate the infants’ physical/emotional situations. Equally important is that these parameters can be reliably estimated from the cry sound by means of modern signal processing/recognition technique.

4. To develop automatic infant cry analysis systems based on our previous investigations and evaluate their performance on computer — In this stage of our research, we will work on the system design of the automatic infant cry analyzer. The to-be-designed system should be capable of automatically analyzing the cry sound and giving reliable estimation of the distress level of the infant. Besides the effectiveness and accuracy of the analysis, we will also pay great attention to the efficient implementation of the cry analysis system;
efficient in the sense of both engineering and clinical practices.

In Chapter 2, we review the history of infant cry study, survey the technology developments in this field, and discuss the main problems facing current infant cry research. Chapter 3 presents a survey of modern speech processing/recognition techniques, with emphasis on those methods which have the most potential for automatic infant cry analysis systems. The feasibility of applying speech processing technology to infant cry analysis and some technical considerations related to such application will be studied in Chapter 4. In Chapter 5, we discuss the problem of computer-based assessment of normal infant distress levels from the cry signal and present a "cry phoneme" system to effectively characterize normal infant cry signals. A cry signal-derived infant distress levels indicator — the H-value — is introduced and examined in this chapter. In Chapter 6 and 7, we present two different implementations of our automatic infant cry analyzer — one based on our newly developed vector quantization-kernel nonparametric classifier, and the other on the hidden Markov modeling technique. The performances of both systems are tested with real cry data. In Chapter 8, we summarize the accomplishments of the current phase of our research, and discuss future directions of automatic infant cry research.
2.1 History

Does infant crying mean anything more than his/her need for attention? This question has stimulated curiosity among physicians, parents, and scientists for a very long time.

As early as the nineteenth century, scientists began to believe that infant cry sounds contained information about the child, his physical and emotional well-being and about the cry-provoking situation. In 1832, Gardiner described the infant cry with reference to its location on a piano keyboard and as having an up-down melodic pattern. Also, when showed a series of photographs depicting various grimaces in the expressions of crying infants, Charles Darwin hinted at the notion that crying contains meaningful information.

In the early twentieth century, researchers started to use the International Phonetic Alphabet to note their perceptions of infant cries. It was not until the invention of the magnetic tape recorder and the sound spectrograph in the early 1950s that more systematic acoustic investigations on infant crying were able to attract wide attention. Lynip [65], in 1951, was probably the first scientist who investigated infant cries with a sound-spectrographic device — the sonograph. He measured the fundamental frequency of the infants’ early utterances and searched for different vowel-like patterns in the cry sound. In 1962, Wasz-Höckert et al. [122] reported their first findings on their sound-spectrographic cry analysis of birth, pleasure, hunger and pain cries. This actually marked the beginning of modern acoustic cry research.
Important fundamental facts were published by Truby, Bosma, and Lind [8, 114] in 1965. Using sound-spectrography, cineradiography, and the recording of the intraesophageal air pressure, they thoroughly described various acoustic phenomena in pain-elicited cries and the motions of different respiratory organs while these cries were recorded. In 1968, a monograph on the statistical analysis of cries of both healthy and abnormal infants was published by Wasz-Höckert et al. [118]. This work is a milestone in infant cry research since it set up the basic methodology of infant cry research which had been followed by many researchers for over two decades. Michelsson et al. [69, 70, 72–78] began using more sophisticated spectrographic techniques in the early 70s, and conducted a series of investigations on cries of infants with various diseases or abnormalities. They found that some characteristics of the cries of abnormal infants differ significantly from those found in normal infants.

Starting in the middle 70s, significant advances in electronics, signal processing theories, and computer technology occurred. These technological developments gradually had their influence in the infant cry research. Tenold et al. [113] in 1974 investigated the variability of fundamental frequency $F_0$ and cry spectra of full-term and premature infants' cries using cepstral and stationarity analysis. Golub and Corwin [33–35] in the early 80s reported their result on computer-aided cry analysis. They utilized computer signal processing techniques. Their findings suggest that the analysis of the infant cry holds promise for detecting a number of abnormalities. They also introduced a physioacoustic model of cry production in an attempt to relate the acoustic properties of the cry to the anatomic and neurophysiologic functioning of the infant. In 1986 and 1988, Fuller et al. [26, 27] reported their findings about the different acoustic characteristics among four types of infant vocalizations. A PDP 11/34 computer was employed in their study. In 1989, Lůdge and Gips [63], using a microcomputer with
specifically designed hardware, studied jitter of the cry uttered by newborn infants.

Recent years have also seen the beginning of the use of some sophisticated signal processing techniques in cry research. For example, the Fast Fourier Transform (FFT) analysis was employed in both normal and abnormal infant cry studies by Rapisardi et al. (1989) [103], Vohr et al. (1989) [116], and Fuller (1991) [25].

Nevertheless, in general, the tremendous improvement in recent electronics, signal processing and computer technology has not yet been fully appreciated and exploited by the researchers in infant cry research field. Some researchers seem to be reluctant to utilize modern signal processing theories and to replace the labor of “reading” spectrograms by advanced analysis techniques employing computers. It is somewhat surprising that the enthusiasm for infant cry research has even faded a little since the 60s and 70s.

To summarize, despite the many years of systematic research and the numerous publications on infant cry research as exemplified above, our knowledge about the infant cry is still surprisingly limited. Furthermore, no widely accepted practical application has emerged from the work of the past three decades in this field. While most of this is apparently due to the fact that the phenomenon of infant cry is of tremendous complexity, we believe that it is also due to the inadequacy of the tools which have been employed in the past. In the next section, a detailed survey of past work and an analysis of the major problem areas facing infant cry research is presented.

2.2 Major Directions in the Study of Infant Cries

Generally, in the past, infant cry research has approached the problem from one of five viewpoints.
1. Psychological Investigations on the Subjective Perceptions of Infant Cries

Cries from both normal (clinically healthy) and abnormal infants are used in such work. The cries in some cases are grouped according to the (sometimes presumed) types of stimulus that caused the cry (e.g., “birth”, “pleasure”, “hunger”, “cooing”, “pain”). The purposes of the investigations are usually aimed to obtain information about:

- the perception of adults of infant cries [9, 10, 39, 83, 115, 118], e.g., the parents’ ability to distinguish different cry-types and to identify the cries from abnormal infants;

- the responses of adults to infant cries [12, 19, 20, 81, 82, 112], e.g., the verbal and behavioral responses of the adults with varying temperament to various types of cries; and

- infant care and nurse training [44, 79, 111, 121];

2. Research on Physiological and Developmental Aspects

It has been postulated that infant crying is a reflection of a variety of complex neuro-physiological functions [34, 51, 114]. Many researches therefore aimed at investigating the infants’ physiological and developmental status by analyzing the cry signals. Such investigations include:

- the cries and the infants’ linguistic development [54, 55];

- the cries and the infants’ respiratory system physiology, and infants’ neurological and psychological development [8, 35, 43, 49, 56, 57, 68, 96, 114, 129];
3. Assessment of Emotional/Physical Situations of the Infant

The importance of assessing the emotional/physical situations of infants is obvious. More demanding is the problem of medical assessment of pain in infants. Approaches of assessment based on the analysis of such parameters as facial expression, heart rate, respiratory rate, transcutaneous oxygen level, body movement, and vocal behavior have been suggested [11, 26, 27, 37, 38, 52, 89]. Among them, methods based on cry analysis have the potential of being the most convenient.

4. Research on Abnormal Infant Cries for Diagnostic Purposes

Most researches of this group utilize spectrographic methods, sometimes with the aid of a computer, to get potentially diagnostically-valuable information from cry sounds of infants with specific diseases or abnormalities. Usually only pain cries, which are commonly induced by stimuli such as a pinch on the infants' arm or ear, are studied in an attempt to control and standardize the intensity of the stimulation. Investigations have been carried out to examine the correlations between various spectral and temporal attributes and the particular medical problems. These include oropharyngeal anomalies, asphyxia neonatorum, symptomless low birth weight, herpes encephalitis, congenital hypothyroidism, hyperbilirubinemia, bacterial meningitis, hydrocephalus, bradycardia, various forms of brain damage, malnutrition, genetic defects, and sudden infant death syndrome (SIDS) [34, 59, 60, 69–78, 90, 103, 109, 116].

5. Detection and Recognition of Infant Cries as a Kind of Warning Signals for Hearing Impaired Parents.

It is very important to alert deaf and hard of hearing parents whenever their baby is crying. Their constant awareness of the baby's emotional and physical situation is critical
for properly caring the infant. Cry research has been expected to provide data and methods which can lead to the building of reliable alerting devices for this purpose.

Unfortunately, little work has been done in this area. Vuorenkoski in 1970 [117] introduced an infant cry detection and analysis device, Cry Detector. The device was actually designed for collecting and analyzing cry samples in a clinical neonatal ward. It compares the measurements of signal energy from two different channels, the total channel of frequency range from 150—7000Hz, and the abnormally high pitch channel of frequency range from 1000—7000Hz, with preset thresholds. Any acoustic signal longer than 400 msec is considered a cry. The Cry Detector was manufactured by Special Instruments, Sweden in 1971, but objective evaluation has indicated that its usefulness in practice was limited [119, page 86].

In 1986, Lundh [64] reported a new baby-alarm based on tenseness information\(^a\) of the cry signal. Instead of the conventional baby-alarm for hearing impaired parents which only activate a flash or a vibrator when the sound emitted by the baby exceeds a predetermined threshold, this new device can give suggestions about the baby’s feelings (happy, crying, or distressed) by illuminating a picture on its panel which shows either a smiling face, a tearful face, or a screaming face. After being tested in ten deaf families, the device was judged to be useful and helpful, but it suffered from problems of identifying the baby’s “feelings” incorrectly and frequent false alarms.

2.3 Techniques and Instruments

In this section, we review the evolution of techniques and instruments employed in the past three decades and evaluate the development of the technology employed in cry analysis works.

\(^a\) — In Lundh’s definition, tenseness is represented by two types of signal: tense and relaxed. Tense sounds are strident and in associated spectrograms a high intensity of upper harmonics can be observed.
2.3.1 Auditory Analysis

The most readily available means for cry analysis is the human ear, and the art of diagnostic listening was already described in ancient times. Flatan and Gutzmann in 1906 used a graphophone \(^b\) to record infant vocalizations, and listened to recordings of cries of 30 neonates. They noted 3 infants with highly pitched phonations. Fairbanks [15] in 1942 listened to gramophone \(^c\) records to study the frequency characteristics of the “hunger wails” of one infant over a period of 9 months. Wasz-Höckert et al. in 1964 [120] found from their tape recordings that hunger, pain, pleasure, and birth cries can be identified auditorily. Valanne et al. in 1967 [115] reported that mothers can recognize the vocalizations of their own infants. Also, Partanen et al. in 1967 demonstrated that the pain cries of healthy infants could be differentiated from the cries of sick babies with certain diseases. They also showed that after a training period of approximate 2 hours, 82 pediatricians could diagnose normal versus pathological cries very accurately [92].

These reports confirm the common-sense experience that some basic information can be obtained by simply listening to cries.

2.3.2 Non-acoustical Analysis

To understand the anatomy of the vocal tract of infants, the functioning of the different respiratory organs during the cries, and the relationship between crying and the infants’ neurophysiological state, many cry studies have utilized methods which analyze and extract information from other than the acoustic signals of the cries.

\(^b\) — A type of phonograph using wax to record.

\(^c\) — Another type of phonograph which records and reproduces sounds by using a metal disk (instead of a wax cylinder) covered with a thin coat of oil or grease.
For example, Truby and Bosma et al. [8, 114] in 1965 used cineradiography, spiography, and intraesophageal pressure recordings together with sound spectrograms to investigate the relation of infants’ cry-sound to cry-act. In 1980 Langlois et al. [49] investigated the respiratory behavior of infants during both cry and non-cry vocalizations with impedance pneumography.

2.3.3 Time Domain Acoustic Analysis

To obtain time domain information about the cry signal, early researchers used direct writing oscillographs and other devices that could graph the sound magnitude or waveforms of the cry as a function of time on a paper chart.

Fisichelli and Karelitz, in the early 60s, used such a device to examine infant cries. They found that infants with diffuse brain damage required a greater stimulus to produce 1 minute of cry [45], and that the mean latency period between pain stimulus and the onset of cry was significantly longer for abnormal infants (2.6 sec) than that for healthy infants (1.6 sec) [16]. Wolff (1967, 1969) [124, 125] , using a similar device, measured inspiratory as well as expiratory phonations and found that in pain-induced cries, the cry units (one expiratory phonation) are longer in the beginning of the cry than at the end.

Time domain instruments usually are easy to operate, inexpensive and reliable. However, they can only provide gross information about the cry signal.

2.3.4 Sound Spectrographic Analysis

Sound spectrographs can provide permanent visual record of the sound, showing the distribution of energy in different frequency bands versus time. The technique was originally invented at the Bell Laboratories in the late 1940s in an attempt to present speech visually
to deaf people. While this goal was not accomplished, due to the excessive time-frequency complexity of speech signals, the technique itself has become very useful and important in many areas of acoustic signal research.

After its invention, the sound spectrographic technique quickly found its way into infant cry studies and became quite popular after 1950. A Scandinavian research team, headed by O. Wasz-Höckert and J. Lind, is most noted for setting up standards for the study of the infant cry with spectrographic techniques. In particular, it was them who introduced the definitions of many of the commonly-used spectrographic features. The following are short descriptions of these features as given by Golub and Corwin [35]:

1. Duration Features:

   - **Latency period**: The time between the pain stimulus applied to the child and the onset of the cry sound. The onset of cry is defined as the first phonation lasting more than 0.5 seconds.
   
   - **Duration**: This feature is measured from the onset of the cry to the end of the signal and consists of the total vocalizations occurring during a single expiration or inspiration. The boundaries are determined by the point on the spectrogram where the sound “seems” to end.
   
   - **Second pause**: The time interval between the end of the signal and the following inspiration.

2. Fundamental Frequency Features:

   - **Maximum pitch**: The highest value of the fundamental frequency $F_0$ on the spectro-
gram over the entire expiratory period.

- **Minimum pitch:** The lowest value of the $F_0$ on the spectrogram over the entire expiratory period.

- **Pitch of shift:** Frequency after a rapid increase in $F_0$ seen on the spectrogram.

- **Glottal roll or vocal fry:** Aperiodic phonation of the vocal folds, usually occurring at the end of an expiratory phonation when the signal becomes very weak and the $F_0$ becomes very low.

- **Vibrato:** Defined to occur when there are at least four rapid up-and-down movements of $F_0$.

- **Melody type:** Either falling, rising/falling, rising, falling/rising, or flat.

- **Continuity:** A measure of whether the cry was entirely voiced, partly voiced, or voiceless.

- **Double harmonic break:** A simultaneous parallel series of harmonics in between the harmonics of the fundamental frequency.

- **Biphonation:** An apparent double series of harmonics of two fundamental frequencies. Unlike double harmonic break, these two series seem to be independent of each other.

- **Gliding:** A very rapid up and/or down movement of $F_0$, usually of short duration.

- **Noise concentration:** High energy peak at 2000-2300 Hz, found both in voiced and voiceless signals.

- **Furcation:** Term used to denote a "split" in the $F_0$ where a relatively strong cry signal suddenly breaks into a series of weaker ones, each one of which has its own $F_0$ contour. It is seen mainly in pathological cries.

---

$d$ — In spectrographic analysis, usually only one expiratory phase of the cry is chosen and analyzed for each infant.
Chapter 2: HISTORY, ADVANCES AND CHALLENGES

- **Glottal plosives**: Sudden release of pressure at the vocal folds producing an impulsive expiratory sound.

A wealth of data and findings, especially concerning the abnormal infant cries, have been reported in past sound spectrographic cry studies. Undoubtedly, sound spectrography has been a useful tool and has significantly contributed to recent advancements in infant cry study. However, there are severe limitations to the technique that have hindered the widespread applications of spectrographic analysis in medical practice. First, the spectrogram has poor dynamic range and often inadequate frequency resolution. Secondly, visual inspection is required to extract acoustic information from the spectrogram. This is usually a long and tedious process that requires much expertise. The results are also subject to the expertise and biases of the inspector. As a result, it is not easy to analyze a large set of cry samples quickly, accurately, and consistently.

2.3.5 Computer-aided Analysis

It becomes more and more evident that computerized methods and highly efficient signal processing techniques are the future of infant cry research. Recently, researchers have begun to apply computer-aided analysis methods to infant cry study [25–27, 33, 63, 103, 116]. The computer and signal processing techniques used by these researchers, however, are usually simple and basic. In some cases, the electromechanical spectrography device was simply replaced by a computer that calculates the spectrogram with the Fast Fourier Transform (FFT) algorithm and displays the results on a terminal screen electronically, instead of printing on the traditional paper charts. While this approach improves the dynamic range of the spectrogram and provides flexibility in observing the signals, the essence of these “new” methods is in fact
identical to that of the traditional approach. The concepts of the spectrographic cry analysis are unchanged, the same types of acoustic features are measured, and the computer system only serves as an improved spectrography machine. This will be discussed in greater details in the next section.

2.4 Remaining Problems

By any standard, the infant cry research is still at the infancy stage itself! As recently stated by Lester [51], "... despite many years of programmatic research and numerous published articles on infant crying, we know surprisingly little about the topic." In this section, we will explore some of the unsolved and controversial issues, and summarize the problems facing infant cry research.

2.4.1 Controversy on How Much Information Infant Cry Conveys

It has become a widely accepted belief that infant crying carries meaningful information that a listener can use to estimate the infants’ physiological and/or emotional state. However, it is important to point out that this is still no more than a postulation. There are still some who argue that infant crying may contain too little information which can be extracted and used for the above purpose.

After comparing and surveying the data and results from a number of previous researches, Hollien (1980) [43, page 28] concludes that:

"... it would appear that the cries of normal infants carry too little perceptual information to permit auditors to identify the condition that evoked them. Therefore, it might be hypothesized that, within the normal home situation, the cry generally acts simply to alert the mother and most (if not all) of her
suppositions concerning the situation that evoked the crying behavior must be based on additional environmental cues.”

As to the perception of infants’ health states, Hollien’s conclusion is rather negative. He says:

“... it must be concluded that the health of the neonate probably cannot be deduced from perceptual analysis of his or her cries.” (ibid., on page 31)

He also surveyed previously published results of acoustic and spectral analyses on infant crying, and indicates that:

“... it appears possible that neonates may exhibit different types of cries and that these cry classes might related to some behavioral or physiological event/condition. Unfortunately, however, there is little or no quantitative (spectral) evidence to indicate what these cries might be or relative to how they might differ, one from another. Hence, the resulting data provide little guidance when the acoustic analyses of neonatal crying are to be considered.” (ibid., on page 39)

Hollien’s opinions are shared by Gardoski (1980) [28] who also discusses the meaningfulness of infant cry. He says (on pages 108–109):

“Previous research does not provide unequivocal support for the contention that the vocalizations of an infant are sufficient to specify the cause of distress. ... Every cry is not necessarily induced by hunger or some immediately obvious external cause. ... If difficulties arise in specifying the exact cause of a cry, even with the observer’s knowledge of the circumstances of the cry-provoking situation, inferring the probable cause of a cry must be
even more difficult without that additional knowledge. The task may become
impossible if neonatal crying in initially an innate mechanism that serves
primarily to promote proximity. . . . It has not been shown that the audible
signal alone can be indicative of anything more than the presence of gross
abnormality.”

In her book The Developing Child [5, page 156], Helen Bee also says:

“Infants may have several cries with somewhat different sound patterns,
but those different sounds do not seem to be related to different kinds of
discomforts or problems. Parents often feel that they can distinguish between
a ‘hunger cry’ and a ‘wet diaper cry’, but when children’s cries are recorded
and played back to parents, the parents are unable to tell the cause of the cry.”

These comments vividly illustrate the difficulties infant cry research now faces. At least,
it seems clear that either there is little situation-related information the infant imposes on
his/her cry sounds besides his/her desire for getting attention, or that the perceptual, as well as
the traditional spectrographic analysis, approaches have failed to effectively extract adequate
information to unveil the relationship between the cry production and the possible cause(s).
It can also be implied from Gardoski’s and Bee’s comments that the relationship between the
cause(s) and the crying may not be as straightforward as being suggested by the classification
of cry types (e.g., pain, hunger, cooing) used in the past.

2.4.2 Problems Facing Modern Computerized Infant Cry Analysis

As we pointed out earlier, computerized analysis is a trend of present infant cry research. It
appears to be the logical avenue to extract detailed information from cry signals so that the
complex relationships between the cause(s) and the cry production can be truly understood. Furthermore, computerization is crucial for bringing the results of infant cry research into clinical applications, a goal that is just as important as the research itself. Therefore, we view the development of computerized automatic infant cry analysis techniques as one of the objectives of present-day cry research.

Unfortunately, many studies in the past were motivated by the academic interest. Thus, the emphasis was on fundamental investigation with little thought given to clinical applications, and the engineering problems associated with the design of an automatic cry analyzer. To realize a practical application which by necessity must involve computerized automatic analysis of infant cries, modifications to the traditional methodology must be made. The following issues need to be reconsidered.

2.4.2.1 Excessive editing of the cry signal before analysis

In many past investigations, only the first expiratory phase of the cry signal was retained for analysis while the rest of the signal was usually discarded. This kind of editing usually deletes short signals, such as the glottal plosives, under 0.4 seconds duration. Such editing was necessary in the 60s since it reduced drastically the complexity and variations of the cry signal. It also made possible the handling of the usually long cry signals with a spectrograph, or sonagraph, which could only analyze few seconds of sound recording at a time. Obviously, this kind of excessive manual editing is not compatible with computer-based automatic cry analysis.

2.4.2.2 Traditional measures not suitable or efficient for automatic real-time analysis

Latency, duration, fundamental frequency (pitch), and formant frequencies have been the
most commonly investigated measures. Their introduction into cry research in the 1960’s was mainly due to the availability of instrumentation and the development of spectrographic techniques. Such measures are the most convenient that a spectrogram reader can take directly from the paper charts. The selection of measures to describe the cry on such basis does not guarantee that they will be suitable and/or efficient for modern computerized signal analysis.

2.4.2.3 Lack of unique definitions for many commonly used measures

Different researchers postulate different definitions for commonly used measures such as the duration and the fundamental frequency ($F_0$). For example, some researchers exclude the hyperphonated periods of the cry signal when they calculate the $F_0$-related measures while others include them. This is likely the cause of some conflicting reports. For example, both Wasz-Höckert et al. (1968) [118] and Fuller et al. (1986) [26] carried out investigations on the mean of $F_0$ of 2-6 month-old normal infants. Wasz-Höckert reported 530Hz $F_0$ mean for pain cries, 500Hz for hunger cries, and 440Hz for pleasure sounds while Fuller obtained 450Hz for pain, 490Hz for hunger, and 355 for pleasure sounds. So Wasz-Höckert concluded that pain cries have the highest mean $F_0$, but Fuller’s data shows that the highest mean $F_0$ is from hunger cries!

2.4.2.4 Past findings hard to be directly used in engineering applications

Most findings in the cry-study literature are reported in terms of statistical analysis. A commonly employed tool is the analysis of variance (ANOVA) and findings often report the statistical differences under certain significance level amongst the different quantities investigated. This kind of presentation of results is popular in fields such as social and behavioral sciences, but its usefulness in signal analysis is very limited.
In engineering applications, according to the theory of pattern recognition [22, 91], it is always desirable that the feature variables possess a separable distribution in the feature space. This enables the variables to be used as reliable indicators to distinguish and classify the different situations. Equally, if not more, important is that there must be an effective and efficient approach to electronically measure the variables of interest. Ideally, such feature variables should not correlate highly with each other. Showing statistical differences with certain significance level, as it is usually does in the cry-study literature, does not necessarily imply that the variables investigated meet the above criteria, especially the practical requirement of having effective and efficient methods to electronically estimate/extract them from the original signal. To sum up, while past findings from the study of cries are definitely valuable in guiding our research into automatic cry analysis, they are not in the form that engineers can directly use to design an automatic cry analysis device.

2.5 Summary

In this chapter, we have reviewed the history of infant cry research and summarized the methodological and technological diversities of the modern cry study. As we have seen, infant cry research has now attracted tremendous enthusiasm from various disciplines of both sciences and engineering.

Since infant crying is a very complex phenomenon, it is not surprising to see that many problems still remain unsolved after so many years of systematic investigations. These problems can probably be partially attributed to the lack of precise and objective analysis methods used in cry research in the past. The techniques used were usually unable to extract subtle differences from the cry signal, which may be characteristic of the conditions that
evoked the cry, and therefore will be crucial for the identification of such conditions from the
cry itself. Also, older techniques often relied on the final interpretations of the measurements
by human observers, making the interpretations subjective and difficult to reproduce.

The introduction of computerized signal analysis techniques to cry research holds great
promise since computerized techniques can often achieve accurate and fast detection of the
acoustic attributes in the cry signal, and enable the extraction of very subtle diagnostic
information that would otherwise be unattainable. Amongst the many advanced signal
processing/analysis techniques we are most interested in those of digital voice and speech
processing, and of computer pattern recognition. Tremendous technical, as well as theoretical
advancements have been achieved in these two fields recently. In the following chapter, we
will review some of the most important and applicable aspects in voice and speech processing
and pattern recognition techniques.
Chapter 3

AN OVERVIEW OF APPLICABLE
VOICE AND SPEECH PROCESSING TECHNIQUES

Like infant cry research, voice and speech processing is also an interdisciplinary subject. It draws on acoustics, linguistics, physiology, psychology, computer science, and engineering. The very purpose of voice and speech processing is to extract information carried in the voice and speech signal. It was not until the introduction of sophisticated digital signal processing technology that voice and speech processing have been able to grow into the realm of practical engineering applications. In this Chapter, we will outline briefly some important aspects of current voice and speech processing theory and technology. These aspects are chosen because of our intention of applying them to infant cry analysis. We will start with an introduction to the human vocalization organs and speech production, since most speech processing theories are based on physiological models. Then we will give a description to the computer speech production model which has guided the development of many modern speech processing algorithms. The rest of this Chapter will discuss the basic principles of pattern recognition, and will survey some important speech processing techniques, including vector quantization (VQ), hidden Markov modeling (HMM), and speech/speaker recognition techniques.

3.1 Human Vocalization Organs and Speech Production

As shown in the schematic diagram in Figure 3.1 [18], the human vocal system can be functionally divided into three main subsystems: 1. lungs and trachea/bronchi, 2. larynx (vocal cords), and 3. vocal and nasal tracts. The lungs and trachea/bronchi are the power
supply of the vocal system: air is compressed by the lungs and delivered to the system by ways of the bronchi and trachea. They also control the loudness of the resulting sound. The larynx contains the principal sound-generating mechanism. Finally the sound is modulated and enhanced by the vocal tract, and sometimes by the nasal tract.

In Figure 3.1, the lungs are represented by the air reservoir at the left. The force of the rib-cage muscles raises the air in the lungs to subglottal pressure $P_S$. This pressure expels a flow of air with velocity $U_G$ through the vocal cord orifice. The vocal cords are represented as a mechanical oscillator composed of a mass, spring, and viscous damping. The passing air flow causes the cords to vibrate and hence interrupt the air flow. The interrupted flow produces quasiperiodic, broad-spectrum pulses — the glottal pulses (see Figure 3.2), which excite the vocal tract. After modulation in the vocal tract, *voiced sounds* are produced. If at the same time the nasal tract is coupled to the vocal tract by opening the trapdoor of the velum, we hear the *nasal sounds*, otherwise *non-nasal* sounds are produced. *Unvoiced sounds*, like fricative sounds, are generated by forming a constriction at some point in the vocal tract, usually toward the mouth end, and forcing air through the constriction to produce
turbulence. A source of noise-like sound is thereby created.

In voiced sounds, the repetition rate of the pulses generated by the interrupted air flow at the vocal cords is defined as the pitch or fundamental frequency \(F_0\).

![Figure 3.2 A typical glottal pulse train.](image)

Excluding its power source (the lungs etc.), the operation of the vocal system comprises two basic functions — excitation and modulation [91], as shown schematically in Figure 3.4. The excitation takes place mostly at the glottis, which is the most important sound-generating organ in the larynx, although other points may also contribute, as mentioned earlier. Modulation is performed by the various organs in the vocal and nasal tracts.

The principal means of modulation is by filtering [91, 101, 107]. In the case of voiced sounds, the glottal pulses can introduce an abundance of harmonics, and the vocal tract, like any acoustical tube, has natural frequencies which are functions of its shape. These natural resonances are called formants, and they are the most important way of modulating the voice. Formants account for the generation of all the vowels and some of the consonants.

Mathematically, if we let the function of glottal pulse waveform be \(g(t)\) and the vocal tract impulse response be \(h(t)\), then the resulting speech signal will be the convolution of \(g(t)\) with \(h(t)\). In the frequency domain, if the spectrum of \(g(t)\) is \(G(f)\) and the vocal tract transfer function is \(H(f)\), the spectrum of the output will be \(G(f)H(f)\). Figure 3.3 shows this schematically.
Figure 3.3 Spectra in speech generation — (a) idealized spectrum of glottal pulse train, (b) frequency response of vocal tract, where the peaks correspond to formants, and (c) spectrum of the resultant speech.

Figure 3.4 A two-stage speech generation model.

The excitation-modulation model shown in Figure 3.4 serves as the basis of modern speech and voice processing theories.
3.2 The Digital Model of Speech Production and Linear Prediction Coding (LPC) of Speech Signals

3.2.1 Time-varying Filter Model of Speech Generation

In computerized digital speech signal processing, the two-stage conceptual model shown in Figure 3.3 and Figure 3.4 has led to the introduction of the following digitalized time-varying filter model of speech generation, as shown in Figure 3.5 [101, 107].

![Diagram of digital time-varying filter model of speech production]

*Figure 3.5 The digital time-varying filter model of speech production.*

In this digital model, excitation is represented by an impulse train generator and a random noise generator. For voiced speech, the digital filter is excited by the impulse train generator that creates a quasiperiodic impulse train in which the spacing between impulses corresponds to the fundamental period of the glottal excitation. For unvoiced sounds, the filter is excited by the random number generator that produces flat spectrum noise. In both cases, an amplitude control, $G$, regulates the intensity of the input to the digital filter, and thus the volume of the resultant speech.

Modulation is furnished by a time-varying digital filter whose coefficients are changed with time so that the transfer function of the filter can always approximate the overall spectral
characteristics of the transmission properties of the vocal and nasal tracts, and the spectral properties of the glottal pulse train.

Usually, such a time-varying system is very hard to manipulate in practice because of its high computational complexity. Fortunately, there is a way to simplify the problem drastically. Since the vocal/nasal tract changes its shape relatively slowly during speech production, it is possible, and also necessary, to assume that over a very short time interval (10–20 msec) the characteristics of the vocal/nasal tracts are almost constant. Thus, within this short time period, the time-varying digital system in Figure 3.5 can be treated as a much simpler time-invariant system [91, 101, 107]. The parameters of this speech generation model, i.e. the digital filter coefficients, pitch period, voiced/unvoiced control, and amplitude, are assumed fixed during each of the short intervals.

3.2.2 Linear Prediction of Speech Signal

The digital filter in Figure 3.5 can be simply chosen as a $p^{th}$ order all-pole system (in statistics, a $p^{th}$ order autoregressive model). Provided $p$ is properly chosen, this assumption is adequate and reasonable for most practical speech applications. Detailed discussion on the all-pole model and its application to speech analysis are found in [67, 101, 107]. More fundamental theoretical analysis on the selection of $p$ is found in [1, 93].

Thus, the digital filter has a transfer function as

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}},$$  \hspace{1cm} (3.1)

where $G$ is the gain parameter which controls the volume of the output speech, and $a_k$ is the coefficient of the filter. The speech output $s(n)$ is related to the excitation $u(n)$ by the
following difference equation,

\[ s(n) = \sum_{k=1}^{p} a_k s(n - k) + G u(n). \]  

(3.2)

For each short-time interval, we can estimate the gain parameter, \( G \) and the filter coefficient, \( a_k \) from the speech samples in a straightforward and computational-efficient manner, as being discussed below.

The method is called linear predictive coding (LPC) analysis. Suppose that we predict the speech signal at time \( n \) with a \( p^{th} \) order all-pole linear predictor with prediction coefficients, \( \alpha_k \), i.e.,

\[ \hat{s}(n) = \sum_{k=1}^{p} \alpha_k s(n - k). \]  

(3.3)

Then the prediction error, \( e(n) \), is defined as

\[ e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n - k). \]  

(3.4)

It can be seen by substituting equation (3.2) into equation (3.4) that, if \( \alpha_k = a_k \) and the speech signal really does obey the model of equation (3.2), then \( e(n) = G u(n) \). This means that, between the excitation impulses of voiced speech, the prediction error should be very small if the predictor coefficient, \( \alpha_k \) is equal to the actual parameter \( a_k \) of the vocal tract transfer function.

Therefore, by minimizing the average squared prediction error

\[ E_n = \sum_m e_n^2(m) = \sum_m (s_n(m) - \hat{s}_n(m))^2 \]  

(3.5)
where \( s_n(m) \) is a segment of the speech waveform that has been selected in the vicinity of sample \( n \), i.e.,

\[
s_n(m) = s(m + n),
\]

we can obtain an estimation of the predictor coefficients \( \alpha_k \) and the gain parameter \( \hat{G} \). Thus,

\[
\hat{H}(z) = \frac{\hat{G}}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}
\]

will be a good approximation to the vocal tract transfer function [91, 101, 107].

Recently, in many areas of speech processing LPC coefficients have been used with excellent results to represent speech signals. This is partly due to the availability of very efficient algorithms which compute the LPC coefficients [91, 101]. Still more important is that from the LPC coefficients we can economically derive many very useful speech parameters, such as the formant frequencies, cepstrum coefficients, PARCOR coefficients, log area ratio coefficients, and others.

### 3.3 Pattern Recognition

Any problem concerning discrimination or classification of signals may be thought of as a pattern recognition problem. In its simplest form, a pattern recognition system consists of two major function blocks: feature extraction and classification, as shown in Figure 3.6.

![Figure 3.6 A pattern recognition system.](image-url)
Prior to feature extraction, we have to determine which features should be measured. The selection of proper features is usually crucial to the success of a pattern recognizer. Features are supposed to be [21, 91, 108]:

1. easy to measure (automatically) and computationally feasible;
2. varying widely from class to class;
3. invariant or insensitive to extraneous variables and the common distortions of the patterns;
4. not correlated with other features; and
5. contain little redundancy.

Unfortunately, at present, there is very little available in the way of a general theory for the selection of features. The decision of what to measure to obtain features is usually application-dependent. This decision is often based on either the importance of the features in characterizing the patterns, or on the contribution of the features to the performance of the recognizer (i.e., the accuracy of recognition) [21, 108]. In some cases, simulation may aid in the choice of appropriate features.

The classification problem can be defined mathematically as follows. Suppose that $N$ features are extracted from each input pattern, as shown in Figure 3.6. These $N$ features compose a vector $\mathbf{x}$, called the feature vector. This vector corresponds to a point in the $N$-dimension feature space $\Omega_\mathbf{x}$. The problem of classification is to assign each possible feature vector, or equivalently the corresponding point in the feature space, to a proper pattern class. This can be interpreted as the partition of the feature space $\Omega_\mathbf{x}$ into mutually exclusive regions, where each region corresponds to a particular pattern class. The boundaries of the partition are called the decision boundaries, which are described by hyperplanes in the $N$-dimension.
feature space $\Omega_X$. Thus, solving the classification problem is equivalent to specifying these hyperplanes. If all the a priori probability information is known, we can determine the hyperplanes analytically, otherwise, we have to get that information by training the classifier with patterns whose classification is already known [21, 108].

3.4 Automatic Speech/Speaker Recognition

3.4.1 Automatic Speech Recognition

Speech recognition problems can be divided into the following categories [91, 101]:

1. **isolated-word recognition** — recognition of words separated by pauses;
2. **word spotting** — the detection of occurrences of a specified word in continuous speech;
3. **continuous-speech recognition** — recognition of words without pauses in between; and
4. **speech understanding** — the extraction of the meaning of an utterance with the use of stored information about the language being spoken.

Among the above, the isolated-word recognition is the least difficult; the pauses between words significantly simplify the problem. However, many of the techniques developed for isolated-word problem have been carried over into word-spotting and continuous-speech recognition.

The following technical aspects are important in the design of speech recognition systems [101].

1. **Feature selection.**

The choice of features is usually application-dependent. Commonly used features include amplitude, zero-crossing rate, gross spectrum balance, LPC coefficients, etc.
Chapter 3: VOICE AND SPEECH PROCESSING TECHNIQUES

2. **Word boundaries detection.**

   This is especially crucial in isolated-word recognition and word-spotting, since these recognizers usually treat each word in the vocabulary as their basic recognition unit to perform pattern-matching.

3. **Time normalization.**

   The traditional method of time normalization is *dynamic time warping* (DTW) which can efficiently normalize the non-uniform distortion on the time axis of the unknown pattern. In recent years, newly developed HMM techniques can achieve the same result with higher efficiency \[50, 91, 99\].

### 3.4.2 Speaker Recognition

There are two related but different areas of speaker recognition: **speaker verification** (SV), and **speaker identification** (SI) \[88, 91\]. Given a speech signal, the former involves verifying whether the speaker is who he or she claims to be, while the latter involves finding the identity of the person most likely to have spoken. Both SV and SI use a stored data base of reference patterns for \(N\) known speakers, and employ analysis and decision-making techniques similar to those used in speech recognition. A speaker recognizer (either an SV or an SI) can work in two different ways — text-dependent or text-independent. A text-dependent recognizer is easier to implement since the templates and the unknown patterns come from the same words \[88, 91, 104\].

Since the meaning of the sentence spoken by the speaker is not important for speaker recognition, the criterion of feature selection may differ from that of speech recognition. In particular, more emphasis could be placed on the features which convey information about
the acoustic differences between different speakers. With this in mind, many features have been explored as potential indicators of a speaker's identity [106, 123], including:

1. vowel formant frequencies and bandwidths and glottal source poles;
2. location of pole frequencies in nasal consonants;
3. pitch contours over a selected sentence; and
4. timing characteristics, specifically the rate of change of the second formant over a selected period of sentence.

As acoustic cues to a speaker's identity are spread throughout each of his or her utterances, many systems utilize templates of averaged parameters [88]. This statistical approach is most useful in text-independent cases, since the time sequences of training and testing utterances do not correspond. Recently, vector quantization (VQ), combined with LPC-related parameters, such as predictor coefficients, cepstral coefficients, reflection coefficients, orthogonal LPC parameters, and log-area ratios has been successfully used in both text-independent and text-dependent recognizers [91, page 338]. The advantage of using these parameters is that very efficient computer algorithms are available to extract these parameters from the speech signal.

3.5 Vector Quantization

Vector quantization (VQ) is a coding technique initially developed for communications. Recently, its applications have been successfully extended into many other fields of digital signal processing, such as classification [46, 126], image and speech data compression [31, 32, 84], automatic speech/speaker recognition [50, 88, 91, 100, 110].

Generally, a vector quantizer, composed of a code-book (or reproduction alphabet) and a definition of distance measure, can map a real input vector onto a discrete symbol [61].
The code-book consists of a set of fixed prototype vectors with the same dimensions as the input vector. To perform the mapping, the distance between the input vector and each prototype vector in the code-book is calculated using the pre-defined distance measure. The prototype vector which has the smallest distance to the input vector is found and becomes the representation of the input vector. A more detailed description of VQ is given in the next section.

3.5.1 Some Details of Vector Quantization

An M-level d-dimension quantizer is a mapping, \( q(\cdot) \), that assigns to each input vector, \( x = (x_0, \cdots, x_{d-1}) \), a reproduction vector, \( y_k = q(x) \), drawn from a finite reproduction alphabet (or code-book), \( \hat{A} = \{y_i; i=1,\cdots,M\} \). The \( q(\cdot) \) is usually chosen as a minimum-distance mapping. This means that the reproduction vector \( y_k \) is chosen such that

\[
d(x, y_k) = \min_{i} d(x, y_i)
\]

(3.8)

where \( d(\cdot, \cdot) \) is any nonnegative distance measure defined on the d-dimension space, and \( d(x, y_k) \) is called the distortion of the quantization. Obviously, a minimum-distance mapping, which is completely defined by the reproduction alphabet \( \hat{A} \) along with the definition of the distance measure, uniquely describes a Dirichlet partition, \( S = \{S_i; i=1,\cdots,M\} \), of the sample space.

An M-level quantizer is said to be optimal if it minimizes the expected distortion

\[
D(q) = E\{d(x, q(x))\},
\]

that is, \( q^* \) is optimal if for any other quantizer \( q \) having \( M \) reproduction vectors, \( D(q^*) \leq D(q) \).

In speech recognition, a feature vector is usually extracted from each frame of speech signal. Then vector quantization is applied to convert the feature vector into an integer (the
index of the nearest prototype vector) which is often only one or two bytes long. In many cases, this drastically compresses the original speech data with a minimal information loss, and significantly simplifies the design of the speech recognition system.

3.6 Hidden Markov Models

The hidden Markov model (HMM) is a stochastic process that has been found to be particularly useful in automatic speech recognition (ASR). As one of the most important pattern-matching approaches used in ASR, the HMM is capable of dealing with the variations in both the temporal structure and the spectral patterns of speech signals.

3.6.1 Description of HMMs

An HMM is a collection of states connected by transitions. Each transition from state $i$ to states $j$ has a transition probability $a_{ij}$. When state $i$ is reached at time $t$, a symbol $O_t$ is emitted, where $O_t$ is from a finite symbol set $\{v_1, v_2, \ldots, v_M\}$, i.e., $O_t \in \{v_1, v_2, \ldots, v_M\}$, where $M$ is the number of all possible symbols in the HMM. Each state is associated with an output symbol probability distribution, which defines the conditional probability of emitting a certain observation symbol given that the state is reached.

Figure 3.7 shows an example of an HMM. The three circles represent the three states of the HMM. At a discrete time instance $t$, the model reaches one of the three states and emits an observation symbol $O_t \in \{v_1, v_2, v_3, v_4\}$. At instance $t + 1$, the model moves to a new state (sometimes itself) by taking a transition and emits another observation symbol, and so on. This continues until a final terminating state is reached at time $T$. All the transition probabilities are tabulated in a $N \times N$ transition matrix, $A = [a_{ij}]$, where $N$ is the number of states in the model, and $a_{ij}$ is the probability of occupying state $j$ at time $t + 1$, given state $i$ at
time $t$. Matrix $B$ is composed of the output symbol probability distributions of all states. Its element $b_{jk}$ is defined as the probability of emitting symbol $v_k$ when state $j$ is reached. The initial state distribution $\pi = [\pi_i]$ determines the probability of occupying state $i$ at time $t = 1$. A convenient notation can be used to indicate an HMM and its parameters, $\lambda = (A, B, \pi)$.

3.6.2 Three Problems

An HMM can be used either as a generator of symbol observation sequences, or as a model for analyzing how a given observation sequence was generated. To make HMMs useful in automatic speech recognition, there are three problems of particular interest [99]:

Problem 1— Given an observation sequence $O = O_1O_2\cdots O_T$, where $O_t \in \{v_1, v_2, \ldots, v_M\}$, and a certain hidden Markov model $\lambda = (A, B, \pi)$, how can $P(O|\lambda)$, the generation probability of the observation sequence, be computed efficiently?
Problem 2— Since an observation sequence $O$ can be emitted by a model $\lambda = (A, B, \pi)$ passing through one of many different state sequences $Q = q_1 q_2 \cdots q_T$, given the observation sequence $O$, and the model $\lambda$, how can the optimal (i.e., the most likely) state sequence $\hat{Q}$ be determined?

Problem 3— How can the model parameters $\lambda = (A, B, \pi)$ be adjusted to maximize $P(O|\lambda)$?

Using the methodology of dynamic programming, solutions to these problems have been found, and efficient computer algorithms exist as well (see Appendix D, and also [95, 99] for more discussions on the algorithms).

3.6.3 Speech Recognition with HMMs

The following postulate is the key to applying HMMs to speech recognition: an utterance is produced by the vocal organs passing through an ordered sequence of stationary states of different durations; the output from each state (which actually forms the utterance) can be regarded as a probabilistic function of the state. This is a crude and drastically simplified description of the complexities of speech; speech is a smooth and continuous process and does not jump from one articulatory position to another. However, the success of HMMs in speech recognition demonstrates that if the hidden Markov model is correctly chosen and its parameters are properly set, it captures enough of the underlying mechanism to be effective.

As we discussed earlier, speech signals are usually represented as sequences of feature vectors, and each vector corresponds to a very short time interval of the speech. The HMMs described above can deal with only a finite number of symbols but the feature vectors are real vectors with their elements continuously valued. Therefore the feature vectors need to
be quantized so that HMMs can be applied. Usually the vector quantization technique is used for this purpose.

The following example shows how HMMs work in isolated-word recognition. Assume we have a vocabulary of \( K \) words to be recognized. For each word, we have a training set of \( L \) tokens. Also, we can have an independent testing set. Before applying HMMs, each of the \( L \) tokens and each of the words in the testing set is represented by a sequence of symbols. To build the recognizer, we must:

1. design (the topology of) one or more HMMs for each of the \( K \) words in the vocabulary, and
2. estimate the parameters (transition and output probabilities) of each HMM, using the sequences of symbols representing the \( L \) tokens of this word (This process is called the *training* of the HMMs).

To perform the recognition we then perform the following steps:

1. for each unknown word in the testing set, characterized by its sequence of symbols, calculate the probability of producing this sequence from all the trained HMMs, and
2. choose the word whose model has the highest probability of generating the unknown word.
Chapter 4

FEASIBILITY STUDIES ON APPLYING SPEECH PROCESSING TECHNIQUES TO CRY ANALYSIS

As it was already mentioned in Chapter 2, the analysis of infant cry has long been considered as having the potential to provide diagnostic information in a clinical setting. Thus, many studies have been undertaken to recognize the characteristics of cries uttered by healthy infants in different situations and by ill infants with various diseases or abnormalities. These studies used various techniques, such as auditory and acoustic analysis, spectrography, and computer-aided analysis. However, extracting situational and diagnostic information from infants' cries using these methods is usually a long and tedious process that requires much expertise, and is sometimes vulnerable to the subjective bias of the observers. Accordingly, there are still many unsolved problems and unanswered questions about such analysis of the infant cry and successful clinical applications are extremely rare.

We will show in this chapter that work over the past three decades has shown that infant cries have many speech-like elements. Furthermore, if it can be shown that models of the generation of infant cries possess enough similarities to those of speech generation, an excellent opportunity exists to apply techniques developed for automatic speech processing/recognition to the analysis and classification of infant cries.

Many new and powerful techniques have been developed recently in the field of automatic speech recognition, including LPC analysis, hidden Markov modeling (HMM), vector quantization (VQ), and others. With the help of these techniques accurate automatic recognition
Chapter 4: FEASIBILITY STUDIES

of continuous speech is becoming a reality [50]. We believe that it is also feasible to develop highly efficient methods for infant cry analysis. In this chapter, we study this feasibility using both theoretical analysis and experimental methods.

4.1 Investigating the Physiological Similarities between Infant Crying and Adult Speech

This investigation is focused on answering the following question: Do infants vocalize in ways which are similar to the ways adults generate speech? Because of our intention of applying speech processing/recognition techniques to analyze infant cries, it is crucial to show theoretically or experimentally that these techniques are valid and applicable to infant cries.

In the following literature survey, we establish the physiological similarities between infant vocalizations and adult speech by describing a physioacoustic model of infant crying which is obviously compatible with the production model of adult speech.

4.1.1 A Survey: Physiology of Infant Crying and a Physioacoustic Model of Cry Production

Infant crying is the result of complex interactions between many anatomic structures and physiological mechanisms. These interactions involve the central and peripheral nervous system, the respiratory system, and a variety of muscle groups [35].

4.1.1.1 Respiration patterns of crying infants

The functions of vocal system are closely integrated with functions of the respiratory system which supplies oxygen to the blood and removes carbon dioxide from the body. Thanks to the cooperation of the many organs and muscular groups in our vocal and respiratory systems, we do not have to stop breathing to speak. These organs and muscles often work for both vocalization and respiration simultaneously. We learn to speak and breathe at the
same time as young children. The breathing accompanied with speaking is called *speech breathing* [49]. The chief physiological requirement for speech breathing is to have short inhalations, and long, controlled exhalations. It is similar for infant vocalizations. Infants must be able to manage their respiratory system to meet the basic criteria of speech breathing. They have to shorten the inspiration and extend the expiration, to quickly intake an adequate volume of air, and to augment or diminish the relaxation pressure generated by the chest wall’s elastic rebound. They also have to constantly adjust the activity of the respiratory muscles to maintain a high and constant pulmonic pressure during expiration. Observations on infant cry respiration suggest that “at a very early age the infant can very well negotiate a short inspiration and a prolonged expiration. . . . The infants’ respiration system is capable of meeting those criteria for adult speech breathing” [49]. It was also pointed out by Lieberman (1985) [55] that “three aspects of the intonation pattern of normal human newborn cry are similar to the patterns that adult speakers usually use” and “many of the linguistically salient aspects of human speech can be seen in the vocal behavior of infants.”

**4.1.1.2 Models of infant cry generation**

Many scientists believe that at birth the infants’ supralaryngeal vocal tract resembles most closely a uniform cross-section tube with a length of approximate 7.5 cm, open at both ends during the production of sound. Although, anatomically speaking, the infant’s vocal apparatus is significantly different from that of the adult, in both position of various organs and the absolute mobility, from the acoustical viewpoint they both operate on similar principles of voice generation [56, 57].
Based on detailed investigations with both sound spectrographs and cineradiographs, Truby and Lind in 1965 [114] suggested an infant vocalization model that is very similar to one commonly used to model adult speech production in speech processing. Their primary concept is: infant vocalization derives from sound source plus resonance, i.e., from the amplification of simple or complex quasiperiodic physical excitation. It can be formulated as:

\[ \text{Vocalization} = \text{Source} + \text{Resonance}. \]  \hspace{1cm} (4.1)

They divided infant vocalization into three categories: (1) phonation or "basic cry", resulting from more or less simple glottal oscillation, with adequate amplification; (2) dysphonation or "turbulence", caused by sometimes simple, sometimes complex glottal excitation, with or without supraglottal excitation, with or without frication; (3) hyperphonation or "shift", resulting from certain constraints on the source plus or minus simple and/or complex glottal participation, with or without supraglottal excitation, with or without frication.

Extending the Truby and Lind’s model, Golub and Corwin introduced in 1985 a physioacoustic model of the infant cry production which is even closer to modern speech generation concepts [35]. As shown in Figure 4.1, their model divides cry production into four parts. The first part is the subglottal (respiratory) system which supplies the necessary air pressure \( P_S(t) \) below the glottis for driving the vocal folds. The second part is the sound source located at the larynx. Mathematically, the sound source can be described, in the frequency domain, as either a periodic source \( S(f) \) or a turbulence noise source \( N(f) \). Frequently these sources operate simultaneously. \( S(f) \) results from the vibration of infants’ vocal folds and the turbulence noise is most likely produced by the turbulence created by forcing air through a small opening left by incomplete closure of the vocal folds [35]. The third part
Figure 4.1 A simplified view of the cry production model.
of the cry model consists of the vocal and nasal tracts located above the larynx. This part functions as an acoustic filter that has a transfer function $T(f)$, determined by the shape and length of the vocal and nasal tracts and the degree of nasal coupling. The fourth part of the system is the radiation characteristic $R(f)$ that describes the acoustic transmission between the mouth of the infant and the auditor.

Figure 4.2 shows the spectra of idealized sound sources, vocal tract transfer function, radiation function, and the spectrum of the output cry sound. The amplitude of the sound is directly related to the subglottal pressure $P_S(t)$.

![Figure 4.2](image)

*Figure 4.2 (A) Spectrum of an idealized periodic source; (B) spectrum of an turbulence source; (C) vocal tract transfer function; (D) spectrum of an idealized radiation characteristic; and (E) spectrum of an idealized output. (after Golub and Corwin [35])*

As to the three different types of infant cries proposed by Truby and Lind, Golub and Corwin explained them in their new cry production model. They supposed that the *phonation*
is produced by the vocal folds vibrating fully at an $F_0$ range of approximately 250–700 Hz; *hyperphonation* results from a “falsetto”-like vibration pattern of the vocal folds with an $F_0$ range of about 1000–2000Hz (likely only a thin portion of the vocal ligament is involved in this mode); and *dysphonation* contains both a periodic and aperiodic sound source and occurs when turbulence noise is generated at the vocal folds.

In summary, Golub and Corwin's cry production model can be described in the frequency domain as:

\[
\text{Output} = \text{Source} \times \text{Filter}. \tag{4.2}
\]

It is not difficult to realize that equation (4.1) and (4.2) describe exactly the same principle the excitation-modulation speech production model shows (see Figure 3.4, Chapter 3). Furthermore, by comparing Figure 3.5 of Chapter 3 with Figure 4.1, we find that, if we set the transfer function of the digital filter of Figure 3.5 to be $S(f)T(f)R(f)$, then the digital model in Figure 3.5 will be equivalent to the cry production model of Figure 4.1.

The above comparisons between the models clearly show that not only does the same excitation-modulation principle apply to infant cry production as it does to adult speech generation, but also that the modulation process can be similarly formulated in terms of a time-varying transfer function. This suggests that the speech generation model shown in Figure 3.5 can also be used to describe the production of infant crying, provided the parameters in the model are appropriately chosen. This will allow the applications of signal

---

*See Figure 4.1, where $S(f)$, $T(f)$, and $R(f)$ are the glottal source spectrum, the transfer function of supraglottal system, and mouth radiation function, respectively.*
processing/recognition techniques, which were initially developed for adult speech analysis, to the analysis of infant cries.

4.2 Formant Frequency Estimate Using an LPC Model

To further examine the validity of representing infant crying with models conformable with the adult speech generation theory discussed in section 3.1, we conducted the following formant frequency estimate experiment.

We first hypothesized that infant cry generation can be adequately modeled by the systems depicted in Figure 3.4 and 3.5, and further, that the digital filter in Figure 3.5 can be defined as having an all-pole transfer function. Then, based on these assumptions, we used the linear prediction coding method to estimate the formant frequency tracks of the cry signal. Finally, we compared the LPC analysis results with the formant frequencies observed on the spectrogram of the same cry. We expected that the LPC results would match the spectrogram observation, if our two above-mentioned assumptions are correct. Our experimental results proved that our expectation was valid.

4.2.1 Theoretical Analysis

By the definition, formant frequencies or formants are the resonances of the vocal tract. Each of the formants will cause a local maximum in the magnitude curve of the frequency response of the vocal tract at the frequency of that formant, as shown schematically in Figure 3.3 (b). The locations of these local maxima or peaks are exclusively determined by the shape of the vocal tract. Therefore, the formant estimate method used in this experiment involves 1) estimating the transfer function of the infants' vocal tract, and 2) locating the peaks in the
magnitude response derived from this transfer function. The locations (frequencies) of the peaks then will be our estimations of the formants.

Suppose that a \( p \)th order all-pole digital filter with a transfer function of

\[
H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}
\]  

is used to model the generation of the cry signal. \( G \) is the system gain constant and has no effect on the formants. To determine this transfer function, we need to estimate the values of all the \( a_k \)'s. As discussed in Section 3.2.2, we can use the LPC method to determine the \( a_k \)'s from the cry signal, and obtain an estimation of the transfer function

\[
\hat{H}(z) = \frac{\hat{G}}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}
\]  

where \( \hat{G} \) and \( \alpha_k \) are estimates of \( G \) and \( a_k \), respectively.

The vocal tract magnitude response can then be determined from the estimated transfer function with the procedure described below.

To establish the procedure, first let

\[
A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k}.
\]  

Then we can rewrite equation (4.4) as

\[
\hat{H}(z) = \frac{\hat{G}}{A(z)}.
\]  

The magnitude of the frequency response of the vocal tract is then given by

\[
|\hat{H}(\omega)| = \left| \frac{\hat{G}}{A(z)} \right|_{z=e^{j\omega}}.
\]
We evaluate \( |\hat{H}(\omega)| \) for \( N \) points over \( 0 \leq \omega < 2\pi \), that is, to compute
\[
|\hat{H}(k)| = \frac{\hat{G}}{A(e^{-2\pi jk/N})}, \quad k = 0, 1, \ldots, N - 1.
\] (4.8)
This can be conveniently effected by finding \( \hat{H}(k)^{-1} \) for \( k = 0, 1, \ldots, N - 1 \) first. From equation (4.4), we have
\[
\hat{H}(k)^{-1} = \frac{1}{\hat{G}}A(e^{-2\pi jk/N}), \quad k = 0, 1, \ldots, N - 1.
\] (4.9)
By defining \( \beta_0 = 1 \) and \( \beta_i = -\alpha_i \) for \( 1 \leq i \leq p \), we can rewrite equation (4.5) as
\[
A(z) = 1 - \sum_{i=1}^{p} \alpha_i z^{-i} = \sum_{i=0}^{p} \beta_i z^{-i}.
\] (4.10)
Then we have
\[
\hat{H}(k)^{-1} = \frac{1}{\hat{G}} \sum_{i=0}^{p} \beta_i e^{2\pi jik/N} \quad \text{(4.11a)}
\]
\[
= \frac{1}{\hat{G}} \sum_{i=0}^{N-1} \beta_i e^{2\pi jik/N}, \quad k = 0, 1, \ldots, N - 1, \quad \text{(4.11b)}
\]
where \( \beta_i = 0 \) for \( i > p \). It should be noted that the right-hand side of equation (4.11b)
is simply the Discrete Fourier Transform (DFT) of sequence \([\beta_0, \beta_1, \ldots, \beta_p, 0, 0, \ldots, 0]\).
Therefore, we evaluate the magnitude of the frequency response of the vocal tract transfer function in the following steps:

1. estimate a set of \( p \) LPC coefficients \([\alpha_1, \alpha_2, \ldots, \alpha_p]\);
2. compute the DFT of the \( N \) point sequence \([1, -\alpha_1, \ldots, -\alpha_p, 0, 0, \ldots, 0]\) with the FFT algorithm. This gives \( H(k)^{-1} \) for \( k = 0, 1, \ldots, N - 1 \); and
3. compute \( |H(k)^{-1}| \) for \( k = 0, 1, \ldots, N - 1 \) and then take the reciprocal of the result to obtain \( |H(k)|, \ k = 0, 1, \ldots, N - 1. \)
4.2.2 Experiment

In this experiment, our samples of infant cries came from a phonograph record found in Wasz-Höckert's monograph (1968) [118]. Before the LPC analysis, the cry signals were low-pass-filtered with a cut-off frequency of 7.5 kHz and digitized at a sampling frequency of 15 kHz with 12 bit resolution. Figure 4.3 shows the flow diagram of our LPC formant estimation experiment.

Figure 4.3 LPC formant estimate experiment of infant cries.
Chapter 4: FEASIBILITY STUDIES

The digitized signal is segmented into frames, each with fixed length (256 points, or 17.1 msec). Silence periods are detected by checking the short-time energy and are deleted in the following analysis. Pre-emphasis removes dc components and flattens the spectrum so as to reduce the effect of the glottal waveform and lip radiation characteristics [18, 36]. Since the autocorrelation method (see Appendix A) is used for the LPC analysis in our experiment, windowing is necessary to reduce the Gibbs effect which is due to the segmentation of the signal [101, 107]. The 14\textsuperscript{th} order predictor is used in our experiment, i.e., \( p = 14 \) in equation (4.3). Thus, after the LPC analysis we obtain a set of 14 LPC coefficients from each frame of the signal. Using the procedure described in Section 4.2.1, we estimate the magnitude response \(|H(\omega)|\) of the vocal tract transfer function (or called the spectrum envelope of the voice signal) from the predictor coefficients. Then, a peak-picking program locates all the local maxima or peaks of the magnitude response, and the possible formant frequencies corresponding to this frame of the signal are estimated from the locations of the peaks.

The above analysis procedure repeats for each frame of the signal, and at the end of the experiment the trajectories of the estimated formants are plotted.

4.2.3 Results

The results from our experiment show that the formant frequencies which were estimated with the LPC method are consistent with those observed on the spectrogram of the cry signal. Figure 4.4 (a) and Figure 4.4 (b) are examples, where Figure 4.4 (a) shows the estimated formant trajectories of a pain cry by using the above LPC method, and Figure 4.4 (b) shows the spectrogram of the same cry, respectively.
Figure 4.4 (a) Estimated formant frequencies by using the LPC-based method on a pain cry. Estimations marked with circles indicate the presence of the hyperphonation. (b) Spectrogram of the same cry. In both graphs, the trajectories of the first formant, second formant, and third formant are labeled by A, B, and C, respectively.
These results serve as evidence that:

1) normal infant crying can be modeled by a system similar to the adult speech generation system depicted in Figures 3.4 and 3.5, and

2) an all-pole model, as well as the accompanying LPC method, is effective in analyzing normal infant crying.

This conclusion is an important experimental support of our theoretical analysis on infant crying modeling in section 4.1.

In addition, Figure 4.4 shows an example of hyperphonation. The hyperphonation occurs during 1.2–1.45 seconds, identifiable by a fundamental frequency ($F_0$) as high as 1–2 kHz. To indicate the presence of the hyperphonation, we mark the estimated formant frequencies in that period with circles in Figure 4.4 (a). During that period, the formant frequency estimates are not reliable because the harmonics of the extraordinarily high $F_0$ dominate the LPC-derived spectrum envelope and overshadow the peaks corresponding to the actual formants.

4.3 Some Technical Considerations of Automatic Cry Analysis

As stated in Chapter 1, the goal of our infant cry research is to develop automatic devices which can analyze and monitor the infants’ physical/emotional situation. In this section we will discuss some technical concerns about the design of such an automatic infant cry analyzer. In many respects, the automatic infant cry analysis problem is analogous to the problem of automatic speech recognition. We can consider cries corresponding to different physical/emotional situations as different “cry sentences” uttered by the infant, and the different time-frequency patterns in different segments of the cry signal as different “cry words”. To determine the physical/emotional situation of the infant, we therefore need to
recognize those "cry words" or "cry sentences." Similar to speech recognition systems, the following technical aspects are of interest in the design of our automatic cry analyzer.

4.3.1 Training of the Automatic Cry Analyzer

Like any pattern recognition system, proper training is the key to the success of our infant cry analyzer. Generally, training process in a pattern recognition system involves establishing a base of standard exemplars to which an unknown (unclassified) input can be compared and then classified. To establish this exemplar base, a set of classified or labeled inputs, called the training set, is needed. The identity information from each labeled input will be analyzed, extracted, and stored to form the exemplar base.

As a compromise between recognition performance and application convenience, two different training strategies are used in automatic speech recognition systems: namely speaker-dependent and speaker-independent training. The speaker-dependent strategy achieves higher accuracy but requires the system be trained on-site by the specified user(s) before using the system for recognition. The speaker-independent system is pre-trained and then applied to other speakers without on-site training. The convenience offered by the speaker-independent system is at some cost in recognition accuracy. This is because the speaker-independent system is not trained by the user's own voice. In order to achieve the required accuracy, the structure of the speaker-independent systems is usually more complicated than that of the speaker-dependent systems.

The same training problem exists in the automatic infant cry analyzer. We can similarly design the automatic cry analyzer to be infant-dependent or infant-independent. But on-site training of a cry analyzer is in most times inconvenient or even unacceptable. Therefore, a
practical and clinically useful automatic infant cry analyzer should be infant-independent. The implementation of a totally infant-independent cry analysis system may pose tremendous technical difficulties, especially if the system is to achieve acceptable accuracy. Some trade-off thus must be made. One possible solution is to introduce age-dependence in an infant-independent cry analyzer. It has been noticed by many researchers that some patterns and acoustic attributes in the cries change with the age of the infant, especially in the first week after birth [51, 97, 118]. If we designed and trained the cry analyzer for only a group of infants whose ages are within a specified range, the variation in the training data may be significantly reduced, and thus the analysis accuracy of the cry analyzer could be improved. We can also train an exemplar base for each of the different age groups of infants, and integrate all exemplar bases into the cry analyzer. In use, the operator of this cry analyzer only needs to specify the age of the infant whose cry is to be analyzed. The analyzer will then match the input cry signal to the exemplar base trained for infants of this age group.

4.3.2 Feature Selection

As we discussed earlier in Section 3.3, there is no generally applicable theory to guide the selection of effective and efficient features. In particular, feature selection is mainly conducted on a trial-and-error basis. Nevertheless, it is reasonable to get help from the published results and findings of infant cry researchers. In our case, we need to examine both the features commonly used in speech processing and those used in “traditional” cry research.
Some of the features which, we believe, are worthy of investigation are listed below:

1. **short-time energy, zero-crossing rate, LPC coefficients and prediction error**

   These are the traditional features used in speech processing/recognition research [101]. For most of these features, efficient algorithms exist to extract them from the voice signal.

2. **LPC-derived cepstrum coefficients**

   LPC-derived cepstrum coefficients are reported to yield the highest recognition accuracy in speaker recognition [88], and are also the most commonly used in recent speech recognition systems [50, 102, 104]. Cepstrum coefficients have the desirable characteristic of being insensitive to any fixed frequency-response distortions in the recording apparatus and in the transmission system [2, 88].

   A highly efficient method exists to compute the cepstrum coefficients. This is to derive them from the linear predictor coefficients by using the following recursive relationships [2],

   \[
   \begin{align*}
   c_1 &= a_1 \\
   c_n &= \sum_{k=1}^{n-1} \left( 1 - \frac{k}{n} \right) a_k c_{n-k} + a_n, \ 1 < n \leq p
   \end{align*}
   \]  

   where \( p \) is the order of predictor, and \( c_i \) and \( a_i \) are the \( i^{th} \) cepstrum coefficient, and the \( i^{th} \) linear predictor coefficient, respectively.

3. **fundamental frequency, formant frequencies, melody types, cry duration, and other attributes used in past infant cry research**

   These features, as well as their relationship to various physical/emotional situations of the infant, have been the focus of infant cry research in the past (see Chapter 2). Many

---

\( f \) — In the simplest form, the (real) cepstrum of the signal \( x(t) \) is defined as the function \( c(t) \), which is the inverse Fourier transform of \( C(\omega) \), where \( C(\omega) = \ln |X(\omega)| \).
published results seem to suggest that certain distinguishable connections exist between these cry attributes and specific physical/emotional conditions of the infant. However, to use those attributes as features in the cry analyzer, we must establish effective and efficient methods to **electronically** extract such features from the cry signal. This is not a trivial problem even for the fundamental and formant frequencies which have been carefully studied from the very beginning of speech processing research. Also, the relationships between these cry attributes and the infants' physical/emotional situations need to be reexamined since our objective is not restricted to the analysis of the cry, but involves the building of an automatic analyzer/monitor of infant cries.
Chapter 5

THE H-VALUE: A MEASURE FOR THE AUTOMATIC ASSESSMENT OF NORMAL INFANT DISTRESS-LEVELS FROM THE CRY SIGNAL

In the previous chapters we have reviewed infant cry research, and discussed the importance and feasibility of introducing modern speech and voice processing/recognition techniques into this field. We found that many problems in infant cry research remain unsolved because highly efficient and precise analytical approaches have not yet been explored. We concluded that the success of applying infant cry analysis techniques in practice will heavily depend on the use of computer-based automatic signal processing methodologies and systems. Through both theoretical analysis and experimental investigations, we have shown that the mechanism or model governing the infant cry generation is comparable both physiologically and acoustically to that of adult speech production. Thus, the methodology and technology developed for advanced voice signal processing and automatic speech recognition are likely applicable to the infant cry analysis after some modifications.

These perspectives and findings suggest that a broad range of opportunities exist for the investigation of using signal processing techniques to extract reliable information from the cry signal, and to make assessment of the physical/emotional situation of the infant based on such information. In addition, new computerized techniques may provide opportunities of transferring accumulated findings and results on infant crying into practical applications.

In the rest of this dissertation, we will demonstrate our concept of automatic infant cry analysis by studying one particular issue which has long challenged infant cry research — the
Chapter 5: A MEASURE FOR INFANT DISTRESS-LEVEL ASSESSMENT

assessment of the levels of distress (LOD) of infants from their cries [127]. In the following chapters we will first present a “cry phoneme” system which effectively represents various time-frequency patterns commonly found in normal infants’ cries. Then we shall report the design of two automatic systems which estimate the LOD of the infant from cry sounds. These systems are based on (1) statistical classification and (2) hidden Markov modeling techniques.

In this chapter, we discuss the concept of infant LOD, introduce our “cry phoneme” system, and define an effective indicator — the H-value — to measure infant LOD from the cry signal.

5.1 The Level of Distress (LOD) Ratings of Normal Infant Cries

It is an important prospect to be able to assess the infants’ physical and/or emotional situation by systematically analyzing their cries. In real life, this assessment is made by parents relying on their intuition and experience formed by everyday dealings with their infants. However, it is still far from being understood how parents assess the infants’ situation from the cry sounds, or what attributes in the cries influence parents’ assessments most. It is therefore of tremendous theoretical and practical value to investigate the process of parents’ assessment of the infants’ situation, and develop effective and systematic methods to automatically make such assessment from the infants’ cry sounds.

In the research on assessing the infants’ physical/emotional situation, one commonly used approach is to classify the cries into different cry types according to the (presumed) stimulus that caused the cry, e.g., hunger and pain. Then, based on the definitions of these cry types, investigations aimed at finding the relationship between the cry attributes and the defined cry types, or at subjectively recognizing these cry types by the caretakers are usually carried
out [9, 10, 26, 27, 118, 120]. The pain, hunger, pleasure, birth, and fussy cry are the most often used cry types [26, 27, 118]. However, this approach implies that the observed cries are a single variable function of the often uncertain and presumed stimulus. It is unlikely that such a straightforward functional relationship always exists. The sleep/waking state, differences in neurophysiological maturity, individual sensitivity, and other uncontrollable internal and external factors may also have important influence on the cry attributes. Without a thorough and proven knowledge of the complex cause-effect relationship of normal infant cry generation, it is not possible to ensure either the meaningfulness of these cry-type definitions, or the reliability of the results obtained. For this reason we choose not to follow the above approach.

We chose a different path in our research. Instead of relating the cry parameters to the possible cause of the cry, we relate them to parents' perceptions of the infants' physical/emotional situation. This involves subjective observations and judgements since after hearing a cry, parents must report their best guess as to the infants' level of distress. The parents' reports are then correlated with the different parameters of the cry. The advantages of this approach are: (1) it is independent of any presumed cry generation mechanism or models, and is therefore free of the uncertainties or risks involved in trying to relate the cry parameters to the cause of the cry, and (2) it can lead to the discovery on how experienced caretakers assess the infants' situations and what information in the cry contributes most to their decision. After identifying the parameters which show the most consistency with the parents' perceptions, it may become possible to build a device that can automatically estimate those parameters and make "humanlike" assessments of normal infant cries.

To evaluate how adult caretakers assess the infants' situation from their cries, several
different systems of ratings have been investigated. Notably are the aversiveness rating items originally used by Zeskind and Lester [130] and the more traditional semantic differential items [87]. The eight aversiveness rating items are urgent, distressing, sick, arousing, grating, discomforting, piercing, and aversive. The semantic differential items include unpleasant, sharp, rugged, awful, fast, heavy, bad, active, and hard. Gustafson and Green [39] cross-checked all these 17 rating items, and found virtually identical results for all ratings. They concluded that, "... at least when applied to the cries of normal infants, the individual items typically used to assess cry perception do not have discrete meanings, instead they may be measures of a single underlying dimension". It was also concluded by Murray [82] that rather than differences of cry types based upon their causes, cries differ along a continuum of intensity. Murray also observed that it is the severity of distress that observers largely discern. Therefore, we decided to use a single item — level-of-distress (LOD) — to describe the perceptual assessment of the normal infants' emotional situation during the cry.

5.2 Definitions of "Cry Phonemes" for Normal Infant Cries

In a fashion analogous to the phoneme definitions in linguistics, we introduce here the concept of "cry phonemes". We use ten "cry phonemes" to encode normal infant cries into cry phoneme sequences, or cry phoneme scripts. These scripts will be later used to calculate a single measure, the H-value.

These ten "cry phonemes" are chosen to satisfy two criteria. The first is that the "cry phonemes" should constitute a basis for normal infant cry signals in the sense that together they cover most time-frequency patterns or variations commonly found in normal infant cries. Thus, any cry from normal infants could be divided into a sequence, of which each member
belongs to one of these cry phonemes. The second criterion is that these phonemes should be detectable and distinguishable using computer signal analysis.

First we describe our definitions of the ten "cry phonemes" as they are observed as different time-frequency patterns.

1. *Trailing (Glottal roll)* — Usually occurs at the end of a long and powerful expiratory phonation. It is characterized by a) a very low, gradually decreasing, and vibrating fundamental frequency $F_0$, and b) a gradually decreasing total energy level.
2. *Flat* — The basic expiratory phonation. Characterized by a) a smooth and steady $F_0$, b) clearly observable harmonics, and c) little energy distribution in between the harmonics.
3. *Falling* — Similar to the flat except for a descending $F_0$.
4. *Double Harmonic break* — A simultaneous parallel series of harmonics in between the harmonics of $F_0$. The in-between harmonics occur suddenly and are usually weaker than the primary ones.
5. *Dysphonation* — Characterized by a) an unstructured energy distribution over all the frequency range, sometimes with a tendency of higher concentration over the middle to high (1–5 kHz) frequency range, or b) an unstructured energy distribution imposing on or in between the barely distinguishable harmonics.
6. *Rising* — Similar to flat except for an ascending $F_0$.
7. *Hyperphonation* — Phonation with an extraordinarily high $F_0$ (often over 1 kHz).
8. *Inhalation* — The sound produced by the infant’s rapid breathing in of air. Usually occurs after an exhaustive expiratory phase.
9. *Vibration* — Characterized by a) clearly observable harmonics but with a vibrating $F_0$, b) no unstructured energy distribution in between harmonics, and c) a normally high total
energy level.

10. **Weak Vibration** — Similar to the vibration except that the total energy level is significantly lower than normal level.

Figure 5.1 shows the spectrograms of the 10 “cry phonemes”. Of each spectrogram, the frequency ranges from 0 Hz at the bottom to 4 kHz at the top.

The above choice of the ten “cry phonemes” are conceptually based on the cry generation models suggested by Truby et al. in 1965 [114] and Golub et al. in 1985 [35]. However, our choice is also governed by the two criteria stated at the beginning of the present section.

In particular, as mentioned in Subsection 4.1.1.2, three different cry modes are suggested by Golub and Corwin [35]. They are: phonation, hyperphonation, and dysphonation. To lay the basis for reliable computer-aided analysis, in our “cry phoneme” definitions we further divide the phonation cry type into five sub-types, namely flat, rising, falling, vibration, and weak vibration, according to the melody types and/or the short-time energy level of the signal. We also define “cry phonemes” for trailing (or glottal roll) and double harmonic break, because of their distinctive time-frequency patterns and their frequent occurrence in normal infant cries. There is no direct explanation of these two phenomena in the cry generation model in [35]. The same applies to the sound created during the inspiratory phases, which we define as a separate cry phoneme: *inhalaion*.

Figure 5.2 shows an example of labeling a three second segment of a pain-elicited cry with the above-defined “cry phonemes.” On the top of the spectrogram in Figure 5.2, which is generated by our specifically designed graphics software, the segmentation of the signal is shown with the number indicating the mode or type of the “cry phoneme” during the interval. The curve at the bottom shows the short-time energy of the cry. There are basically two
Figure 5.1 The time-frequency patterns of the 10 "cry phonemes" shown in the computer-derived spectrograms. They are: (1) trailing, (2) flat, (3) falling, (4) double harmonic break, (5) dysphonation, (6) rising, (7) hyperphonation, (8) inhalation, (9) vibration, and (10) weak vibration, respectively.

expiratory and one inspiratory vocalizations in this cry signal. The first expiratory phase
5.3 The Relationship between “Cry Phonemes” and the Parents’ LOD Ratings

To determine the relationship between each of our “cry phonemes” and the parents’ assessment of the LODs of the cries, we conduct some experiments. We also deduce a single measure — the H-value — to quantify the LOD of the infant.
Chapter 5: A MEASURE FOR INFANT DISTRESS-LEVEL ASSESSMENT

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>S.D.</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (min)</td>
<td>130.25</td>
<td>40.00</td>
<td>57-253</td>
</tr>
<tr>
<td>Birthweight (g)</td>
<td>3540.83</td>
<td>532.26</td>
<td>2540-4900</td>
</tr>
<tr>
<td>Birth length (cm)</td>
<td>51.25</td>
<td>1.83</td>
<td>48-55</td>
</tr>
<tr>
<td>Gestation (weeks)</td>
<td>38.94</td>
<td>1.01</td>
<td>37-41</td>
</tr>
<tr>
<td>Apgar at 5 min</td>
<td>9.28</td>
<td>0.57</td>
<td>8-10</td>
</tr>
</tbody>
</table>

Table 5.1 Characteristics of the 36 infants.

5.3.1 Experiments

Subjects

The subjects include 36 newborn infants recruited from the labor and delivery unit of a major metropolitan maternity hospital.\(^8\) Criteria for inclusion were spontaneous vaginal delivery or planned caesarian section, gestational age 37–42 weeks, birth weight above 2500 g, 5 min Apgar of 8, 9, or 10, and infant judged clinically healthy (normal). Table 5.1 shows the characteristics of the infants [38].

Procedure

The data collection was conducted in a quiet room near the nursery. A color camera with a unidirectional microphone was used to record both the facial activities and the vocalizations of the infant with 3/4” videotape (the recorded video signal was not used in our study). During the recording the microphone was suspended 8 inches from the infant’s mouth; a distance that was maintained for all infants.

Each infant responded by crying to the following three procedures: injection of vitamin K into the thigh, application of triple dye disinfectant solution to the umbilical cord, and

---

\(^8\) The infant recruitment, criteria setting, and data collection was conducted by Dr. R. V. E. Grunau et al. of the B.C. Children’s Hospital and University of British Columbia, and was supported by a grant to Dr. K. D. Craig from NSERC. For more details on data collection, see [38].
swabbing the thigh opposite the one receiving injection. The order of the procedures was randomly set for each infant.

The cry recordings were first transferred onto 2-track audiocassettes and then digitized with a sampling rate of 8 kHz after lowpass-filtered at 4 kHz. A 12 bit resolution was used for the digitization.

After deleting some noise-corrupted samples and splitting those long cry samples which were often interrupted by periods of silence, we finally collected 103 cries from the recording. Amongst these 103 cries, 58 were randomly chosen as our testing sample set and were used in this experiment. The remaining 45 cries form our training set that were later used in designing our cry analyzers.

Then, each of the 58 cries in the testing set was divided and labeled into a cry phoneme sequence as discussed in the previous section. This was accomplished with the aid of our computer graphics program which provides a convenient work space consisting of a high resolution spectrogram and an energy curve of the signal (as seen in Figure 5.2). To compute the spectrogram and the short-time energy, the signal was cut into frames by a 32-msec-wide sliding Hamming window which advanced in steps of 10 msec each. Then, for each cry we calculated the percentage of the total time of each cry phoneme out of the entire vocalization time of the cry.

The same cries were later evaluated by 20 parents (10 married couples with at least one young child each). These parents were recruited from the graduate student population of our university. The parents were instructed to repeatedly listen to the recordings of the 58 infant cries, and to report their assessments of the infants' level of distress. They were asked to rate each cry into one of five LOD grades. Grade 1 to grade 5 in the rating represents an increasing
degree of LOD from the least to the most. To prevent preconceptions, the precise causes of
the cries were not disclosed to the parents prior to (and during) conducting the experiment.

5.3.2 Results

As mentioned above, each of the 58 cries was segmented and labeled into sequences of
“cry phonemes”, with the aid of our computer graphics tool. Also, the cries were separately
assigned by the parents to one of the five LOD grades. Then, to find the correlation between
the “cry phonemes” and the parents’ LOD ratings, scatter graphs in Figure 5.3 were computed.
Each scatter graph contains 58 dots corresponding to the 58 cries and shows the relationship
between a specific “cry phoneme” and the parents’ LOD ratings. The abscissa in a scatter
graph is the percentage duration of the respective cry phoneme in each cry sample. Each
cry in a scatter graph is represented by a line with a dot in its middle. The dot indicates the
mean LOD assigned by the parents and the line shows the standard deviation of the ratings
given to that cry. Please note that a dot in the vertical axis of a scatter graph indicates that
the corresponding “cry phoneme” is not present in the cry signal represented by the dot.

Amongst the 10 “cry phonemes”, the dysphonation, hyperphonation, and inhalation show
strong positive relationship with the parents’ LOD ratings, and the flat and weak vibration
show strong negative relationship. Also, the trailing and double harmonic break show weaker
yet still clear positive relationship. No strong correlation is observed with the falling, rising,
and vibration.

The dysphonation “cry phoneme” shows the most consistent positive relation with the
parents’ ratings. From our experiment data, all the cries containing more than 10% of
Figure 5.3 The relationship of parents’ LOD ratings to the percentage occurrences of different “cry phonemes”. (Continued . . . )
Chapter 5: A MEASURE FOR INFANT DISTRESS-LEVEL ASSESSMENT

Figure 5.3 The relationship of parents' LOD ratings to the percentage occurrences of different "cry phonemes". (Continued . . . )
Chapter 5: A MEASURE FOR INFANT DISTRESS-LEVEL ASSESSMENT

Figure 5.3 The relationship of parents' LOD ratings to the percentage occurrences of different “cry phonemes”.

(Continued...)

Figure 5.3 The relationship of parents’ LOD ratings to the percentage occurrences of different “cry phonemes”. (Continued...
Figure 5.3 The relationship of parents' LOD ratings to the percentage occurrences of different "cry phonemes". (Continued . . . )
Figure 5.3 The relationship of parents' LOD ratings to the percentage occurrences of different "cry phonemes".
dysphonation were given a mean rating between the middle to the highest LOD. This is consistent with the findings of Gustafson and Green [39] who analyzed the cry on the basis of single expiratory phases, and also agrees with the definition and descriptions about dysphonation in Truby and Lind [114].

5.4 The Introduction of the H-value and Its Application in Quantifying LOD Assessments

For expediency and efficiency, in the study of the LOD rating, we wish to characterize the 10 “cry phonemes” by a single composite parameter — the H-value. The H-value of a “cry phoneme” sequence is defined in terms of the phonemes showing positive correlations with the parents’ LOD ratings. We propose that the H-value be defined as the percentage time duration of the combination of the following five “cry phonemes” to the whole time duration of the vocalization of the cry: trailing, double harmonic break, dysphonation, hyperphonation, and inhalation, i.e.,

\[ H = \left( \frac{D_T + D_B + D_D + D_H + D_I}{D_{total}} \right) \times 100 \text{ (\%)} \]  (5.1)

where \( D_T, D_B, D_D, D_H, \) and \( D_I \) are the durations of the above five “cry phonemes”, respectively, and \( D_{total} \) is the total time of vocalization in the whole cry. Figure 5.4 shows the correlation between the H-values of the 58 cries and the parents’ LOD ratings.

It is evident from Figure 5.4 that the H-value shows a very high consistency with the parents’ LOD ratings, i.e. high H-values always correspond to high LOD ratings of the parents while low H-values always correspond to low LOD ratings. This makes it possible to use the H-value as a reliable measure for evaluating the LOD of the normal infant. An automated cry analyzer may thus calculate the H-value of the cry under consideration and then give an estimation of the LOD of the cry.
5.5 Discussions and Remarks

We follow the approach of correlating the cry attributes to subjective observations by parents, or experienced caretakers. At present, we believe this is a more feasible approach than that of classifying the cry into types corresponding to the different stimuli that may have caused the cry. The latter approach will become more productive after we gain a thorough and proven understanding of the cause-result relationship in the infants’ cry generation process.

In this chapter, we introduced (1) the use of the “cry phonemes” and the “cry phoneme” sequences or scripts to characterize the normal infant cries, and (2) from these “cry phoneme” scripts we extract a single parameter — the H-value — to evaluate the underlying LOD of the infant during the cry. In our experiment with 36 newborn infants and 20 parents, we found
that the H-values derived from the cry signals show clear correlation with the parents' ratings on the LOD of the infants. The next challenge is to find effective and efficient methods to automatically estimate the H-value from the cry signal. Then it will become feasible to build reliable and robust devices which can make humanlike assessments of the infants' LOD by automatically analyzing their cries. In the following two chapters, we describe two such methods, respectively — one of which is based on nonparametric statistical classifier, and the other on the hidden-Markov-model technique.
Chapter 6

CRY ANALYZER I: A NONPARAMETRIC
STATISTICAL CLASSIFIER-BASED METHOD

In this chapter, first we construct a cry analysis system based upon the traditional statistical classifier. Then, we evaluate the performance of this cry analysis system using cry samples from normal infants, by comparing the estimated values of the infants' LOD which this system yields with the LOD reported by parents who listened to the same cries.

The idea behind the design of this cry analyzer is to use statistical classification methods to automatically identify the 10 different "cry phonemes" of the cry signal defined in Chapter 5. Then we calculate the H-value of the cry by using the definition given in Section 5.4. The H-value serves as an estimation of the infants' LOD inferred by the cry, as discussed in the previous chapter.

6.1 A Survey of Traditional Statistical Classification Methods and Their Limitations

Theoretically, a classifier assigns an object whose value is observed or known, but whose class is unknown, to one of many predetermined classes or groups. Traditionally, statistical classifiers are based on Bayes' Theorem (see Appendix C). The Bayes Theorem is stated as follows,

\[
\text{assign the object to group } \omega_k \text{ if } \\
p(x|\omega_k)P(\omega_k) \geq p(x|\omega_j)P(\omega_j) \text{ for all } j \neq k, \quad (6.1)
\]
where $\omega_k$'s ($k = 1, 2, \cdots, N$) are $N$ different groups to which the unknown observation $x$ is to be assigned, and $p(x|\omega_k)$ is the conditional probability density function of observation $x$ given group $\omega_k$, and $P(\omega_k)$ is the a priori probability of group $\omega_k$. A more detailed discussion about Bayes' Theorem is found in Appendix C.

To build a classifier, it is thus important to determine or estimate the conditional probability density function (pdf) $p(x|\omega_k)$ for each group $\omega_k$, $k = 1, 2, \cdots, N$. In some cases, we can analytically deduce these pdfs from some a priori knowledge about the problem. However, in most practical situations, the pdfs can only be estimated from a set of samples with known group assignments. This approach of estimating pdfs is very commonly used, and is referred to as the training of the classifier.

The pdfs estimating procedure, as well as the resultant classifiers, are divided into two different classes — parametric and nonparametric — depending on whether or not the function form of the pdfs is predetermined prior to the training. For parametric classifiers, such as the linear classifiers and the quadratic classifiers [22, 40], the function form of the pdfs is predetermined. Hence we only need to estimate the parameters of the pdfs.

On the other hand, nonparametric classifiers are built without any assumption about the form of the pdfs function. This makes the nonparametric classification one of the most useful approaches in statistical pattern recognition [22, 40]. When dealing with problems with unknown or complicated distributions, nonparametric classifiers frequently show much higher classification accuracy than those achieved by parametric classification approaches. Even when the data are from normally distributed populations, the nonparametric algorithms sometimes outperform their parametric counterparts [85, 86].

Generally, when we do not possess enough knowledge to allow us to make reasonable
assumption about the functional form of the pdfs, we must give the nonparametric methods the first consideration. This is exactly the case in our infant cry analysis problem, where little or no knowledge is available about the distributions of the features of infant cries in either the temporal or the spectral domains. We, therefore, decided to use the nonparametric classification approach to build our automatic infant cry analyzer.

For all its advantages and power, the application of nonparametric classifiers also suffers from various difficulties, especially as the size of the problem increases.

The common disadvantages of classical nonparametric approaches (kernel estimator, kNN classifier, etc.) are their computational complexity and their requirement of large amounts of computer storage to hold the design sets. Large design sets are always desirable when nonparametric approaches are used. This is because the feature vectors in the problems requiring the use of nonparametric methods usually possess very complex or unknown probability distributions. This means that a large sample set is required to obtain sufficient statistical information. Because of these difficulties, nonparametric classification methods are usually complex and slow, and their on-line application is rare. Their uses are often limited to situations where the computational time is not a crucial factor, such as in the estimation of the Bayes error and data structure analysis [22].

A solution to the above problems is to reduce the size of the design set while insisting that the classifiers built upon the reduced design set perform as well, or nearly as well as the classifiers built upon the original design set. This idea has been explored for various purposes over a period of time and resulted in the development of many algorithms for the kNN classifier design using reduced design sets. Examples of this type of classifier are the condensed NN (CNN) [42], the reduced NN (RNN) [29], and the edited NN (ENN) [13].
these algorithms, iterative processes are used to test the effect on the classification performance of each individual vector in the design set as the vector is moved in and out of the design set. Then only the "pertinent" vectors are retained in the reduced design set. For very large design sets, these methods are often tedious and difficult to implement since a new classifier is in fact built and evaluated every time a vector is moved in or out of the design set. The most serious disadvantage is that the resulting reduction rate is usually low (i.e., not as good as hoped). Also the reduction rate is usually not under the control of the algorithms, but depends entirely on the nature of the design set to be reduced.

Recently, two nonparametric data reduction algorithms for the Parzen's kernel classifier and the NN classifier design have been proposed by Fukunaga et al. [22–24]. Their algorithms find the optimal reduced design set in the sense that the difference between the estimated probability density functions of the reduced set and the original set is minimized. Bearing some similarities to the traditional reduced data kNN algorithms, their algorithms iteratively move each individual vector in and out of a tentatively chosen reduced design set and test the resultant effect on the criterion function. To avoid an exhaustive search of all possible subsets, the optimization scheme used in Fukunaga's algorithms achieves a local optimum. The computational complexities of these algorithms are considerable. Moreover, the initial guess of the reduced sample set is of crucial importance in Fukunaga's reduced NN algorithm. So far, an intuitively developed initial assignment procedure for only the 2:1 reduction rate case has been published [24].

6.2 The Development of the VQ-based Nonparametric Classification Approach

We introduce a new approach for nonparametric data reduction using the vector, or block,
Chapter 6: CRY ANALYZER I: STATISTICAL METHOD

quantization technique. In a recent publication "Vector Quantization Technique for Nonparametric Classifier Design", we have already demonstrated the effectiveness of this approach [126].

While our VQ-based method is developed for the purpose of realizing efficient automatic infant cry analysis, it should be noted that the method is, in fact, a general purpose nonparametric classification method. Hence it is applicable to other statistical pattern recognition problems. Therefore, in this section we will present, investigate, and discuss such properties as the accuracy, speed and computational complexity of this new method within the broader context of statistical pattern recognition. Then in Section 6.3 we will apply this approach to the design of an automatic cry analyzer system.

Vector quantization is a mathematical process which has already been widely used in various areas of engineering, such as digital signal processing, communication, and speech recognition, as was mentioned in section 3.5. In particular, VQ has been used as an effective data and image compression technique. Our research indicates that this technique can also be effectively utilized to perform the desired data reduction in the design of nonparametric classifiers. Combining vector quantization with each of the classical Parzen's kernel and the \( k \)NN approaches, we develop two new algorithms of reduced nonparametric classifier design, which we shall call the VQ-kernel and the VQ-\( k \)NN methods.

In these new algorithms, we first construct an optimal vector quantizer for each class of the training data. For each class, its corresponding original design set is used as the training sequence to determine the optimal reproduction alphabet of its vector quantizer. This reproduction alphabet, or code-book, is then retained as the reduced design set to represent the original set of this class. Using the reduced sets, a classifier utilizing either the Parzen's
kernel or the \( kNN \) methods is then built.

For our purposes, we find that the reproduction alphabet serves as a good representative of the original design set. The obtainable reduction rate can be significantly high and may be preset freely in the algorithms. As illustrated by the examples given later, large original design sets containing hundreds of vectors can be well represented by a reproduction alphabet with only a few vectors. Despite the high reduction rate, our VQ-based classifiers perform as well or better than classical reduction methods in terms of classification error rate. This remarkable classification performance probably results from the fact that our VQ-based method does not restrict the search for the members of the representative (reduced) design set to vectors in the original design set. Instead, our VQ-based method creates the representative set which retains information from all the vectors in the original set. This separates our VQ-based method from the traditional methods which always restrict a representative vector to belong to one of the original design set vectors.

In our VQ-based methods, we use the algorithm developed by Linde et al. [61] to design the vector quantizer. This algorithm has the advantage that it is well developed and widely applied in various fields, and its efficiency in implementation and its convergence properties are proven. A summary of Linde’s algorithm [61] is found in Appendix B.

As an alternative, the learning vector quantization (LVQ) algorithm, which was proposed recently by Kohonen [46, 47], may be used to generate the quantizer. The overall performance of the LVQ technique has been shown to be comparative to that of Linde’s VQ algorithm in image compression applications [84].
6.2.1 Development of the Algorithms

Since our new algorithms use vector quantization as the first stage, it is helpful to repeat briefly some of the definitions of vector quantization presented in Section 3.5. Then in Subsection 6.2.1.2 we will discuss an important property of vector quantization. Our algorithms are then described in Subsection 6.2.1.3.

6.2.1.1 Definition of vector quantization

An $M$-level $d$-dimension quantizer is a mapping, $q(\cdot)$, that assigns to each input vector, $x = (x_0, \cdots, x_{d-1})$, a reproduction vector, $y_k = q(x)$, drawn from a finite reproduction alphabet (or code-book), $\hat{A} = \{y_i; i = 1, \cdots, M\}$. The $q(\cdot)$ is usually chosen as a minimum-distance mapping. This means that the reproduction vector $y_k$ is chosen such that

$$d(x, y_k) = \min_{i} d(x, y_i)$$

(6.2)

where $d(\cdot, \cdot)$ is any nonnegative distance measure defined on the $d$-dimension space, and $d(x, y_k)$ is called the distortion of the quantization. Obviously, a minimum-distance mapping, which is completely defined by the reproduction alphabet $\hat{A}$ along with the definition of the distance measure, uniquely describes a Dirichlet partition, $S = \{S_i; i = 1, \cdots, M\}$, of the sample space.

An $M$-level quantizer is said to be optimal if it minimizes the expected distortion $D(q) = E\{d(x, q(x))\}$, that is, $q^*$ is optimal if for any other quantizer $q$ having $M$ reproduction vectors, $D(q^*) \leq D(q)$. 

6.2.1.2 Distribution Property of the Reproduction Vectors
and the VQ Data Reduction Method

In the following, we argue that the distribution of the reproduction vectors in an optimal vector quantizer possesses desirable properties that make the VQ technique a promising approach for nonparametric data reduction. Before doing so, however, it is helpful to point out a generalization concerning the nonparametric classifier design.

The philosophy guiding the development of most traditional nonparametric classification methods is as follows: the statistical information contained in a set of pre-classified samples (or design set), \( \omega_k \), is used for finding a good approximation of the actual underlying probability density function, \( p(x|\omega_k) \). Then the classifier is built by applying the Bayesian rule (see section 6.1). This philosophy is also explicitly employed in the development of Fukunaga’s reduced Parzen and NN classifiers. In these classifiers, as mentioned earlier, the reduced design set is selected in such a way that the difference between the density estimate \( \hat{p}(x|\omega_k) \) obtained from this reduced set and that obtained from the original design set is minimized. However, for achieving high classification performance, this approximation to \( p(x|\omega_k) \), while it is obviously sufficient, is not necessary. An illustration of this is that any good approximation to \( [p(x|\omega_k)]^\alpha \), where constant \( \alpha > 0 \), will yield the same Bayesian classifier as that achieved by approximating \( p(x|\omega_k) \) itself.

In [30] Gersho addresses the properties of the reproduction vectors of an optimal vector quantizer. The density function of the reproduction vector in a \( d \)-dimensional quantizer is defined as

\[
g_M(x) = \frac{1}{MV(S_i)}, \quad \text{if } x \in S_i, \quad \text{for } i = 1, 2, \ldots, M. \tag{6.3}
\]
where $V(S_i)$ is the volume of $S_i$. Gersho shows that for an optimal quantizer, in the asymptotic situation where $M$ is sufficiently large, $g_M(x)$ will closely approximate a continuous density function $\lambda(x)$ which is proportional to $[f(x)]^{1/1+\beta}$, where $f(x)$ is the actual underlying density function of the input random vector and $\beta$ is a constant determined by the dimension $d$ and the distance measure. Therefore, if we generate the optimal quantizer by using the training vectors from class $\omega_k$, the density function of the resulting reproduction alphabet (of class $\omega_k$) will closely approximate a function that is proportional to $[p(x|\omega_k)]^{1/1+\beta}$.

This finding, along with the general consideration discussed at the beginning of this section, strongly indicates that the reproduction alphabet in an optimal quantizer could be used as an effective design set for building classifiers, provided the level of the quantizer $M$ is sufficiently high.

Using VQ algorithms such as Linde's and Kohonen's, we can estimate an either locally or globally optimal reproduction alphabet from the training data set. This leads us to the following data reduction approach for the nonparametric classifier design: using the original design set as the training data, find the optimal reproduction alphabet of this set and use it as the reduced design set. In the following section we describe our data reduction methods in full detail.

The most important difference between our data reduction method and the other traditional methods is that our reduced set is not necessarily a subset of the original design set. This is because in our VQ-based method, the vectors in the reduced set are created (but not selected) by surveying all the information carried by every known vector. These vectors are created so that they best represent all the vectors in the original set in the sense of minimizing the average distortion rate of quantization.
6.2.1.3 Classifier Design Algorithms Using Vector Quantization

The VQ-kernel Classifier

In this method, we propose that vector quantization be first applied to each of the original design sets of all classes. The reproduction alphabets of the resultant optimal quantizers are then retained as the reduced design sets. Then the kernel method is applied as usual except that the reduced design sets are used.

The following steps explain this method in greater detail.

Considering an N-class problem in d dimension, assume that for each class \( \omega_i \) \((i = 1, \cdots, N)\) an original design set \( \{x_j^{(i)}; j = 0, \cdots , n_i - 1\} \) is given.

1. For class \( \omega_i \), as in the algorithm in [61] we start with an initial \( M_i \)-level reproduction alphabet \( \hat{A}_{0i} \). This initial alphabet can be generated by using the “splitting” approach proposed in [61].

2. Find the \( M_i \)-level optimal quantizer for class \( \omega_i \) using the quantizer design method in [61]. Suppose the final reproduction alphabet for class \( \omega_i \) is \( \hat{A}_i = \{y_j^{(i)}; j = 1, \cdots , M_i\} \).

3. Apply the Parzen’s kernel method to the reproduction alphabet \( \hat{A}_i \) to get an estimate of \( p(x|\omega_i) \), i.e.

\[
\hat{p}(x|\omega_i) = \frac{1}{M_i} \sum_{j=1}^{M_i} K_i(x - y_j^{(i)})
\]  

(6.4)

where \( K_i \) represents the kernel function of class \( \omega_i \).

4. Repeat step 1 to step 3 for each class and get estimates of \( \hat{p}(x|\omega_i) \) for \( i = 1, \cdots, N \).

5. Finally a Bayes classifier is built upon all the estimated class-conditional pdfs \( \hat{p}(x|\omega_i) \).

The Bayes classifier assigns the unknown observation \( x \) to class \( \omega_m \) if \( p(x|\omega_m)P(\omega_m) \geq p(x|\omega_l)P(\omega_l) \) for all \( l \neq m \), where \( P(\omega_l) \) is the known a priori probability of class \( \omega_l \).
For the special case when the levels of quantizers are chosen equal to the number of vectors in the original design sets, i.e. $M_i = n_i$ ($i = 1, \cdots, N$), the above VQ-kernel method becomes equivalent to the traditional kernel method.

It should be noted that the selection of the kernel function as well as its parameters is important, yet difficult. This disadvantage is inherent to the classical kernel approach. Theories and conclusions developed in the literature on the classical kernel methods can be directly used to guide the selection of proper kernels and smoothing factors. A thorough discussion on this topic is found in [22, 41].

**The VQ-kNN Classifier**

This algorithm combines the VQ technique and the $k$NN method.

For an $N$-class problem, we assume that for each class $\omega_i$ ($i = 1, \cdots, N$), a design set $\{x_j^{(i)}; j = 0, \cdots, n_i - 1\}$, and an initial $M_i$-level reproduction alphabet $A_0i$ are given (for the selection of $A_0i$ see [61]).

1. For each class $\omega_i$ ($i = 1, \cdots, N$), find the $M_i$-level optimal quantizer using the algorithm in [61]. Suppose the final reproduction alphabets are $A_i = \{y_j^{(i)}; j = 1, \cdots, M_i\}$ ($i = 1, \cdots, N$).

2. Combine all the reproduction alphabets $A_i$'s of all classes into one single set $\hat{A}$ of $M$ vectors. That is, $\hat{A} = \bigcup_i A_i$ and $M = \sum_i M_i$.

3. The classification rule is as follows: assume $x$ is the new observation point to be classified,

   a) For $x$, find the first $k$ nearest reproduction vectors in $\hat{A}$. Suppose that amongst these $k$ reproduction vectors there are $k_m$ vectors from class $\omega_m$. Then the classification rule is:
assign \( x \) to class \( \omega_m \) if
\[
P(\omega_m) \frac{k_m}{M_m} \geq P(\omega_l) \frac{k_l}{M_l} \quad \text{for all } l \neq m.
\] (6.5)

b) If the *a priori* probability \( P(\omega_i) \) is unknown, we can approximate it from the design set as
\[
\hat{P}(\omega_i) = \frac{n_i}{\sum_{l=1}^{N} n_l}.
\] (6.6)

The classification rule, then, becomes:
assign \( x \) to class \( \omega_m \) if
\[
\frac{n_m k_m}{M_m} \geq \frac{n_l k_l}{M_l} \quad \text{for all } l \neq m.
\] (6.7)

The classical \( k \)NN method becomes a special case of our new method when \( M_i \) is chosen to be equal to \( n_i \) \((i = 1, \cdots, N)\). If \( k = 1 \) is used, we get the VQ-NN classifier, which is very effective in computation since, when all the *a priori* probabilities are equal, the unknown observation is simply assigned to the class to which its nearest reproduction vector belongs.

### 6.2.2 Experiments of the Application of Our VQ-based Classifiers

We give now two examples to demonstrate the use of our VQ-kernel and VQ-NN classifiers. The data used in the first example has a known probability distribution, in particular a mixture of Gaussian distributions. In the other example, we test our VQ classifiers with real speech data of unknown probability distribution. For the reason of comparison, we also test the performances of traditional reduction algorithms including the CNN, RNN, ENN, as well as Fukunaga’s reduced Parzen method.
6.2.2.1 Example 1 (synthetic data from Gaussian mixture)

We adopt Fukunaga's 8-dimension data model from [22] (page 555, Experiment 11-7). In this model, each of the two classes consists of two Gaussian distributions. The \textit{a priori} probabilities are equal for the two classes. The class-conditional pdfs are \( p(x|\omega_1) = 0.5N(\mu_1, I) + 0.5N(\mu_2, I) \) and \( p(x|\omega_2) = 0.5N(\mu_3, I) + 0.5N(\mu_4, I) \) respectively, where the covariance matrices are equal to the identity matrix \( I \), and the mean vectors are \( \mu_1 = [0 \ 0 \cdots 0]^T \), \( \mu_2 = [6.58 \ 0 \cdots 0]^T \), \( \mu_3 = [3.29 \ 0 \cdots 0]^T \), and \( \mu_4 = [9.87 \ 0 \cdots 0]^T \), respectively. The Bayes error of this data model is \( \varepsilon = 7.5\% \). The Euclidean metric system was used as the distance measure. The original design set contains 150 vectors for each class, and similarly the testing set has 150 vectors from each class.

Fig. 6.1 shows the result of our experiment in which we tested our VQ-kernel and VQ-NN classifiers as well as other traditional reduced data classifiers, including the CNN [42], the RNN [29], and the ENN and the edited-condensed nearest neighbor (EC-NN) [13]. Fukunaga \textit{et al.} reported the performance of their reduced Parzen classifier with the same data model in [22]. We include their results in the graph as well. The same normal kernel function, with a constant covariance matrix of \( 1.5^2 \times I \), which was used in Fukunaga's reduced Parzen classifiers, was also used in our VQ-kernel method. Each point in the graph represents the average result after 10 trials.

In both Fukunaga's and our methods, all 1-level reduced data classifiers failed due to the fact that the underlying probability density of each class is composed of two separate Gaussian distributions. When 2 or more representatives are used in the reduced sets, both our new VQ-based classifiers showed excellent performance. In particular, our VQ-kernel classifier achieved the best performance amongst all the classifiers. It gave a classification
accuracy extremely close to the Bayes error, the theoretical lowest bound, at almost all the reduction rates, and its performance showed little correlation to the reduction rate.

Our VQ-NN classifier evidently also outperformed all other modified NN classifiers in terms of both the reduction rate and the classification accuracy. This is with the exception of the ENN which showed a better accuracy but at a low reduction rate of only 1.67:1. It is very interesting to notice that at high reduction rates ($M<40$), our VQ-NN classifier showed
significantly better performance than it did at the lower reduction rates. We believe that this is due to the specific structure of the distribution density functions (mixtures of two Gaussian distributions) underlying the experimental data.

6.2.2.2 Example 2 (speech data with unknown distribution)

This example demonstrates the classification performance of our new VQ-based classifiers with real data extracted from speech signals.

The 12-dimension speech data measured the first 10 cepstrum coefficients, the short-time zero-crossing rate, and the short-time energy of the voice signals.

The data for class 1 (2247 vectors in total) came from male speaker 1 (Chuck), and the data in class 2 (2256 vectors in total) were for male speaker 2 (Gray). For each trial in the experiments, we first randomly drew a 500 vector design set from each class, and used the rest to form the test set.

For our VQ-kernel classifier and Fukunaga’s reduced Parzen classifier, we used the Euclidean metric system and normal kernel functions with their covariance matrices estimated from the original design sets, that is, the kernel function in equation (6.4) becomes

\[ K_i(x - y) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp \left[ -\frac{1}{2}(x - y)'\Sigma_i^{-1}(x - y) \right], \quad i = 1, 2 \]  

(6.8)

with \( \Sigma_i = C_i \), where \( C_i \) is the covariance matrix of the original design set of class \( i \). For our VQ-NN and the traditional reduced data NN classifiers, the variance-weighted distance measure was used, i.e.

\[ d(x, y) = \sum_l |x_l - y_l|^2 / \sigma_l, \]  

(6.9)
where $x_l$ and $y_l$ are the $l^{th}$ element of the $x$ and $y$, respectively, and $\sigma_l$ is the variance of the $l^{th}$ element of the training vectors.

Fig. 6.2 shows the result of our experiments. The best performances were produced by our VQ-kernel classifier at all the reduction rates. The flatness of the curve shows that the classification accuracy of our VQ-kernel classifier is almost independent of the reduction rate. Comparing to Fukunaga’s reduced Parzen classifier, the improvement on the classification accuracy of our VQ-kernel classifier is obvious, especially at the high reduction rates (> 25 : 1, or $M < 20$).

Figure 6.2 The classification error rates of our VQ-based classifiers, the traditional reduced data NN classifiers, and Fukunaga’s reduced Parzen classifier for the speech data.

When operating at the same range of reduction rate (for $M \geq 16$), our VQ-NN classifier significantly outperformed the other reduced-data NN classifiers. When the reduction rate
used in our VQ-NN classifier becomes very high (62.5 : 1 to 500 : 1, i.e., $1 \leq M \leq 8$), our VQ-NN classifier still achieved classification accuracies comparable to those of the other reduced-data NN classifiers operating at much lower reduction rates, ranging from 1.75:1 for the ENN to 23:1 for the EC-NN. Please note that, for the other reduced-data NN classifiers (CNN, RNN, ENN, and EC-NN) the reduction rate is not selectable, but solely depends on the distribution of original design sets. Only our VQ-NN classifier has the capacity of choosing different reduction rates.

In addition to the classification accuracy and the reduction rate, the computational demands of Fukunaga’s algorithms are much larger than that of our VQ-kernel algorithm. The computational complexity of finding the reduced set in Fukunaga’s reduced Parzen algorithm can be shown to be of the order of $rN^2k^2$ while that of our VQ-kernel algorithm is of the order of $rNk$, where $r$ is the size of the reduced set, $N$ is the size of the original set, and $k$ is the dimension of the data. In our above speech data experiment, a single trial of our VQ-kernel classifier took only about 3.92 minutes of CPU time, while a trial of Fukunaga’s reduced Parzen classifier for exactly the same data and on the same SUN Sparc 2 computer took over 56 hours of CPU time! Amongst the NN-based algorithms, our VQ-NN classifier was also found to be the fastest. Table 6.1 shows the CPU time used in a single trial by each of the tested algorithms for constructing the reduced design sets in the above speech data experiment. The CPU time for our VQ-NN was measured with $M=128$.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>VQ-NN</th>
<th>CNN</th>
<th>ENN</th>
<th>ECNN</th>
<th>RNN</th>
<th>VQ-kernel*</th>
<th>Fukunaga’s kernel*</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU Time (min)</td>
<td>3.41</td>
<td>5.31</td>
<td>21.17</td>
<td>22.17</td>
<td>51.34</td>
<td>3.92</td>
<td>3378.45</td>
</tr>
</tbody>
</table>

* including CPU time used for classification

Table 6.1 CPU time used for finding the reduced set in the speech data experiment.
6.2.3 Summary

In this section we introduced the vector quantization technique into the area of nonparametric classifier design and showed that vector quantization is an extremely effective approach to data reduction.

Using the vector quantization data reduction technique, two methods of nonparametric classifier design, namely the VQ-kernel and the VQ-kNN, are proposed and tested with both synthetic and real data. Compared to other known nonparametric data reduction algorithms, the new methods are found:

1) to give much better results in terms of both the classification accuracy and the data reduction rate;
2) to have significantly less computational complexity in general;
3) to have control over the reduction rate; and,
4) to achieve a classification accuracy which is only moderately dependent on the reduction rate.

Theoretically, for highly nonparametric data the classification accuracy of our VQ-based classifiers will increase as $M_i$, the number of representatives in the reduced set, is increased. When the value of $M_i$ approaches that of the size of the original design set, the performance of our VQ-based classifiers will approach that of the basic (without reduction) Parzen and NN classifiers. This tendency is shown clearly in our experiments above. Therefore, the selection of $M_i$ becomes a trade-off: a larger $M_i$ generally yields higher accuracy but lower reduction rate, and a smaller $M_i$ yields lower accuracy but higher reduction rate.
6.3 Automatic Cry Analyzer Based on the VQ-kernel Methods

In this section, we build a system which automatically estimates the H-value of the cry signal. The H-value is a measure which indicates the level of distress of the infant who uttered the cry. We detail the system design and important technical considerations of our automatic cry analysis system. This system is built around the new VQ-based nonparametric classifier developed in the previous section. In particular, we choose the VQ-kernel classifier because of its superior classification accuracy over other nonparametric classifiers, as illustrated in the previous section.

6.3.1 System Design

Figure 6.3 shows the block diagram of our automatic cry analysis system. The system can be readily divided into three functional parts — preprocessing/feature extraction, VQ-kernel classifier, and H-value calculation.

6.3.1.1 Data preprocessing and feature extraction

As illustrated in Figure 6.3, the signal of the cry sound is first digitized at a sampling frequency of 8 kHz and a resolution of 12 bits per sample. Then a pre-emphasis filter with a transfer function of $1 - 0.95z^{-1}$ is applied to the digitized signal to remove its dc-component and to flatten the spectrum in order to remove the effect of glottal waveform and lip radiation characteristics [17, 36]. Finally, at the end of the data preprocessing, the cry signal is divided into overlapping frames by a 32 msec-wide sliding Hamming window which advances every 10 msec. Every frame has 256 samples.

The feature vector used in our automatic cry analyzer is composed of 24 elements. The first twelve elements are the zero-crossing rate, the energy, and the first ten LPC-derived
Figure 6.3 Automatic infant cry analysis system based on the VQ-kernel classifier.

Cepstrum coefficients (see subsection 4.3.2). To take into account the sequential characteristics of the cry attributes, the feature vector also includes the temporal derivatives of the first twelve elements. For reasons of simplicity, we approximate the derivative of each attribute by its first-order difference. Thus two consecutive frames are used to create the temporal
derivatives of the original twelve attributes, i.e.,

$$\frac{\partial \tilde{v}}{\partial t} \bigg|_{t=t_k} \approx \tilde{v}(t_k) - \tilde{v}(t_{k-1})$$  \hspace{1cm} (6.10)

where $\tilde{v}(t_k)$ is the vector of the first 12 attributes at a discrete time $t_k$. Thus, the final feature vector used in our present system has 24 elements, as shown in Figure 6.3. This means that each frame of 256 samples, or equivalently 32 msec, of the cry signal will be represented by a 24 element feature vector.

6.3.1.2 The VQ-kernel classifier and H-value calculation

The 24 element feature vectors obtained from the data preprocessing/feature extraction stage are used as inputs to the VQ-kernel classifier. The latter is designed so as to classify two types of input vectors — the E-type and H-type vectors, which will be defined below. The classification results are used in estimating the H-value of the cry.

We define the H-type vectors as feature vectors representing frames of one of the trailing, double harmonic break, dysphonation, hyperphonation, and inhalation cry phonemes. This definition is based on our H-value definition in equation (5.1) in Section 5.4. Similarly, the E-type vectors are vectors representing frames from the remaining five cry phonemes, i.e. the flat, falling, rising, vibration, and weak vibration cry phonemes (see Chapter 5 for the definitions of all cry phonemes).

With the above H-type vector definition and our definition of H-value in Section 5.4, it is clear that the H-value can be estimated by counting the number of H-type vectors (or, equivalently, the number of H-type frames) and then dividing it by the total number of non-silence frames (or vectors, accordingly) in the cry, i.e.

$$H\text{-value} = \frac{\text{the number of H-type vectors}}{\text{the total number of non-silence vectors}}$$  \hspace{1cm} (6.11)
The non-silence frames are easily detected by the energy in the feature extraction phase.

To design the VQ-kernel classifier, two training sets are generated; the first is composed of all the H-type vectors from the 45 training cry samples (see Chapter 5), and the second set is composed of all the remaining E-type vectors. Then vector quantization is applied to each of these sets to obtain the reduced training (design) sets. The reduced sets are then used in the design of the VQ-kernel classifier, as discussed in the previous section. Figure 6.4 shows the generation of the two reduced training sets.

**6.3.2 Determination of VQ-kernel Classifier Parameters**

As all kernel classifiers, the VQ-kernel classifier also requires the determination of the kernel function (refer to Subsection 6.2.1.3). Other important parameters that need to be determined for the VQ-kernel classifier are the decision threshold of the classifier [22] and the size of the code-books.
6.3.2.1 The kernel function

The VQ-kernel classifier uses the normal kernel function. Thus, the equation for kernel function is

$$K(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} x^T \Sigma^{-1} x \right], \quad (6.12)$$

where $d$ is the dimension of sample space, which is equal to 24 in our case, and $\Sigma$ is the covariance matrix. The value of the covariance matrix has a significant effect on the properties of the kernel function [22]. We use two different covariance matrices, $\Sigma_1$ and $\Sigma_2$ for the two different classes of feature vectors: the H-type and the E-type vectors. To choose the values of the two covariance matrices we conducted a series of experiments. We found that the highest classification accuracy was given by using the covariance matrix of the vectors of each training data set as the $\Sigma_1$ and $\Sigma_2$ of the respective class, i.e., for H-type signal we use $\Sigma_1 = \Theta_H$, where $\Theta_H$ is the covariance matrix of the H-type vectors in the original (before reduction) training data, and similarly, for the E-type signal we use $\Sigma_2 = \Theta_E$.

Thus, we get the following estimated conditional pdfs (see Subsection 6.2.1.3):

$$\hat{p}(x|\text{H-type}) = \frac{1}{(2\pi)^{d/2} N_H |\Theta_H|^{1/2}} \sum_{j=1}^{N_H} \exp \left[ -\frac{1}{2} (x - y_{Hj})^T \Theta_H^{-1} (x - y_{Hj}) \right] \quad (6.13a)$$

$$\hat{p}(x|\text{E-type}) = \frac{1}{(2\pi)^{d/2} N_E |\Theta_E|^{1/2}} \sum_{j=1}^{N_E} \exp \left[ -\frac{1}{2} (x - y_{Ej})^T \Theta_E^{-1} (x - y_{Ej}) \right] \quad (6.13b)$$

where $N_H$ and $N_E$ are the numbers of training vectors in the reduced H-type training set and the reduced E-type training set, respectively, and $y_{Hj}$ and $y_{Ej}$ are the $j^{th}$ feature vector in the reduced H-type training set and the reduced E-type training set, respectively.
6.3.2.2 The decision threshold

As indicated in [22, Chapter 7], when the kernel density estimate method is used, the likelihood ratio classifier becomes

\[
\ln \frac{\hat{p}(x|\text{H-type})}{\hat{p}(x|\text{E-type})} \geq t
\]

where \( t \) is the decision threshold. Note that when the a priori probabilities \( P(\text{H-type}) = P(\text{E-type}) \) (no information suggesting otherwise in our case), the likelihood ratio classifier is equivalent to the Bayes classifier (see [22]).

The decision threshold \( t \) is simply a means of compensating for the bias inherent in the density estimation procedure. In [22], Fukunaga suggests methods for determining \( t \) in the Parzen classifier with a normal kernel function. To determine the threshold \( t \) in our VQ-kernel classifier, we use the following approach.

First, the reduced training sets are generated via the vector quantization method. Then we use the resulting VQ-kernel classifier with a fixed value of \( t \) to classify each training vector in the original (not the reduced) design sets into the H-type or E-type class. The classification accuracy is computed after all the training vectors were classified. The above is repeated for different decision threshold \( t \). The decision threshold which achieves the highest classification accuracy is chosen in our classifier (6.14). This procedure is illustrated in Figure 6.5.

Figure 6.6 shows the curve of the classification error rate vs. \( t \), as \( t \) is varied from \(-4.5\) to \(4.5\). Figure 6.6 is the result of applying the above procedure to the 45 cry samples in our training set. In this experiment a code-book of 256 vectors was used for each reduced training set. As shown in Figure 6.6, the lowest error rate (14.71\%) occurs at \( t = 1.5 \). We then use this decision threshold in our final VQ-kernel classifier.
Chapter 6: CRY ANALYZER I: STATISTICAL METHOD

Reduced Training Set

Original Training Set \[\rightarrow\] Feature Extraction \[\rightarrow\] VQ-kernel Classifier \[\rightarrow\] Classification Error Estimation

Decision Threshold

Figure 6.5 Determination of the decision threshold, $t$.

Figure 6.6 The estimated classification error rate vs. the decision threshold $t$.

It is worthy to point out that while this procedure can find a reasonably good value for
t, the estimated classification error rate (as shown in Figure 6.6) is in fact the resubstitution error as defined in [22].

6.3.2.3 The size of code-books

Besides the parameters of the ordinary kernel method, as discussed in Section 6.2, the size of the code-books, i.e., the number of representative vectors in each of the reduced training sets, is the only additional parameter that needs to be predetermined in our VQ-kernel classifier. Theoretically, the larger the size of the code-book $M$ the higher the classification accuracy, and the higher computational complexity in the classification operation. However, as demonstrated by the examples in Section 6.2, the increase in the classification accuracy “saturates” as the size of the code-book reaches some point. Beyond this point, the accuracy will no longer show any significant improvement no matter how large the increase in the size of code-book is.

To determine the appropriate code-book size of the VQ-kernel classifier in our infant cry analyzer, we tested the VQ-classifier with code-book sizes of 32, 64, 128, 256, and 512 on each of our two training data sets. We found that the classification accuracy “saturates” when 256-vector code-books were used, and no further significant increase in the accuracy was observed when the code-book sizes were increased to 512. We therefore decided to use code-books of sizes of 256 each.

6.3.3 Training of VQ-kernel Classifier

The determination of the parameters such as the $t$, $M$, and the $\Sigma$'s, in fact, constitutes part of our training process. However, most of the computational effort in the training process is consumed by the estimation of the 256-level optimal code-books.
After the preprocessing/feature extraction phase, about 18,000 feature vectors for each of the two classes (H-type and E-type) are obtained from the original training data. Each feature vector is of dimension 24. Without our VQ-kernel algorithm, it would be impossible to apply the kernel method to such a huge training data set. We used the procedure described in Subsection 6.2.1.3 to condense each original training set containing 18,000 vectors, into a reduced design set which has only 256 vectors. It took about 100 hours of CPU time on a SUN SparcStation 2 computer to generate the two reduced sets. It is obvious from Table 6.1 that the computer time consumption would have been prohibitive (orders of magnitude larger) if the Fukunaga’s Parzen method was used to find the reduced sets.

6.4 Results and Discussions

We tested our infant cry analyzer with the cry samples in the testing set (see Chapter 5). Our infant cry analyzer gives as its output an estimated H-value after each of the cry samples in the testing set is processed. Figure 6.7 shows the relationship between the measured H-values (equation 5.1) and the H-value which is automatically estimated by our infant cry analyzer. Each small circle in the diagram represents a cry in the testing set. The vertical distance from a circle to the diagonal dashed line is the error of the estimated H-value. The mean error between the estimated H-value and the measured H-value is 14%. It is seen in Figure 6.7 that the automatic infant cry analyzer gives H-value estimations very close to the measured ones when the values of the actual measurements are high, and gives estimations higher than the measured values when the latter have low values ($\leq 20$).

In Figure 6.8, the H-values which are automatically estimated by the infant cry analyzer are related to the parents’ LOD ratings of the same cry samples in the testing set. Comparing
Figure 6.7 The H-values which are estimated by our VQ-kernel classifier-based cry analyzer versus the actual measurements. The estimation errors appear as the vertical distances from the circles to the diagonal dashed line.

Figure 6.8 with Figure 5.4, we can easily see the effect of the errors in the low H-value range we mentioned above, which appears to “scatter” the points in that region and move some of them towards the higher H-value region (the ideal locations of the data points should be around the diagonal from the lower left corner to the upper right corner on the scatter graph).

As expected, Figure 6.8 clearly shows a consistency between the parents’ LOD assessments of the cries and the H-values estimated by our automatic infant cry analyzer. This result is preliminary but is of significant importance because it establishes for the first time a
relationship between the parents’ perceptual assessment of the infants’ LOD conveyed in the cry sound and the automatically derived parameter (the H-value) of the LOD in the cry signal. More important is that we now have an effective and efficient method, which is based on digital signal processing and pattern recognition techniques, and which automatically extracts the H-value parameter from the cry sound. At the same time, the above results are a strong confirmation of the physical meaningfulness of 1) our use of the cry phonemes to represent the time-frequency patterns of normal infant cries and 2) of our use of the H-value to quantify the parents’ perception of the infants’ physical/emotional situations from the cry sound.

Figure 6.8 Relationship between the H-values estimated by the VQ-kernel classifier-based cry analyzer and the parents’ LOD ratings.
The performance of our automatic infant cry analyzer is expected to improve if the various parameters, especially those in the feature vectors are fine-tuned, and if the sequential characteristics of the cry patterns are further exploited. In the next chapter, we present another approach for designing our proposed automatic infant cry analyzer, which is based on the hidden Markov modeling technique, and which gives even better performance with our present infant cry data base.
Chapter 7

CRY ANALYZER II: A HIDDEN MARKOV MODEL (HMM) -BASED METHOD

In this chapter, we apply the hidden Markov model (HMM) technique to normal infants' cry sound analysis [128]. As in Chapter 6, the purpose here is the automatic estimation of the infant LOD from the cry signal. However, we now base our estimation on the HMM technique. As discussed in Chapter 3, the HMM technique is a powerful stochastic modeling approach. Recently it has been successfully applied to many automatic speech recognition systems as an alternative to the traditional dynamic time warping (DTW) method. In particular, the HMM is capable of dealing with the randomness in both the temporal and spectral structure of the signal.

Considering that infant crying is a special kind of human utterance, and that the generation of infant crying can be modeled via an approach which is similar to that of the generation of adult speech (see Chapter 4), lead us to believe that it is possible to utilize the hidden Markov modeling technique in our proposed automatic infant cry analysis system.

The main characteristic which distinguishes our HMM-based cry analysis system from the VQ-kernel classifier-based system described in the previous chapter is that here the recognition of the different time-frequency patterns is based on the structural information of the signal rather than on its statistical information.

In the following, we will first describe the structure of our HMM-based cry analyzer. Then we will discuss the training and recognizing procedures of the HMMs and the system as a whole. Finally, we will present experimental results and some concluding remarks.
7.1 System Design

Our HMM-based infant cry analysis system is composed of the functional blocks shown in Figure 7.1. These are 1) data preprocessing/feature extraction to transfer the signal into a long sequence of feature vectors, 2) vector quantization to encode the sequence of feature vectors into a long sequence of symbols, 3) segmentation to divide the long symbol sequences into short segments suitable for the HMMs, 4) the application of HMMs to recognize the different cry phonemes, and finally 5) the H-value calculation.

![Figure 7.1 The block diagram of our HMM based cry distress level analysis system.](image)

7.1.1 Data Preprocessing and Feature Extraction

The data preprocessing/feature extraction is similar to that in our VQ-kernel classifier-based cry analyzer discussed in the previous chapter. As before, after the cry sound signal is lowpass-filtered and digitized at 8kHz, a pre-emphasis filter with a transfer function of $1 - 0.95z^{-1}$ is applied to flatten the spectrum so that the effect of glottal waveform and lip radiation characteristics can be removed. Then, a sliding Hamming window with a width of 32 msec
Chapter 7: CRY ANALYZER II: HMM METHOD

Figure 7.2 Vector quantization.

is applied to extract frames of 256 points each from the digitized signal. The Hamming window's sliding step is 10 msec, thus the consecutive frames overlap.

Then, the procedure diverges from the feature vector construction in the VQ-kernel classifier-based cry analyzer. In particular, we now extract only a twelve-dimension feature vector from each signal frame. The feature vector is composed of the normalized short-time energy, the zero-crossing rate, and the first 10 cepstrum coefficients. The other 12 features used in the previous cry analyzer which represent the time difference are not included.

Silence periods or pauses in the cry sound are detected by comparing the short-time energy to a preset energy threshold. If the short-time energy is below the energy threshold, the frame is considered and marked as a silence frame.

7.1.2 Vector Quantization of the Feature Vectors

Vector quantization is an important technique for reducing data redundancy in HMM-based voice processing systems. In particular, it allows the use of discrete HMMs so that the implementation of the system is greatly simplified [50, 95, 98].

We use a code book with 128 prototype vectors, or code-words, to quantize every feature vector. The code book was obtained by using Linde et al.'s vector quantizer design algorithm [61]. As shown in Figure 7.2, after quantization each feature vector (representing a frame) is represented by a code-word index. Thus, at the output end of the vector quantizer, the
cry signal is transformed into a sequence of code-word indices and silence indicators which mark the pauses in the cry signal.

7.1.3 Segmentation of the Signal into Recognition Units

In speech recognition, the nature of spoken sentences provides an opportunity for using a hierarchical decomposition approach. This approach is based on the fact that there is always some grammar governing the generation of all the allowable sentences. A sentence itself can always be further decomposed into a sequence of words, and finally, a word can be further decomposed into a sequence of phonemes. This hierarchical structure is very suitable and efficient in the implementation of HMMs-based systems [50, 98].

Unfortunately, this kind of clear hierarchical structure (grammatical — syntactic — phonetical) is not available in infant crying. It is not easy to use the cry phonemes as basic recognition units on which we can base our HMM system. The first reason for this difficulty is that we have no effective method to detect the start and end points of cry phonemes, making it very difficult to separate them in a long continuous cry signal. Secondly, even if we could effectively isolate each cry phoneme in the cry signal, it would still be difficult to establish an effective pattern analysis method for the recognition of the cry phoneme. This is because the duration of a cry phoneme often varies widely (up to 10 folds), and cry phonemes can be interrupted by unpredictable short pauses or uncharacteristic distortions in the time-frequency pattern.

To cope with the above problems, we propose dividing each cry phoneme into more fundamental recognition units. We found that, despite great variations in duration and short interruptions or distortions within a cry phoneme, the time-frequency characteristics of a cry
phoneme boundaries

\begin{align*}
distortion/interruption & & & & & & & \\
\text{Segmentation} & & & & & & & \\
a micro-segment & & & & & & & \\
\end{align*}

c.p. — cry phoneme
* — micro-segments affected by phoneme boundary or distortion

**Figure 7.3 The segmentation of a cry signal.**

phoneme show a "pseudo-stationary" pattern. Therefore, if we divide a cry phoneme into short "micro-segments" of fixed length, these micro-segments show great similarity to each other in terms of their time-frequency characteristics. On the other hand, micro-segments from different types of cry phonemes will be dissimilar because different cry phonemes are defined on the basis of distinguishable time-frequency characteristics. Since all the micro-segments are of the same duration, it is easier and more reliable to analyze a micro-segment, and to identify the cry phoneme to which the micro-segment belongs. This concept of segmentation is illustrated in Figure 7.3. Figure 7.3 shows an example of a part of a cry signal containing five cry phonemes (c.p. 0, c.p. 1, c.p. 2, c.p. 3, and c.p. 4) each with varying durations. After segmentation, the signal is divided into micro-segments with fixed durations. We choose the duration of the micro-segments to be short enough to ensure that the shortest cry phonemes can still be further divided into a series of micro-segments.
The boundaries between each cry phoneme need not be detected in this scheme. As shown in Figure 7.3, a phoneme boundary will only affect the micro-segment to which the boundary belongs. Such an affected micro-segment will likely be misrecognized since it contains mixed information from two different cry phonemes. Similarly, if a distortion or interruption occurs within a cry phoneme (as illustrated in c.p. 1 in Figure 7.3), the few micro-segments to which this part of the signal belongs will be affected, and will likely be misrecognized later. However, these corrupted micro-segments only constitute a tiny portion of the entire signal. The majority of the remaining micro-segments will be reliably recognized.

As shown in Figure 7.1, the above segmentation scheme is implemented after performing vector quantization, i.e., we apply our segmentation scheme to the code-word indices representing the feature vectors of the overlapping frames of the cry signal. The overlapping frames are shifted by 10 msec intervals. Thus for every 10 msec there exists a code-word index. From our experiment, we observed that the duration of a cry phoneme is typically around 1 second (but may vary between 0.5 to 3 seconds). Thus, each cry phoneme is typically represented by 100 code-word indices. The sequence of code-word indices is segmented into micro-segments (silence periods are skipped). Each micro-segment contains ten code-word indices. This ensures that a cry phoneme is segmented into at least five micro-segments. The micro-segments are then used as the basic recognition units in our HMMs-based recognizer.

7.1.4 Configuration of the HMM-based Analyzer and the Calculation of the H-value

To identify the cry phoneme to which each micro-segment belongs, one HMM is assigned to each of the ten cry phonemes (the definitions and basic operations of HMMs are briefly explained in Section 3.6 and in Appendix D, also see [95, 98] for a full description). A 3-state left-to-right topology, as shown in Figure 7.4, is used for every HMM. Each state in an HMM
has two outgoing transitions, one to the state itself and the other to the next state. At any instant of time only one state is occupied or reached. When a state is reached, a symbol is emitted. The symbol is one of the 128 code-word indices of the prototype vectors in the VQ code book. The decision as to which symbol is emitted is governed by the output probability density function of the occupied state. In our HMMs, the output probability density function of a state is represented by a table, which consists of 128 probabilities corresponding to the 128 code-word indices. Each of the two outgoing transitions is also associated with a transitional probability. If the model at time $t$ occupies state $i$, the transitional probability defines the probability of the model reaching state $j$ at time $t + 1$.

The values of all the above model parameters such as the transitional probabilities and the output probabilities are determined by training. Every HMM is trained with only the micro-segments associated with the cry phoneme the HMM represents. For example, the HMM representing trailing is trained with all the micro-segments of all the trailing cry phonemes in the training set.

Given an HMM and a micro-segment, the probability of the model generating this micro-segment is determined by using algorithms such as the forward-backward procedure in [95, 98]. This probability is known as the model generation probability (see Appendix D). As a result of the training, the parameters of each HMM are set such that the model generation

---

**Figure 7.4 Topology of the hidden Markov models for micro-segment identification.**
probability over all the micro-segments belonging to the cry phoneme represented by this HMM will have the highest mean value.

Since we want to estimate the H-value (and not to identify each micro-segment), we divide the ten HMMs into two groups. This is done after each HMM is individually trained, as described above. The first group is composed of the five HMMs corresponding to the five cry phonemes involved in the definition of the H-value, (namely the trailing, double harmonic break, dysphonation, hyperphonation, and inhalation cry phonemes, see Section 5.4). The other group is composed of the remaining five HMMs representing the flat, falling, rising, vibration, and weak vibration cry phonemes. Figure 7.5 shows how the first group of HMMs are assembled to form a big HMM. We call it the H-HMM. The other big HMM, which we call the E-HMM, has the same topology as that of Figure 7.5, but is composed of the other group of HMMs representing the flat, falling, rising, vibration, and weak vibration cry phonemes.
From To H-valueSegmentation Calculation

Figure 7.6 The HMMs-based classifier.

When a micro-segment and a big HMM are presented to the forward-backward procedure (see Appendix D and [95, 98] for detailed description of the procedure and corresponding algorithms), the probability of this big HMM generating the micro-segment is determined. Because of the topology of our big HMMs, it is straightforward to show that if the micro-segment belongs to one of the five cry phonemes in this big HMM, then this big HMM will more likely give a higher model generation probability than the other big HMM.

Our cry analyzer is built around the H-HMM and the E-HMM, as shown in Figure 7.6. Each micro-segment is presented to both the H-HMM and the E-HMM simultaneously. The likelihood of a micro-segment being generated by each of the H-HMM and E-HMM is computed using the forward-backward procedure. The micro-segment is then classified as either H-type or E-type according to which model generation likelihood is larger.

The classification results from the HMM analyzer are collected to calculate the H-value of the cry signal. From its definition in Section 5.4, the H-value is estimated as

$$\text{H-value} = \frac{\text{the number of H-type micro-segments}}{\text{the total number of micro-segments in the whole signal}}.$$
7.2 Training of the HMMs

The training of our HMM-based cry analyzer system takes two steps: first, the training of each of the ten individual HMMs, and then the training of the H-HMM and E-HMM.

There are a total of 2804 micro-segments in the 45 cries in our training set. Table 7.1 shows the distribution of these micro-segments over the ten different cry phoneme groups in the training data.

Uniform initialization is used for all the transitional probabilities in all the ten individual HMMs, i.e., all transitions from a state are considered equally likely. The output probability density table of each of the three states in an HMM is initialized as follows: each of the micro-segments in the group corresponding to that HMM was divided into three parts — the beginning, middle, and end parts, and the frequency of occurrences of each of the 128
code-word indices in these three parts was used as the initial values corresponding to that symbol in state 1, 2, and 3 of that HMMs, respectively.

After initialization, the forward-backward algorithm (see Appendix D) was used to optimize the parameters in the HMMs. For each HMM, we ran the forward-backward algorithm iteratively with the micro-segments belonging to the cry phoneme of this HMM as the training data. The training of an HMM ended when the averaged model generation probability over this group of micro-segments showed no further significant increase.

In the second step of the training, the H-HMM and E-HMM were initialized with the optimized parameters from the ten individually trained HMMs. Using the same training procedure described above and a joint training data set from the five cry phonemes group which defines the H-value, the parameters in the H-HMM, consisting of all the parameters in the five individual HMMs, were refined. The same procedure was applied to refine the parameters in the E-HMM, except that the joint training data set from the other five cry phonemes group was used.

This second training is proven to be important to our HMM-based cry analyzer as our experiment shows that the accuracy of classifying the individual micro-segments in the testing set increased from 80.15% to 82.67% after the second training.

7.3 Experiment Results and Discussion

With a similar procedure we used to evaluate our VQ-kernel classifier-based infant cry analyzer, the performance of our HMM-based infant cry analyzer was tested with the 58 cries in the testing set. The system automatically gave an estimation of the H-value after each of the 58 cries was analyzed.
Figure 7.7 The computer estimations versus the actual measurements of the H-values of the 58 cries in the testing set. The estimation errors appear as the vertical distances from the circles to the diagonal dashed line.

Figure 7.7 shows the 58 estimated H-values versus their true measurements. Each small circle in the diagram represents a cry in the testing set. The vertical distance from a circle to the diagonal dashed line is the error of the estimated H-value. The mean absolute error between the HMM-based system estimation and the actual measurements is 12.9%. This demonstrates that our HMM-based method is effective in estimating the H-values from the cry sounds.
Figure 7.8 shows the relationship between the H-value estimated with our HMM system and the parents’ level-of-distress (LOD) rating. All the estimated H-values, except the one whose coordinates are [69, 1.05], do show a clear trend of consistency with the LOD ratings given by the parents.

Comparing this result (Figure 7.8) with that of our VQ-kernel classifier cry analyzer reported in the previous chapter (Figure 6.8), we can see that the HMM-based method gives better estimations when the cry samples under analysis have low LOD ratings. This is shown in Figure 7.8 as that cries with low LOD rating (≤ 3) are consistently assigned to low H-
values (≤ 50) by our HMM-based cry analyzer. However, for cries with high LOD rating (≥ 3), the VQ-kernel classifier-based method shows slightly better performance, as shown in Figure 6.8. Overall, the performance of our HMM-based method is better than that of our VQ-kernel classifier-based method. The mean error of the estimated H-value in our HMM-based system is 12.9%, while the mean error rate is 14.0% in our VQ-kernel classifier system. We believe that the better performance of our HMM-based system is due to the more effective exploitation of the sequential (structural) information of the time-frequency characteristics in the cry sound.
8.1 Accomplishments and Contributions

Since the 19th century, scientists have believed that infant cry sounds convey information about the baby’s physical, physiological and emotional situation. As a result of generations of research, a wealth of observations and findings about infant crying have been achieved. However, as our literature survey in Chapter 2 shows, practical applications of these achievements remain scarce.

The recent advancements in digital signal processing theories and techniques, coupled with the rapid development of computer technology, lead us to believe that we now are in a position to investigate the development of methods and means by which infant cry analysis could be automated. In this work, we began such development by combining what we already know about infant crying with modern signal analysis techniques.

A fundamental accomplishment of our research is the establishment of a platform, based on modern speech processing/recognition theories and techniques, for the automatic analysis of infant cry signals.

In doing so, we first identified the problems and difficulties facing the study of infant cries, particularly those hindering the utilization of automated analysis methods. Then, to solve and alleviate these problems, we proposed some modifications to the traditional methodology of infant cry analysis. These included the use of modern signal analysis approaches, as well
as the introduction of an efficient measure to characterize the cry signals. We argued that efficiency here must be defined in terms of reliable detection and recognition of this measure by automated means.

In particular, our contributions are the following. We have,

1. **Demonstrated the compatibility of infant crying with adult speech generation models**

   This is important since most successfully-applied modern speech processing techniques are based on models of adult speech generation. In showing this compatibility, we discussed the physiological similarities of infant crying to adult speech, surveyed and analyzed infant cry generation theories and models proposed previously by others. We have also compared these models with the speech generation models commonly used in modern digital speech processing/recognition theories. To further support this compatibility, we conducted experiments to estimate the formant frequencies of infant cries using the linear prediction coding (LPC) method. LPC is a method which is heavily based on the definition of the signal generation model and which is well proven in speech processing applications. The results of the experiments clearly showed that the LPC-based method gives good estimation of the formant frequencies of the infant cries (see Chapter 4).

   Our investigation concluded that: a) the same *excitation-modulation* voice generation principle which applies to adult speech also applies to infant crying, b) the modulation process of the infant cry generation can be formulated as a time-varying filter system, in a fashion similar to that of the adult speech generation, and c) an *all-pole* digital model, as well as the LPC method based on such a model, is effective in analyzing and representing normal infant cries.
This investigation set up the basis for applying the techniques and methods originally developed for speech signals to that of infant cry sounds analysis and processing.

2. **Discussed the problems of the automatic cry analysis and the automatic assessment of normal infants' distress levels**

   We addressed one particular issue of normal infant cry research — the automatic assessment of the infants’ distress level via analyzing the infants’ cry signals. We argued that, to build a device which can assess the infants’ physical/emotional situations, a new approach is required. The conventional approach classifies cries according to some probable or presumed cry stimuli, such as pain, fussy, and hunger. At present, this approach is not as reliable and practical as that of correlating the attributes of the cry signal to the parents’ and other caretakers’ subjective perceptions about the infants’ general situation conveyed in the cry. We also suggested the use of a single parameter — the level-of-distress (LOD) — to evaluate the cry-implied physical/emotional situation of the infant.

3. **Introduced the cry phonemes concept and the H-value parameter to characterize the cry sound, and to evaluate the infants’ LOD conveyed in the cry**

   To effectively and efficiently characterize and represent the varieties of the infant cry signals, we introduced a set of “cry phonemes”. This set is based on the infant cry generation models and the typical time-frequency patterns which we have identified in normal infant cries. Our set of cry phonemes was chosen to meet two important criteria: a) any normal infant cry signal can be divided into, and thus represented by, a sequence of cry phonemes belonging to this set; and b) the cry phonemes in the set can be reliably recognized by means of automatic signal processing and recognition techniques.
With our definition of the cry phonemes set, we conducted experiments to analyze the relationship between the individual cry phoneme and the LOD perceptions by parents who have listened to the recording of the cry. Based on the results of our experiments using recorded cries from 36 newborns, we defined an indicator — the H-value — as a quantitative measure of the infants’ LOD conveyed in the cry. From the experiments we found that the H-value measured from a cry signal always shows clear consistency with the parents’ LOD rating on that cry.

4. Developed and tested two automatic infant cry analyzers — the VQ-kernel classifier-based cry analyzer and the HMM-based cry analyzer

Based on our cry phonemes set and the H-value, we developed two automatic infant cry analyzers which are capable of making humanlike assessment of the normal infants’ LOD.

The principle behind both of our two designs is to estimate the H-value of the cry by detecting and calculating all occurrences and durations of each of the ten different cry phonemes in the cry signal.

In the first system, we used a nonparametric classification method to classify cry phonemes of the cry signal after dividing the signal into frames. Because of the huge amount of data involved, we developed a set of highly efficient vector quantization-based nonparametric classification algorithms. These new algorithms, while retaining a high classification accuracy, can greatly reduce the storage requirement and drastically boost up the classification speed. The new algorithms we developed are general in nature, and are therefore suitable to handle any large scale problem, which may cause difficulties for traditional nonparametric methods.
Our second automatic infant cry analyzer uses the hidden Markov modeling (HMM) technique. This technique has become very popular recently and has been shown to possess many advantages when applied to automatic speech recognition. Comparing this HMM-based cry analyzer to our nonparametric classifier-based one, the HMM-based approach puts more emphasis on the sequential structure of the time-frequency patterns in the cry signal. In the HMM-based cry analyzer we introduced an approach based on "micro-segments" of the signal representation. This makes it possible to associate the cry signal with a hierarchical structure which allows the efficient implementation of the HMM method.

The performances of our two different automatic infant cry analyzers were tested and evaluated with a set of 58 cry samples from normal infants. The test results show that both of our two infant cry analyzers perform well. In terms of the estimation accuracy of the H-value, the HMM-based analyzer achieves an accuracy of 87.1%, while the nonparametric classifier-based analyzer gives an accuracy of 86.0%. When the automatic estimations of the H-value given by these analyzers are related to the LOD ratings given by the parents, clear consistency between them can be observed.

In conclusion, our research clearly shows that it is feasible to utilize efficient modern digital signal analysis methods to detect subtle characteristics and features in the infant cry signals. These features and distinctions are otherwise undetectable with the traditional methods used in the past. Our research also establishes relationship between the objective measurements of information conveyed in the cry signals (determined automatically) and the subjective human perceptions of that information.
The above achievements and contributions show that applying modern signal processing techniques to the study of infant cries promises success. We have only opened the door to a very exciting area of research where there is much room for improvements of methods and are many important applications. Some of these are further discussed in the following section.

8.2 Some Topics for Future Study on Automatic Infant Cry Analysis

8.2.1 Refining the Techniques and Methods We Have Developed

We have shown, both theoretically and experimentally, the feasibility of automating infant cry analysis by using modern speech processing/recognition techniques, and developed two systems to demonstrate this concept. However, time and scope limitations have prevented us from further refining many technical aspects of the design and the implementation of the infant cry analysis systems reported in this dissertation.

Opportunities for technical refinements and further investigations exist in the following areas:

- **Feature selection and extraction:**

  In our infant cry systems, there are still many promising feature candidates that may be studied. Examples of such features are the fundamental frequency, the formant frequencies, and the transitional cepstrum coefficients. Note that while some very coarsely defined transitional cepstrum coefficients (equation 6.10) are used in our VQ-kernel classifier-based infant cry analyzer, but none is used in our HMM-based cry analyzer. It is worthwhile investigating whether or not these features could provide improved performance over those used in the systems reported in this thesis.
Chapter 8: CONCLUSIONS

- More sophisticated model structures and topologies for the HMMs-based system:
  In our HMMs-based infant cry analysis system, for simplicity we use the basic structure and basic model operation of the HMMs. More sophisticated model structures however are worthy of investigation in the future to obtain better system performance and efficiency. Among the many other models, the continuous density HMMs and the HMMs with explicit state duration density [98] deserve special attention. In the continuous density HMMs, the observations are not symbols chosen from a finite alphabet. Instead, they are real vectors with continuous values. This eliminates the possibility of performance degradation caused by the quantization process in the discrete HMMs-based systems. The HMMs with explicit state duration density attempt to overcome the major weakness of conventional HMMs, which is the modeling of the state duration. Each of these two alternative versions of the conventional HMM has been successfully used in many speech recognition applications [102, 105].

  Further refinement may also be possible in the selection of HMM topologies used in our infant cry analysis system, especially the one we chose for the ten individual HMMs.

- Improved system training and performance evaluation:
  A larger infant cry sample database is always desirable so as to train and test our infant cry analysis systems better. Like any other classification/ recognition systems, the performances of our systems are greatly dependent upon the training process. The larger the training sample database, the better trained our cry systems will be.

  Another key factor to the success of our cry analyzer design is the performance evaluation. Since we judge the system performance solely against the parent rating database, the quality of this rating database becomes crucial. The performance evaluation
of our infant cry analysis systems will be improved by using a larger testing sample set and an improved parent rating database formed from a larger number of parents.

- **New techniques for the representation and processing of signals:**

  Many new theories and technologies are recently being applied to research in speech/voice processing, image processing, pattern recognition, computer vision, and other signal processing fields. Especially noteworthy are the neural network [62, 94, 80], the fuzzy logic, the wavelet transformation [48, 58, 66], and the instantaneous frequency (IF) representation theories [6, 7]. Most of these techniques have been shown to possess some advantages in one aspect or another when applied to voice and speech processing related problems. The feasibility and merits of incorporating these new technologies into the infant cry analysis are certainly worthwhile pursuing.

**8.2.2 Extending Our New Methodology to Other Infant Cry Research Topics**

Although our automatic infant cry analysis systems are designed and trained with the cry data from a specific age group of normal infants, it is important to note that the methodology we have developed is not restricted to this case. Our methodology can be generalized and applied to other age groups of infants and to problems other than the automatic assessment of infants' LOD. In particular, our methodology applies to the problem of automated diagnosis of specific diseases and/or abnormalities in sick infants.

**8.2.2.1 Developing analysis/recognition/diagnosis systems for abnormal infants' cry signals**

With the sound spectrograph and, recently, with some computer-aided analysis methods, many scientists have carried out investigations in the area of abnormal infant cry analysis. Cries
from abnormal infants with various diseases, especially those which affect the baby’s central nervous system, have been studied (see the survey in Chapter 2).

These investigations have generated a wealth of findings about the abnormal infant crying. In many occasions, differences amongst the various attributes of the cries of sick infants and those of the clinically normal infants are reported [34, 63, 69–78, 103, 109, 116]. Thus, it will be worthwhile to extend our signal processing-based infant cry analysis technique to the analysis of abnormal infant cries.

From a technical viewpoint, the problem of abnormal infant cry analysis is better defined than that of normal infants’. The physical situation of the infant (either the disease, or the abnormality, or both) can usually be determined/diagnosed by using conventional (though more costly, complex, or inconvenient) medical methods. In addition, in abnormal infant cry research it is the “pain cry” that is of interest. By studying the “pain cry” only, it is possible to standardize and unify the situations, under which the cry signals are acquired, and the stimuli inducing the cries. It has also been pointed out that pain is the maximal stimulus to the nerve system of the infant, hence may significantly overshadow the influence of other factors in the process of cry generation. Because of these factors, we expect the methodology we developed in our two automatic infant LOD assessment systems to be effective, especially the approach of identifying typical time-frequency patterns and defining the corresponding cry-phonemes. Different diseases or abnormalities may need different definitions of cry-phoneme sets to best characterize the particular problem.

In these types of studies, the cooperation with pediatricians and medical staff will be essential. This is specially true for rare and difficult-to-diagnose diseases.
8.2.3 Developing New Products Based on Our Computerized Infant Cry Analysis Techniques

We believe it is now feasible to develop automatic cry analyzing devices based on the techniques we have developed, by using modern advanced VLSI and computer hardware techniques. In the following, we list a few possible products:

1. **Intelligent baby alarm systems with the ability of assessing the baby’s LOD**

   Our automatic infant LOD assessment techniques could be incorporated into ordinary baby alarm systems. In ordinary baby alarm systems, the device is usually triggered when the sound level in the baby’s room exceeds a preset threshold. This simple strategy is not only subject to false alarms, but it is also unable to provide the parents with more detailed information about the infants’ physical/emotional situation. The incorporation of our automatic infant LOD assessment techniques into ordinary baby alarm systems will give the alarm device some intelligence; it will be able to distinguish infant cries from various background noise sources so as to avoid false alarms, and estimate and report to the parents the infants’ LOD situation. This kind of device will be particularly helpful to deaf and hard of hearing parents.

2. **Monitoring systems used in hospital wards which can detect the cries, and automatically analyze and record the LOD situations of each infant**

   This kind of equipment can make the supervision of infants more efficient by supplying the caretaking environment with accurate information about each infant’s situation, and automatically registering that information for later analysis.

3. **Computer cry analysis systems to assist physicians in the diagnosis of infant diseases**
This has been one of the main objectives of research into the abnormal infant cry. Infant disease diagnosis by cry analysis will facilitate diagnostic procedures, and make them convenient, more economical, and faster. Such a system may contain different sets of parameters and databases for different kinds of diseases, and for different age groups of infants. This will increase the accuracy of analysis and, at the same time, enable the system to diagnose different diseases and handle infants of different ages.

The development of such systems obviously requires further research on abnormal infant cries.
Appendix A

Linear Predictive Coding (LPC) Algorithms

Linear predictive analysis is one of the most powerful speech analysis techniques developed in recent years. The importance of this method lies both in its ability to provide accurate estimates of the speech parameters, and in its computational effectiveness.

We have already summarized the basic principles of linear predictive analysis in Section 3.2. Here we will outline some commonly used algorithms in LPC analysis.

As indicated in Section 3.2, to get an estimate of the vocal transfer function

\[ \hat{H}(z) = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \]  

we need to estimate the coefficients \( \alpha_k \)'s by minimizing the average squared prediction error

\[ E_n = \sum_m e_n^2(m) = \sum_m (s_n(m) - \hat{s}_n(m))^2 \]  

where \( s_n(m) \) is a segment of the speech waveform that has been selected in the vicinity of sample \( n \), i.e., \( s_n(m) = s(m + n) \). Substitute

\[ \hat{s}_n(m) = \sum_{k=1}^{p} \alpha_k s_n(m - k) \]

into equation (A.2) and set \( \partial E_n / \partial \alpha_i = 0, \quad i = 1, 2, \ldots, p \), so as to find the values of \( \alpha_k \) that minimize \( E_n \). We obtain

\[ \sum_m s_n(m - i)s_n(m) = \sum_{k=1}^{p} \alpha_k \sum_m s_n(m - i)s_n(m - k) \quad 1 \leq i \leq p \]  

If we define

\[ \phi_n(i, k) = \sum_m s_n(m - i)s_n(m - k) \]  

(A.5)
then equation (A.4) can be written more compactly as
\[ \sum_{k=1}^{p} \alpha_k \phi_n(i, k) = \phi_n(i, 0) \quad i = 1, 2, \ldots, p \] (A.6)
This set of \( p \) equations in \( p \) unknowns (the predictor coefficients \( \{\alpha_k\} \)) can be solved in an efficient manner.

One approach solving this equation set, called autocorrelation method, assumes that \( s_n(m) = s(m + n)w(m) \), where \( w(m) \) is finite length window (e.g., a Hamming window) that is identically zero outside the interval \( 0 \leq m \leq N - 1 \). This approach also defines the average squared prediction error as
\[ E_n = \sum_{m=-\infty}^{+\infty} e_n^2(m). \] (A.7)
Since \( s_n(m) \) is nonzero only for \( 0 \leq m \leq N-1 \), from equation (A.2) and (A.3), we can notice that \( e_n(m) \), for a \( p \)th order predictor, will be nonzero only for \( 0 \leq m \leq N-1+p \). Thus,
\[ E_n = \sum_{m=-\infty}^{+\infty} e_n^2(m) = \sum_{m=0}^{N-1+p} e_n^2(m). \] (A.8)
With similar analysis, we can show that
\[ \phi_n(i, k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k) \quad 0 \leq k \leq p \] (A.9)

We define the autocorrelation function of \( s_n(m) \) as
\[ R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)s_n(m+k) \] (A.10)
Thus, the right-hand side of equation (A.9) can be seen as the autocorrelation of \( s_n(m) \) for \( (i-k) \). That is,
\[ \phi_n(i, k) = R_n(i-k). \] (A.11)
Since $R_n(k)$ is an even function, equation (A.6) can be expressed as
\[ \sum_{k=1}^{p} \alpha_k R_n(|i-k|) = R_n(i) \quad 1 \leq i \leq p \] (A.12)
or in matrix form as
\[
\begin{bmatrix}
R_n(0) & R_n(1) & R_n(2) & \cdots & R_n(p-1) \\
R_n(1) & R_n(0) & R_n(1) & \cdots & R_n(p-2) \\
R_n(2) & R_n(1) & R_n(0) & \cdots & R_n(p-3) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
R_n(p-1) & R_n(p-2) & R_n(p-3) & \cdots & R_n(0)
\end{bmatrix}
\begin{bmatrix}
\alpha_1 \\
\alpha_2 \\
\alpha_3 \\
\vdots \\
\alpha_p
\end{bmatrix}
= 
\begin{bmatrix}
R_n(1) \\
R_n(2) \\
R_n(3) \\
\vdots \\
R_n(p)
\end{bmatrix}
\] (A.13)
The $p \times p$ matrix of autocorrelation values is symmetric and is also a Toeplitz matrix.

In the following, we give a recursive algorithm for solving equation (A.13) by exploiting its Toeplitz nature. This procedure, called Levinson-Durbin’s recursive procedure, is one of the most efficient method known for solving this particular system of equations [14, 53]. The procedure can be stated as follows:

\begin{align*}
E_n^{(0)} & = R_n(0) \\
E_n^{(i)} & = \frac{R_n(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R_n(i-j)}{E_n^{(i-1)}} \\ 
\alpha_i^{(i)} & = k_i \\
\alpha_j^{(i)} & = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \\
E_n^{(i)} & = (1 - k_i^2) E_n^{(i-1)}
\end{align*}
(A.14)-(A.18)
Equations (A.14)-(A.18) are solved recursively for $i = 1, 2, \ldots, p$ and the final solution is given as
\[ \alpha_j = \alpha_j^{(p)} \quad 1 \leq j \leq p \] (A.19)
Therefore, at time $n$, we obtained estimates of the digital filter coefficients $\{a_k\}$ of the vocal transfer function stated in equation (A.1). These estimates are optimal in the sense of minimizing the average squared error of linear prediction over the segment of speech of $\{s(m+n), \ 0 \leq m \leq N - 1\}$. More detailed discussions on LPC algorithms and application can be found in [91, 101].
Appendix B: Linde-Buzo-Gray's Algorithm of Vector Quantizer Design

Linde-Buzo-Gray's Algorithm of Vector Quantizer Design

Linde-Buzo-Gray's (LBG) method actually contains two different algorithms: one for problems with known distribution function, and the other for problems in which the distribution properties of the input random vectors are unknown. Since we are only interested in the latter case, the algorithm for known distribution is omitted in this appendix.

The LBG algorithm (for unknown distribution) is stated as follows:

(0) Initialization: Given $N$, number of levels, the distortion threshold $\epsilon \geq 0$, an initial $N$-level reproduction alphabet $\hat{A}_0$, and a training sequence $\{x_j; j = 0, \ldots, n - 1\}$. Set $m = 0$ and $D_{-1} = \infty$.

(1) Given $\hat{A}_m = \{y_i; i = 1, \ldots, N\}$, find the minimum distortion partition $\mathcal{P}(\hat{A}_m) = \{S_i; i = 1, \ldots, N\}$ of the training sequence: $x_j \in S_i$ if $d(x_j, y_i) \leq d(x_j, y_l)$ for all $l$, where $d(x, y)$ is the distortion measure (or distance) between $x$ and $y$. Compute the average distortion

$$D_m = D\left(\mathcal{P}(\hat{A}_m)\right) = n^{-1} \sum_{j=0}^{n-1} \min_{y \in \hat{A}_m} d(x_j, y). \quad (B.1)$$

(2) If $(D_{m-1} - D_m)/D_m \leq \epsilon$, halt with $\hat{A}_m$ as the final reproduction alphabet. Otherwise continue.

(3) Find the optimal reproduction alphabet $\hat{x}(\mathcal{P}(\hat{A}_m)) = \{\hat{x}(S_i); i = 1, \ldots, N\}$ for $\mathcal{P}(\hat{A}_m)$. Here, $\hat{x}(S_i)$ can simply be the centroid of $S_i$ (under the distance definition $d(x, y)$). Set $\hat{A}_{m+1} \triangleq \hat{x}(\mathcal{P}(\hat{A}_m))$. Replace $m$ by $m + 1$ and go to (1).
Appendix B: Linde-Buzo-Gray's Algorithm of Vector Quantizer Design

For the choice of the initial alphabet \( \hat{A}_0 \), Linde et al. suggest a "splitting" approach, as stated below.

(0) Initialization: Set \( M = 1 \) and define \( \hat{A}_0(1) = \hat{x}(A) \), the centroid of the training sequence.

(1) Given the reproduction alphabet \( \hat{A}_0(M) \) containing \( M \) vectors \( \{y_i; i = 1, \ldots, M\} \), "split" each vector \( y_i \) into two close vectors \( y_i + \theta \) and \( y_i - \theta \), where \( \theta \) is a fixed perturbation vector. The collection \( \hat{A} \) of \( \{y_i + \theta, y_i - \theta, i = 1, \ldots, M\} \) has \( 2M \) vectors. Replace \( M \) by \( 2M \).

(2) Is \( M = N \)? If so, set \( \hat{A}_0 = \hat{A}(M) \) and halt. \( \hat{A}_0 \) is then the initial reproduction alphabet for the \( N \)-level quantization algorithm. If not, run the algorithm for an \( M \)-level quantizer on \( \hat{A}(M) \) to produce a good reproduction alphabet \( \hat{A}_0(M) \), and then return to step (1).

For detailed discussion of above algorithms, see [61].
Appendix C

Bayes’ Theorem for Statistical Classifier Design

Consider an N-dimensional classification problem with a feature vector \( X = [x_1, x_2, \ldots, x_N] \). For each class \( \omega_j, \ j = 1, \ldots, m \), assume that the conditional probability density function of \( X \), \( p(X|\omega_j) \), and the probability of occurrences of \( \omega_j \), \( P(\omega_j) \), are known. On the basis of the a priori information \( p(X|\omega_j) \) and \( P(\omega_j), \ j = 1, \ldots, m \), the fundamental function of a classifier is to perform the classification task by minimizing the risk resulting from its decisions (in the simplest case, the probability of misrecognition).

Let \( L_X(\omega_i, d_j) \) be the loss incurred by the classifier if the decision \( d_j \) is made when the input vector, \( X \), is actually from \( \omega_i \). The conditional loss (also called conditional risk) is

\[
r(\omega_i, d) = \int_{\Omega_X} L_X(\omega_i, d)p(X|\omega_i)dX \quad (C.1)
\]

where \( \Omega_X \) is the N-dimensional sample space. For a given set of a priori probabilities \( P = \{P(\omega_1), P(\omega_2), \ldots, P(\omega_m)\} \), the average loss (or average risk) is

\[
R(P, d) = \sum_{i=1}^{m} P(\omega_i)r(\omega_i, d) \quad (C.2)
\]

Substitute equation (C.1) into equation (C.2) and let

\[
r_X(P, d) = \sum_{i=1}^{m} \frac{L_X(\omega_i, d)P(\omega_i)p(X|\omega_i)}{p(X)} \quad (C.3)
\]

then equation (C.2) becomes

\[
R(P, d) = \int_{\Omega_X} p(X)r_X(P, d)dX \quad (C.4)
\]

\( r_X(P, d) \) being defined as the a posteriori conditional average loss of the decision \( d \) for given feature measurements \( X \).
The problem is to choose a proper decision $d_j$, $j = 1, \ldots, m$, to minimize the average loss $R(P, d)$. The optimal rule which minimizes the average loss is called the Bayes’ rule.

From equation (C.4), it is clear that $R(P, d)$ can be minimized if for each particular feature measurement $X$ we choose a decision $d^*$ so as to minimize the a posteriori conditional average loss $r_X(P, d)$. That is, for a given $X$, we choose $d^*$, so that

$$r_X(P, d^*) \leq r_X(P, d_j) \quad j = 1, 2, \ldots, m$$ (C.5)

or,

$$\sum_{i=1}^{m} L_X(\omega_i, d^*) P(\omega_i)p(X|\omega_i) \leq \sum_{i=1}^{m} L_X(\omega_i, d_j) P(\omega_i)p(X|\omega_i) \quad j = 1, 2, \ldots, m$$ (C.6)

A simple yet very useful loss function is the (0,1) loss function, i.e.,

$$L_X(\omega_i, d_j) = 1 - \delta_{ij} = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases}$$ (C.7)

In such cases, the average loss is essentially the probability of misrecognition. The Bayes’ decision rule then becomes

$$d^* = d_i, \text{ i.e., } X \sim \omega_i \text{ if } P(\omega_i)p(X|\omega_i) \geq P(\omega_j)p(X|\omega_j) \text{ for all } j = 1, 2, \ldots, m$$ (C.8)

This is the decision rule we used in our classifier designs in Chapter 6.
The HMM Computation and Training Algorithms

First we define a set of more formal notations. An HMM can be defined as $\lambda = (A, B, \pi)$. $A$ is the state transitional probability distribution matrix, defined as

$$A = \{a_{ij}\}, \quad 1 \leq i, j \leq N$$

where $N$ is the number of states in the HMM and $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$, i.e., the probability of entering state $S_j$ at time $t + 1$, given the HMM is in state $S_i$ at time $t$. The observation symbol probability distribution matrix, $B$, is defined as

$$B = \{b_j(k)\}, \quad 1 \leq j \leq N, \quad 1 \leq k \leq M$$

where $b_j(k) = P(v_k \text{ at } q_t = S_j)$, i.e., the probability of emitting the $k^{th}$ observation symbol, $v_k$, in the finite alphabet of $M$ distinct symbols, given the HMM is in state $S_j$ at time $t$. The initial state distribution $\pi = \{\pi_i\}$, where $\pi_i = P(q_1 = S_i), 1 \leq i \leq N$.

Before presenting the solutions, we list below once again the three general problems in HMM operations and training:

Problem 1— Given the observation sequence $O = O_1 O_2 \cdots O_T$, and a model $\lambda = (A, B, \pi)$, how to efficiently compute $P(O|\lambda)$, the probability of generating the observation sequence from the model?

Problem 2— Given the observation sequence $O = O_1 O_2 \cdots O_T$, and a model $\lambda = (A, B, \pi)$, how to determine a state sequence $Q = q_1 q_2 \cdots q_T$ which is optimal in some meaningful sense?

Problem 3— How to adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $P(O|\lambda)$?
Solution to Problem 1

The method is called forward-backward procedure. However, for solving Problem 1, either the forward part, or the backward part, of the algorithm is needed.

Define the forward variable $\alpha_t(i)$ as

$$\alpha_t(i) = P(O_1 O_2 \cdots O_t, q_t = S_i | \lambda)$$

i.e., the probability of the partial observation sequence, $O_1 O_2 \cdots O_t$, (until time $t$) and state $S_i$ at time $t$, given the model $\lambda$. We can solve for $\alpha_t(i)$ inductively, as follows:

1. Initialization:

$$\alpha_1(i) = \pi_i b_i(O_1), \quad 1 \leq i \leq N. \quad (D.4)$$

2. Induction:

$$\alpha_{t+1}(j) = \left[ \sum_{i=1}^{N} \alpha_t(i)a_{ij} \right] b_j(O_{t+1}), \quad 1 \leq t \leq T - 1, \quad 1 \leq j \leq N \quad (D.5)$$

3. Termination:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \quad (D.6)$$

The computation requirement of the above procedure is in the order of $N^2 T$ calculations.

Solution to Problem 2

Different way of defining the “optimal” state sequence associated with the given observation sequence leads to different solution to this problem.
The most widely used optimization criterion is to find the single best state sequence (path), i.e., to maximize $P(Q|O, \lambda)$. A formal technique for finding this best path exists, based on dynamic programming methods, and is called Viterbi algorithm.

First we need to define the quantity

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P(q_1 q_2 \cdots q_t = i, O_1 O_2 \cdots O_t | \lambda)$$  \hspace{1cm} (D.7)

i.e., $\delta_t(i)$ is the best score (highest probability) along a single path, at time $t$, which accounts for the first $t$ observations and ends in state $S_i$. By induction we have

$$\delta_{t+1}(j) = \left[ \max_i \delta_t(i) a_{ij} \right] b_j(O_{t+1})$$  \hspace{1cm} (D.8)

To actually retrieve the state sequence, we need to keep track of the argument which maximized equation (D.8), for each $t$ and $j$. We do this via the array $\psi_t(j)$. The complete procedure can be stated as follows:

1. **Initialization:**
   $$\delta_1(i) = \pi_i b_i(O_1), \quad 1 \leq i \leq N$$
   $$\psi_1(i) = 0.$$  \hspace{1cm} (D.9)

2. **Recursion:**
   $$\delta_t(j) = \max_{1 \leq i \leq N} \left[ \delta_{t-1}(i) a_{ij} \right] b_j(O_t), \quad 2 \leq t \leq T$$
   $$\psi_t(j) = \arg \max_{1 \leq i \leq N} \left[ \delta_{t-1}(i) a_{ij} \right], \quad 2 \leq t \leq T$$  \hspace{1cm} (D.10)

3. **Termination:**
   $$P^* = \max_{1 \leq i \leq N} [\delta_T(i)]$$
   $$q^*_T = \arg \max_{1 \leq i \leq N} [\delta_T(i)].$$  \hspace{1cm} (D.11)
4. Path (state sequence) backtracking:

\[ q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T - 1, T - 2, \ldots, 1. \tag{D.12} \]

At the end of the procedure, the best path (state sequence) is found as \( Q^* = \{q_1^*, q_2^*, \ldots, q_T^*\} \), and \( P^* \) is the probability of the model \( \lambda \) taking the best state sequence \( Q^* \) and emitting the observation sequence \( O \).

**Solution to Problem 3**

In a manner similar to equation (D.3), we first define a **backward variable** \( \beta_t(i) \) as

\[ \beta_t(i) = P(O_{t+1}O_{t+2}\cdots O_T | q_t = S_i, \lambda) \tag{D.13} \]

i.e., the probability of the partial observation sequence from \( t + 1 \) to the end, given state \( S_i \) at time \( t \) and the model \( \lambda \). We also need to define \( \xi_t(i, j) \), the probability of being in state \( S_i \) at time \( t \), and state \( S_j \) at time \( t + 1 \), given the model and the observation sequence, i.e.,

\[ \xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda). \tag{D.14} \]

From the definitions of the forward and backward variables, we can write \( \xi_t(i, j) \) in the form

\[
\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)} \\
= \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)} \tag{D.15}
\]

where the numerator term is, in fact, \( P(q_t = S_i, q_{t+1} = S_j, O|\lambda) \) and the division by \( P(O|\lambda) \) gives the desired probability measure. Another definition necessary for solving Problem 3 is
the probability of being in state $S_i$ at time $t$, given the observation sequence and the model, $\gamma_t(i)$. $\gamma_t(i)$ can be associated with $\xi_t(i,j)$ as

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j).$$  \hfill (D.16)

The follow interpretations are necessary to the solution we are going to present later,

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{expected number of transitions from } S_i$$  \hfill (D.17)

$$\sum_{t=1}^{T-1} \xi_t(i,j) = \text{expected number of transitions from } S_i \text{ to } S_j.$$

Using Equations (D.15)-(D.17), we can have a set of reasonable re-estimation formulas for $\pi$, $A$, and $B$ as follows:

$$\tilde{\pi}_i = \text{expected frequency (number of times) in state } S_i \text{ at time (} t = 1 \text{)}$$

$$= \gamma_1(i)$$

$$\tilde{a}_{ij} = \frac{\text{expected number of transitions from state } S_i \text{ to state } S_j}{\text{expected number of transitions from state } S_i}$$

$$= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$  \hfill (D.18)

$$\tilde{b}_j(k) = \frac{\text{expected number of times in state } j \text{ and observing symbol } v_k}{\text{expected number of times in state } j}$$

$$= \frac{\sum_{t=1, O_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}.$$

equation (D.18) gives us a re-estimated model $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$. It has been proven by Baum et al. [3, 4] that either 1) the initial model $\lambda$ defines a critical point of the likelihood
function, in which case $\lambda = \lambda$; or 2) model $\lambda$ is more likely than model $\lambda$ in the sense that $P(O|\lambda) > P(O|\lambda)$, i.e., we have found a new model $\lambda$ from which the observation sequence is more likely to have been produced.
Appendix E

Instruction to the Parents in the Infant LOD Rating Experiment

Dear participants:

We are working in a project which is aimed to use computer technology to automatically evaluate the degree of distress of an infant by analyzing his/her cry sound. In the tape, we recorded dozens of cries uttered by different babies. We would like you to listen to each of them and grade your feeling about the baby’s distress in 5 levels. Level 1 can be used to indicate that you feel the baby is crying with little distress, while level 5 can be used to indicate that the segment of cry sounds to you as if the infant is trying to express the greatest distress. Level 2 to level 4 can be used if you feel the baby’s distress situation is between the above two extremes.

Please also note:

1. You can, and are highly encouraged, to listen to the cry recording repeatedly if you find it is necessary for you to make your assessments.
2. Please make your assessment based on ONLY the segment of cry you just listened, but NOT try to predict the tendency of the baby’s distress situation before or after the recorded cry segment.

Thank you very much for your participation.

Qiaobing Xie
Graduate Student, EE Department, UBC
Bibliography


