National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

Improved Processing and Classification Techniques for Infant Cry Vocalizations

Marco Petroni

B. Eng. (Electrical Engineering, McGill University) 1989,
M. Eng. (Electrical Engineering, McGill University) 1991.

Department of Electrical Engineering
McGill University
Montréal
May, 1995

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

ISBN 0-612-05774-7

Canada

Ad Majorem Dei Gloriam

Abstract

Advances in cry research and understanding have been limited due to the lack of available analysis and classification methods which can adequately deal with the particulars of this simple, yet effective, communication medium. This thesis presents new processing and classification methods for infant cry signals. First, a new method of accurately extracting the vocal fundamental frequency from cry signals is proposed. This multi-step crosscorrelation vector-based method accurately tracks rapid changes in the fundamental frequency in these utterances, is not limited to any particular range of pitch values, and allows a more detailed view of this important parameter for further analysis. The benefits of this method are not limited to infant cry vocalizations, however. This new method can be employed by any application that requires accurate and detailed pitch extraction, as well as being suitable for pitch synchronous analysis of a voiced signal. Then, a novel application of artificial neural networks is presented: the automatic classification of anger, fear, and pain cries. A comparison of five different input data sets derived from two different parametric representations, applied to four different neural network architectures is presented. From the classification rates obtained, the use of artificial neural networks would seem well suited to the classification of these types of infant cries and warrants future investigations. Some future work is outlined prior to the concluding remarks outlining the contributions of this dissertation.

Sommaire

Les progrès dans le domaine de la recherche et de la compréhension des cris des nouveaux-nés étaient limités à cause d'un manque de méthodes qui pouvaient traiter adéquatement ce simple moyen de communication qui se trouve à être efficace pour attirer de l'attention aux besoins du bébé. Cette dissertation présente des nouvelles méthodes pour le traitement et la reconnaissance des cris de bébés. En premier lieu, une nouvelle méthode pour l'extraction de la fréquence fondamentale de ces signaux est proposée. Cette nouvelle méthode, consistuée de plusieurs étapes, est basée sur les vecteurs de corrélations croisés qui servent à suivre l'évolution des valeurs de la fréquence fondamentale avec une grande précision dans ces signaux. En plus, cette méthode n'est pas limitée à une portée fixe de valeurs de fréquences fondamentales, et peut aussi permettre une vue plus détaillée de ce paramètre qui peut servir pour d'autres analyses. Les bénéfices de cette méthode ne se limitent pas aux cris. La méthode peut être employée par toutes applications qui demandent une analyse précise et détaillée de la fréquence fondamentale, ou qui demandent une analyse des signaux synchronisé au début de la période fondamentale. Ensuite, une nouvelle conception et mise en application des réseaux neuroniques pour la classification des cris fâchés, de peur, et de douleur sera présenté. Cinq différents groupes de données qui dérivent de deux représentations parametriques différentes du signal sont présentées à quatre architectures différentes et comparées. Suite aux résultats obtenus de ces expériences, les réseaux neuroniques semblent bien classifier ces types de cris et encouragent une l'expérimentation continuée sur ce système de classification. Les grandes lignes des expansions futures sont discutées avant de conclure.

Acknowledgements

At the conclusion of this long, but exciting process, my first thoughts go out to my parents. It was they who provided me with the support and understanding I needed to get through this experience in graduate studies. To them, I owe everything. I am forever grateful for their help, and for showing me that hard work and perseverance have their rewards. The financial support of NSERC is also gratefully acknowledged.

Thanks must also be expressed to my supervisor, Alfred Malowany, for allowing me to discover the interesting field of Graduate Studies, and for the inspiring discussions that assisted me to the very end. I would also like to acknowledge the help and input of Prof. C. Celeste Johnston. I thank her for allowing me to discover the exciting and challenging domain of infant cry analysis, and for sensitizing me to the issues and considerations specific to infants and their perception of pain; our collaboration has been both enriching and exciting. Thanks also goes to Prof. Bonnie Stevens, a former PhD student of Prof. Johnston's who is now on staff at the University of Toronto, who along with Prof. Johnston, recorded the cry signals which were used in this study.

A special thanks goes out to the other members of my PhD committee, namely, Prof. Howard C. Lee, for his input, and to Prof. Renato De Mori, for his frequent and insightful comments, and for his guidance.

Thanks also goes out to the system staff of the Center for Intelligent Machines (CIM) Jan Binder, Steve Robbins, and Mike Parker, who have kept the computers up and running through thick and thin, and who have been responsible for maintaining an excellent computing environment.

Another special thanks to all my fellow students who, over the years have

passed through the CIM, and also a very special thanks to the "veterans", who have all passed through room 463, especially fellow PhD students Damian Haule, Majid Noorhosseini, and Kumbesan Sandrasegaran. They deserve praise for their help, suggestions, and friendship.

Last, but not least, an extra special thanks goes out to my girlfriend Joyce Di Turi for her patience, support, and understanding.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

As is the case for the newborn offspring of other mammals in the animal kingdom, the human infant is very helpless and defenseless, relying on its parents to tend to its needs. For the human infant, these needs typically consist of feedings, diaper changes, affection, and, in some cases, prompt medical attention due to a possibly unknown problem. Unlike older infants, neonates do not possess the command of language, and thus must rely on other methods to signal their needs to their care-giving environment. The most common and most primal of these signalling methods are cry vocalizations.

These vocalizations are a very effective means of eliciting a response from care-givers, who often judge their effectiveness in tending to the infant's needs by gauging how soon the crying stops after care is administered [Donovan and Leavitt, 1985b]. Due to the complex neurological and physiological processes involved in cry production, it is thought that a more subtle form of information may be contained in these vocalizations in addition to simply serving to attract attention [Lester, 1984, Porter *et al.*, 1986].

In the vast majority of cases, parents of infants learn over time to distinguish between the different types of cries of their infant. From the cry, they can determine whether the infant is hungry, hurt, or just wants to be held [Golub and Corwin, 1985]. This knowledge is also applicable across infants as well since it has been observed that after parents have "learned" to identify the meanings behind the various types of cries from their infant, they do better at distinguishing between the cries of other infants than do other adults [Lester and Boukydis, 1985]. This observation thus leads to the belief that there are similar or common features present in cries uttered by infants who are in the same state, be the state hunger,

pain, or fussy, for example.

It is not only parents that learn to distinguish between different types of cries, however. Studies have shown, that in a clinical setting, the cries of healthy, or so-called "normal" infants, can be differentiated from those which have genetic disorders, such as Down's Syndrome or 15-15 trisomy, for example, or from those which have had traumatic birth histories [Zeskind and Lester, 1978, Lind *et al.*, 1970] Typically, the latter cries are said to be "harsher" sounding than the former. Consequently, from these auditory discriminations, a crude, simple, yet accurate determination of pathological diagnosis can be made of an infant's state or condition, based on the characteristics of the cry utterance. If a precise diagnosis cannot be made from a cry utterance, one might at least say that something is wrong, prompting a further and more detailed medical examination of the infant.

The idea of listening in order to assist in the determination of pathology, also known as *diagnostic listening*, is not a new one, however, and dates back to about 400 B. C. when it was originally proposed by Hippocrates [Golub and Corwin, 1985]. Over 2000 years passed before this idea re-surfaced in the domain of infant crying. In the latter part of the 19th century, Charles Darwin, the father of the theory of Natural Selection, treated the subject through a series of drawings and descriptions of different types of infant cries uttered in different situations [Darwin, 1872].

More recently, however, a number of research groups have attempted to determine the discriminating features in the cry in the hope that these differences could be quantified for the eventual development of an automated classification system [Wasz-Höckert *et al.*, 1985, Johnston and O'Shaughnessy, 1988, Benini *et al.*, 1993]. This work arose from earlier work done on the analysis of infant cry signals. Although some success has been achieved in visually or auditorily discriminating between different cry types, this information has yet to be used in an automated system.

The work done to date on the analysis of infant cries which has been docu-

mented in the literature, has used methods borrowed from the domain of speech processing, since the cry can be considered to be a form of speech, and since the mechanisms which produce a cry are similar to those which produce speech [Petroni *et al.*, 1994a]. Although these methods borrowed from the speech processing domain have opened the door for work to be undertaken in this domain, they are not always useful on all types of cry vocalizations due to the peculiar characteristics of certain vocalizations. Consequently, no one method yet exists which can correctly and consistently deal the full spectrum of infant cries.

This is especially true when dealing with the analysis and treatment of the vocal fundamental frequency (F_0). This parameter, also referred to as the pitch, represents the rate at which the vocal folds, located in the larynx, vibrate during voiced portions of the signal. A number of studies have determined that vocal fundamental frequency, and its progression over time and over the length of an utterance are important indicators of both infant state and neurological organization not only in infants, but in adults as well [Anand *et al.*, 1989, Colton and Steinschneider, 1980, Fuller, 1991, Hollien, 1980]. Despite the importance of F_0, none of the methods borrowed from the speech domain, adequately deal with the wide range of infant vocal fundamental frequency values, which can go from values as low as 150 Hz to values of over 2500 Hz. It should be noted that adult speech has F_0 values which typically fall below 600 Hz.

Ideally, a method specifically, tailored for correctly dealing with this range of F_0 values would have to be used for the purposes of correct extraction. This F_0 extraction method should be able to track the progression of the fundamental frequency in cry utterances which undergo rapid changes and during events such as double-harmonic break episodes. These events are common in certain types of cries, such as pain cries, and it is important that these events be properly handled. As well, a method which would be capable of producing a value of F_0 for every pitch period in the signal would be desirable, so that the prosodic progression of F_0 could be accurately tracked. This would enable the subsequent identification

and quantification of the specific F_0 characteristics of different types of cries. Other potential benefits of this method would be to perform *pitch-synchronous* extraction of other parameters, such as the formant values, for example, in addition to the aforementioned F_0 analysis [O'Shaughnessy, 1987, Medan and Yair, 1989].

Although the identification and extraction of parameters from cry vocalizations is important, it would also be desirable, to determine a set of features which could be used to accurately classify or discriminate between different infant states or pathological conditions. The cry is a readily available and non-invasive parameter, the latter making it particularly appealing for use in a clinical setting, as was previously mentioned, especially if certain clinical situations are considered. For example, such a classification system would be useful for care givers who are responsible for a number of infants at any one given time. When tending to one infant, they would be able to identify whether another infant who starts to cry is doing so because it requires immediate attention resulting from a problem, or is crying simply to relieve some stress. An automated classification system could then be useful to assist a care giver in determining if the needs of the infant who is crying are greater or require more prompt attention than those of the current infant being tended to before leaving the current infant in order to tend to the crying infant. Also, there have been documented cases where all pathological signs in a certain infant are normal, but an abnormal sounding cry is present [Zeskind and Lester, 1978]. In these cases, an automatic cry classification system could alert physicians and indicate that an infant requires a more detailed observation of his or her condition, and that something may be wrong.

Prompt medical attention is especially important in the development of a newborn infant [Keating, 1980]. Also, prompt identification of infants who are said to be *at risk*, due to a traumatic birth or a low birth weight, for example, can lead to a faster and more successful medical treatment which will enable the infants to proceed along a normal path of development in the shortest delay possible.

This dissertation attempts to address the issues of processing and classification in the hope that the methods and the results presented here will bring the state of the art in the cry domain one step closer to achieving the goals of accurate fundamental frequency extraction and correct classification of infant state from the cry signal.

This chapter has given a brief introduction to the cry analysis and classification problem statement and has mentioned the importance and the potential applications of this work.

Chapter 2 presents background information on cry analysis from its early origins to the presentation and discussion of more recent developments. Methods borrowed from the speech domain which have been used to date on cries and documented in the literature are outlined. The motivations for looking at specific parameters in the cry signal are also presented. In particular, the extraction of vocal fundamental frequency from both speech and cries will be discussed. The chapter concludes with a presentation on automatic classification methods used in speech and on cries.

Chapter 3 addresses the extraction of one parameter of particular interest from the cry signal, namely, the extraction of vocal fundamental frequency. This chapter begins with a detailed presentation of a new and accurate F_0 extraction method especially suited for infant cry signals. The results of this method are then compared to those obtained from other F_0 extraction routines adopted from the speech domain and tested on cry utterances. Computational considerations, potential benefits, and spin-offs of the new method are discussed as well.

Chapter 4 focuses on experiments performed on the automatic determination of infant state using a number of different artificial neural networks (ANNs) architectures and learning methods. The merits, strengths, and weaknesses of the different methods, and of different input feature sets are discussed.

Chapter 5 focuses on future work which can be undertaken as a result of the

work presented in this dissertation, in both the analysis and classification domains.

Concluding remarks, including a concise presentation of the contributions of this dissertation are presented in chapter 6.

Chapter 2 Background and Related Work

This chapter presents the background and related work of topics which are directly related to this dissertation. First, the topic of cry analysis will be presented, from its early treatments to the more recent analysis of this signal and its features. In this section, the motivations for looking at the cry and the information which has been extracted to date from this signal, will be outlined. As well, some recent attempts at augmenting the cry with other parameters for the purposes of accurate classification of infant state will be presented. Following this, the evolution of the analysis methods of a particular parameter in the cry signal, namely the vocal fundamental frequency, will be addressed. The methods of choice for the analysis of this parameter in the speech domain will be presented. The last section of this chapter addresses the issue of neural network-based classification, outlining the methods developed by other research groups to deal with the classification of time-varying signals such as speech.

2.1 Cry Analysis and Classification

As was mentioned in chapter 1, the idea behind the examination of the infant cry for the purposes of determining the state of an infant is not new. The development and proliferation of less expensive computers with stronger computational engines, coupled with more sophisticated signal processing, classification, and visualization techniques will most likely provide the impetus for future research in the analysis and classification of infant cries. However, much about the cry has been learned from research performed over the past 30 to 40 years, and the advances in the analysis and understanding of the cry expression has paralleled the development

of techniques which have assisted in its recording and analysis.

Significant research activity in this domain began at the turn of the century with the advent of devices which allowed the permanent recording of sounds. At that time, two German investigators noted that certain infants in their test group had notably higher pitch values in their utterances than did other infants [Flatau and Gutzmann, 1906]. Later, as tape recorders appeared in the 1920s, research in this domain began to spread. The analysis and classification techniques used during this initial period of work undertaken in this domain, consisted exclusively of auditory methods. One particular researcher attempted to determine the meanings behind cries using auditory techniques to identify relevant sounds in certain utterances [Sherman, 1927]. Other researchers also began to focus on the identification of particular features in the cry which would allow certain types of cries to be differentiated from each other. Fairbanks, in 1942, published an article detailing his studies of the fundamental frequency, or pitch, of hunger vocalizations [Fairbanks, 1942]. Even in these initial studies into the analysis of infant cries, the fundamental frequency emerged as an important parameter for discrimination.

The development of sound spectrograms in the late 1940s prompted increased interest into this domain and a number of research groups used this tool to further advance the understanding of cries. Spectrograms gave researchers the opportunity to visually identify features of certain types of cries. In 1968, a Scandinavian group published a book detailing their research efforts in attempting to identify relevant features in the cries of both healthy and sick infants [Wasz-Höckert *et al.*, 1968]. When these first studies attempted to identify certain features in cry utterance using the spectrogram, they discovered that there was no nomenclature available for certain types of cries. Consequently, much of the definitions and description of many cry characteristics were developed in the 1960s.

This monograph by Wasz-Höckert, Lind, Vuorenkoski, and Partanen provided a

comprehensive presentation of the spectrograms and harmonic patterns in the cries of infants with genetic or pathological disorders, as well as for hungry infants or infants in pain. Prior to the publication of this book, however, Truby and Lind had identified three types of pain cries and classified them according to the fundamental frequency values [Truby and Lind, 1965]. The first type was referred to as the basic cry, and contained F_0 values between 200 Hz and 600 Hz. The second type was the turbulent, or disphonation, utterance which was caused by an overloading of the vocal tract, resulting in aperiodic vibrations of the vocal folds. The third type of pain cry was hyperphonation, and contained F_0 values between 1000 Hz and 2000 Hz, which was thought to correspond to extreme distress.

In this early period of cry research, there was no real focus of groups in general to tackle a particular cry type. Researchers would publish observations which would serve in the future as a stepping stone to improving the observations made previously, or proposing the correlation between certain physiological effects and certain attributes in the cry. In addition to the aforementioned researchers, Parmelee [Parmelee, 1962] noticed that there were certain differences between the cries of healthy infants, and those which suffered from neurological disorders. Also, Bosma, Truby, and Lind stated that neurological maturity is revealed by the stability of laryngeal coordination and vocal tract mobility, since the production of the vocalizations involves varying control of vocal articulators [Bosma *et al.*, 1965]. This tie between neurology and attributes would prove to be important in future studies of neurological stability and central nervous system insult, for example, and cry attributes [Anderson-Huntington and Rosenblith, 1976]. The latter study also showed that these abnormal cries could be used as indicators of future developmental problems.

In 1959, Davis published an article noting that the human auditory system is particularly sensitive to frequency values about 800 Hz, which in turn implies that humans are particularly sensitive to cries with high fundamental frequency values [Davis, 1959]. By the late 1960s, a number of studies quoted that fundamental

frequency and duration were important parameters for the discrimination of cry types. Insofar as adult responses to cries were concerned, Korner and Goldstein noticed that the cry would elicit attention and a visual scanning of the infant for further indications as to why the infant was crying [Korner and Grobstein, 1966].

Aside from an eliciting attention from adults, it was noticed that certain types of cries had specific harmonic patterns associated with them, and that these cries could also be differentiated auditorily [Partanen *et al.*, 1967]. The recordings of this particular study included cries of infants with asphyxia, brain damage, hyperbilirubinemia, and Down's syndrome. The studies published in and around that time could not provide quantifiable measures of pain, pathology, or genetic problems from spectrograms, due to the lack of available methods for extracting measures from cry signals. Consequently, only non-parametric statistical analysis was performed at this time.

One of the results of this study was the development of a device called the "Cry Analyzer" which would screen cries obtained from a neonatal ward [Vuorenkoski *et al.*, 1970]. On the cry utterances that would differ from the normal types of cries, further spectrographic analysis would be performed. This device recorded the fundamental frequency of the cry as well as the heart and respiration rates of the infant. An objective evaluation of the Cry Analyzer indicated that its use in practice would be limited and subsequent studies using this device were discontinued. However, this method was a first attempt at automating the acquisition of "interesting" cry recordings for further analysis.

Along with the identification of fundamental frequency related patterns in certain types of cries, Ostwald, Phibbs, and Fox showed that elements in the cry such as fundamental frequency and utterance duration could also be useful in predicting the occurrence of future health problems, or to provide an "early-warning" for certain disease types [Ostwald *et al.*, 1968]. Of these two features, fundamental frequency was found to have the most reliable diagnostic value.

In the 1970s, cry analysis branched out into the particular study of the cries of infants with other conditions, and to the study of factors which influence the cry as well. Lieberman, Harris, Wolff and Russel [Lieberman *et al.*, 1971] compared the characteristics and communicative significance in the cries of human infants with those of non-human primates.

Katarina Michelsson continued research in the identification and analysis of the relevant characteristics of certain types of cries. In one particular study, Michelsson, Sirvio, Koivisto, and Wasz-Höckert investigated the cry and fundamental frequency characteristics of pain cries of neonates both with and without feeding tubes [Michelsson *et al.*, 1974]. A couple of years later, Michelsson and Sirvio examined the cries of infants with congenital hypothyroidism [Michelsson and Sirvio, 1976] and determined that the fundamental frequency characteristics of these cries differed from those of healthy infants. The following year, Michelsson, Sirvio, and Wasz-Höckert published other articles outlining the difference between the fundamental frequency and duration of pain cries of healthy and asphyxiated infants [Michelsson *et al.*, 1977a], and in cries of healthy infants and of those with bacterial meningitis [Michelsson *et al.*, 1977b].

Moreover, Michelsson, along with Juntunen, and Sirvio, performed sound spectrographic investigations of infants with severe malnutrition [Juntunen *et al.*, 1978]. They found that the maximum and minimum fundamental frequency values of the cries of malnourished infants were higher than those of healthy infants. Lester had published the results from a similar study two years earlier [Lester, 1976]. Both studies concluded that the spectrogram could indeed be a useful tool in assisting the determination of the level to which the brain is affected by malnutrition.

The spectrographic and auditory studies of cries uttered by infants with different conditions by Katarina Michelsson continued in association with other researchers. Thodén and Michelsson illustrated the difference in fundamental frequency values between healthy infants and those with Krabbe's disease spectrographically

[Thodén and Michelsson, 1979]. The compilation of the spectrographic cry analysis studies, in which Michelsson participated in, were presented in 1980 when a book on the significance of infant communication using cry and early speech was published [Murry and Murry, 1980]. This book also presented the current state of cry analysis research in addition to the results of the aforementioned spectrographic studies. In one particular chapter of this book, Michelsson illustrated the spectrographic differences in the fundamental frequency values and harmonic patterns of cries uttered by infants with physiological or genetic problems [Michelsson, 1980]. In another article, the significance and potential benefits of cry analysis for determining the probability of an infant at risk or of asymptomatic infants with neurological problems [Michelsson and Wasz-Höckert, 1980].

In this same book, two other members of this Scandinavian research group published an article detailing the acoustic attributes of pain cries in normal infants as seen in a spectrogram [Thodén and Koivisto, 1980]. Here, Thodén and Koivisto detailed how certain features, fundamental frequency values, and harmonic structure were present in pain cries, but not in non-pain cries.

Following the publication of these articles, research into identifying differences between healthy infants and those with other pathological or genetic conditions continued for this particular researcher. Michelsson, Tuppuranien, and Aula [Michelsson *et al.*, 1980] noticed that infants with an abnormality of chromosome 4 or 5 had cries with significantly higher fundamental frequency values than those of healthy infants. As well, infants with "Cri-du-Chat" syndrome had flat and monotone melody types, and that infants with 13- or 18-trisomy had hoarse and low-pitched cries. The infants with chromosomal abnormalities also had different pain cries than other infants who suffered from central nervous system disorders. These researchers concluded that the cry can be a useful indicator of chromosomal abnormalities.

Raes, Michelsson, and Dehaen later published an article where a number of

spectrographic characteristics were compared between the pain cries of healthy infants and infants with infectious or congenital disorders of the larynx, once again noting that certain spectrographic features occur more often in cries of infants with central nervous system disorders. The results of a similar study conducted on infants with congenital hydrocephalus, cerebral malformations, and healthy infants was published two years later [Michelsson *et al.*, 1984]. This study also noticed that certain spectrographic features and melody types were common to certain types of cries.

The mid-1980s saw the publication of yet another collection of cry analysis articles dealing in a wide variety of topics of interest to this domain [Lester and Boukydis, 1985]. Here, Wasz-Höckert, Michelsson, and Lind described the research undertaken by Scandinavian researchers over the past 25 years [Wasz-Höckert *et al.*, 1985], and, as well, new work was presented by Thodén, Järvenpää, and Michelsson on the spectrographic analysis of pain cries in premature infants [Thodén *et al.*, 1985]. This article noted that the more premature the infant, the higher pitched the cry.

Recently, the focus of the research of this group has focused on the crying patterns of infants and adult perceptions of cries. In a recent article, Michelsson, Rinne, and Paajanen noted that the length of crying bouts decreases as an infant gets older, and adult's perceptions of cries also changes over time [Michelsson *et al.*, 1990].

For all the articles published by this group of Scandinavian researchers, the cry features quoted result from a visual analysis of spectrograms and lack precise quantification of particular events in these signals. Nevertheless, this group has perhaps contributed the most in identifying certain acoustic events and fundamental frequency melodies and the occurrence and frequency of these events in different cries over the past 35 years. However, theirs was not the only work undertaken during this time period. A number of research groups analyzed the cry or noticed that there were differences between different cry types.

In 1973 Stark and Nathanson [Stark and Nathanson, 1973] compiled an article detailing the cry attributes and facial gestures present in cries uttered by infants for no apparent reason; that is, cries which were not a result of pathology or following the application of a specific stimulus. Two years later, these researchers published a spectrographic and fundamental frequency analysis of the cries of infants who later died of sudden infant death syndrome (SIDS) noting that there were differences between the spectrograms of normal infants and those with SIDS [Stark and Nathanson, 1975]. These acoustic differences later proved to be contradictory [Colton and Steinschneider, 1980] and inconclusive [Colton *et al.*, 1985].

In a study where the effects of toxic factors on an infant were examined, Ostrea Jr., Chavez, and Strauss, noticed that certain central nervous system manifestations, one of which was a high pitched cry, were present in infants whose mothers used heroin during the last trimester of the pregnancy [Ostrea Jr. *et al.*, 1975]. This observation was also made by Finnegan [Finnegan, 1985]. Lester and Dreher [Lester *et al.*, 1989], noticed that there were durational and other spectral differences between the cries of healthy infants, and those whose mothers used marijuana during pregnancy. Another study noticed that there were differences between the duration of cries, the number of cry utterances, and the number of hyperphonated cries between healthy control infants and those whose mothers took cocaine during pregnancy [Corwin *et al.*, 1992]. The more recent studies state that the effects of narcotics have an effect on the neurological development of the infant and that this is in turn manifested in the characteristics of the cry.

The cry has also been used as an indicator of infant development. Tenold, Crowell, Jones, Daniel, McPherson, and Popper, used cepstral and stationary analysis to determine that there was greater variability in the fundamental frequency and in the spectra of premature infants than those of full-term infants [Tenold *et al.*, 1974]. This study hypothesized that the greater variability observed in the cries of premature infants related to the underlying neurophysiological maturity. Prescott [Prescott, 1975], illustrated that there were differences in the melody of infant cries

in the first two months following birth. The melody of infants during this time showed more variability than it did shortly after birth.

Also, in 1978, Zeskind and Lester published a comprehensive article on infant crying [Zeskind and Lester, 1978]. In this article they stated that high fundamental frequency values were indicators of stressed infants, adding that harmonic and temporal features may also be present in these types of cries. As well, they remarked that neurodevelopmental impairment has also been shown to contribute to acoustic and temporal features in cries. Moreover, infants with serious complications, due to fetal malnutrition, for example, have certain cry-related features, one of which was a high F_0 value. Zeskind and Lester went on to state that certain cry patterns may reflect the risk states of the infant.

Other studies also presented the correlation between development and cry features, as well as commenting on the neurological implications of these findings. Hollien presented some developmental aspects of neonatal vocalizations [Hollien, 1980] and Illingsworth discussed the developmental factors which affect infant vocalizations in the first year of life [Illingsworth, 1980].

In 1984, Lester [Lester, 1984] stated that the characteristics of cries are a direct measure of the integrity of the central nervous system. In the same article, he also proposed his biosocial model of infant crying, which is a neural model of the cry production process. Citing previous studies, Lester presented the idea that neurological maturity is revealed by the coordination and stability of vocal tract articulators. In this article, the most important neural contribution to the production of the cry is the effect of the vagus nerve, which also related to cardiovascular activity and also serves a sensory role for abdominal fullness [Kennedy III and Kuehn, 1989, Lemme *et al.*, 1989]. Lester also states that there are a number of biological factors which affect the cry, and adds that there are social aspects to cry utterances as well; most notably, that the cry serves to attract attention and to prompt a visual scan of the infant by adults to determine why the infant

is crying.

In this article, Lester proposes a binocular view of the cry. First the features in the cry reflect aspects of the neurophysiological function of the infant that are important for later developmental outcome, which also make it useful as a diagnostic tool, and secondly, the cry functions to signal to care givers that the infant is in jeopardy, resulting in a response from the care givers. Lester notes that humans rely on prosodic features to communicate with each other, especially in the first utterances. These aspects overall motivate the more careful examination of vocal fundamental frequency in cry utterances.

Zeskind gives a complete treatment on the developmental aspects of the cry in a later article [Zeskind, 1985]. Another article published in 1989 added to this information stating that the fundamental frequency characteristics change over a two year span, starting from birth [Robb *et al.*, 1989]. Recently, an article by Johnston, Stevens, Craig, and Grunau treated the developmental changes in the pain expressions of premature, full term, two-, and four-month infants [Johnston *et al.*, 1993]. In their study, the cry parameter was augmented by the use of facial expressions in order to determine the behavioural responses to pain stimuli of these various infants. This article noted that higher pitched cries were one of the significant attributes the pain expressions of premature infants in comparison with those of other infants.

In addition to the studies performed by the Scandinavian research group mentioned earlier, a number of other studies also observed that there were correlations between certain genetic disorders and certain cry characteristics [Stallard and Juberg, 1981, Beemer *et al.*, 1984]; the most common attribute being a very high-pitched, or cat-like cry. Other, more recent studies have also noticed that the presence of abnormal or high-pitched cries are accurate indicators of chromosomal abnormality [Murayama *et al.*, 1991, Chernos *et al.*, 1992].

The 1980s also marked a rise in the number of groups studying the ef-

fects of infant cries on human listeners. Gladding [Gladding, 1978] reported on the effects of listener empathy, gender, and training on the identification of infant cries. Donovan [Donovan, 1981] studied the maternal response of mothers of young infants to varying degrees of control over the termination of infant crying, contributing additional insight into this topic four years later [Donovan and Leavitt, 1985b, Donovan and Leavitt, 1985a], and again in 1989 and 1990 [Donovan and Leavitt, 1989, Donovan *et al.*, 1990]. Other researchers noted that some adults may respond negatively to the prolonged exposure to cry utterances, and that there may not be correlations between perceived urgency on the part of the listener and the actual state of the infant [Boukydis, 1985, Frodi, 1985, Murray, 1985].

Other research groups have published their findings related to the presence of abnormal cries with the occurrence of pathological conditions. One research group in England noticed that the presence of a weak cry, accompanied by general weakness and poor feeding, following constipation in a 24-week-old girl, characterized the incidence of botulism [Turner *et al.*, 1978]. These abnormal cry characteristics were also confirmed by another research group 12 years later [Jagoda and Renner, 1990].

In 1982, Golub and Corwin treated the topic of the use of the infant cry for diagnostic purposes [Golub and Corwin, 1982]. In this article they tested the postulation that the infant cry is a reflection of complex neurophysiological functions by using a model of cry production which related the acoustic properties of the signal, to anatomical and physiological characteristics of the infant producing the cry. This initial pilot study was expanded to include more cries for further examination if the cry production model could indeed predict cry utterance parameters occurring as a result of a physiological condition [Golub and Corwin, 1985]. As the desired goal of their long-term study, where different acoustic parameters extracted from the cries are to be examined, the authors hope that the screening of infants using cry analysis for the classification of pathology will become as common in

hospitals as blood tests are.

One major focus of recent efforts from researchers, is the problem in the determination and classification of infant pain, since strong evidence now exists that neonates possess the necessary anatomical functional components for the perception of pain [Anand and Hickey, 1987]. Despite the initial studies which attempted to differentiate between pain and other cries in the late 1960s by the aforementioned group of Scandinavian researchers, the past 10 years has seen a surge in the number of publications which attempt to find the discriminating characteristics of infant pain.

In 1986, two studies were published on the observable effects that pain had on infants. Porter, Miller, and Marshall [Porter *et al.*, 1986] published a study reporting on the cry features observed during various stages of a circumcision procedure. The analysis of these cry recordings from newborn males used spectrograms, and features from these recordings were derived from these spectrograms, such as duration of vocalizations, pitch patterns, and the number of identifiable harmonics, to name just a few of the features examined. It was determined from this study, that the most invasive procedures generated significantly longer crying episodes, higher peak fundamental frequency values, fewer harmonics, and greater F_0 variability. It should be noted that no anesthetic was used in any of these procedures. These cries were also presented to adult listeners for a subjective judgment of the perceived urgency of the cries. Porter, Miller, and Marshall noticed that the cries from the most invasive procedures were judged as being the most urgent. As well, adult listeners all seemed to judge the cries along the lines of harmonic, temporal, and pitch characteristics.

In that same year, Johnston and Strada published an article detailing a description of acute pain response in infants undergoing a routine immunization procedure [Johnston and Strada, 1986]. This study not only examined cry features, as Porter, Miller, and Marshall did, but also looked at other measures such as heart rate, body

movements, and facial expressions. The analysis of the cry recordings in this study was performed using spectrograms. The authors remarked that there was wide variability across spectrographs but that facial expressions remained consistent across infants. Johnston and Strada go on to identify a particular pattern which emerged as a result of an initial response to pain in the heart rate, cry features, body movements, and facial expressions of the infants investigated in the study.

In the following year, Grunau and Craig investigated pain expressions as a result of a heel-lance for blood sampling purposes using measures of facial expressions and cries [Grunau and Craig, 1987]. This study was conducted on infants who were asleep and on infants who were awake at the time of the procedure, in order to gauge if the expression of pain would differ depending on the functional state of the infant. It was discovered that facial expressions differed if the infant was asleep or awake at the time of the heel-lance, but that the fundamental frequency of the cry was not related to this state. Once again, this study used spectrographic techniques to analyze the cries.

Fuller and Horii published an article which attempted to determine an indicator of distress in infant cries in 1988 [Fuller and Horii, 1988]. In this study, the authors cite the promise of using features extracted from four types of cry signals, namely pain, fussy, hungry, and cooing, which were guided from the "stress-arousal framework" which states that levels of stress and arousal in infants will be reflected in the cry characteristics. Some F_0 related features, such as fundamental frequency jitter and amplitude shimmer, showed no variability across cry types, whereas other measures, which were generated to model the "tenseness" of the vocal tract, such as the mean spectral energy of the cries, proved to be a useful tool in differentiating cries. These parameters were extracted from a window-based analysis of the utterances.

Also in 1988, Johnston and O'Shaughnessy published an article where they extracted the position and energy values of the second formant in order to de-

termine if these values, which reflect the excitation and tenseness in vocal tract, could be identified as different physiological responses to different emotional states [Johnston and O'Shaughnessy, 1988]. In this study, pain, fear, and anger cries were examined, with pitch extraction performed using a modified version of the simplified inverse filter tracking (SIFT) method [Markel, 1972b], and narrowband spectrograms were employed to examine parameters such as duration, harmonic structure, and melody. Formant structure was analyzed using wideband spectrograms [O'Shaughnessy, 1987]. The authors found that there was greater intensity and higher second formant frequency values in pain cries, a result which was found to be consistent with the stress-arousal model of cry production mentioned by Fuller and Horii.

In the following year, Johnston published a review article on infant pain assessment and management techniques [Johnston, 1989]. In this article, the relevant physiological measures of pain observed in infants were cited as being cardiovascular and hormonal changes, whereas the relevant behavioural responses to pain were cited as being both facial expressions and cry characteristics. Pharmacological and nonpharmacological methods of controlling pain in infants were also discussed, with the aforementioned measures being used to determine the relative effectiveness of these pain management techniques.

In that same year, Anand, Phil, and Carr, published a comprehensive article on the neurological, anatomical, and chemical processes and responses evoked by pain, stress, and analgesia in infants [Anand *et al.*, 1989]. The changes in these processes and responses as an infant developed were also presented, with a description of the cry as being an important manifestation of the underlying state of the infant.

In 1990, Grunau, Johnston, and Craig focused on the relevance of facial expressions and fundamental frequency and other cry characteristics for the tracking the response of infants to invasive and non-invasive procedures [Grunau *et al.*, 1990].

The cry related features were all extracted using spectrographic analysis.

Although a number of researchers used the cry as one of the parameters investigated as an indicator of pain, few used techniques other than spectrographic ones. Fuller [Fuller, 1991] used measures extracted using a somewhat more elaborate extraction of parameters from the cry signal, using discriminant function analysis to determine the relevance of certain extracted features for the classification of pain, fussy, and hunger cries. The method of choice for coding cry characteristics for the assessment of infant pain still seems to be the spectrogram even if a number of other parameters have emerged in the consideration of relevant parameters measured from infants, such as facial expressions [Maikler, 1991, Benini *et al.*, 1993].

In recent years, however, a number of other parameters are being extracted from cry utterances in order to assist pain determination in a multidimensional parametric representation of features. These parameters include energy values for premature infant pain [Stevens *et al.*, 1994], and formant frequencies [Hadjistavropoulos *et al.*, 1994].

Recent efforts have also focused on the automatic classification of infant state based on the characteristics of the cry. Published efforts at automating this discrimination basically begin with Lundh [Lundh, 1986]. In this article, the author presents the development of a baby alarm which determines the tenseness of the cry, based on energy characteristics of the signal, to characterize happy, crying, and distressed. This device was tested by 10 deaf families who compared the picture illuminated by the device which was meant to convey the state determined by the device, with the actual state of the infant. Although useful, this device generated a number of false alarms.

Seven years later after the development of this baby alarm were documented, Xie, Ward, and Laszlo published a brief article outlining their attempts at classifying an infant's level-of-distress from cry signals using hidden Markov models (HMMs) [Xie *et al.*, 1993]. This measure was determined from several adult perceptions

regarding the aversiveness, or perceived urgency, of several cries. Although the article does not reveal specific implementation details, a correct classification rate of over 80% was quoted for the level-of-distress classification measure for the HMM.

Despite the number of parameters extracted from cry signals and investigated for the classification of different cry types, the most commonly used, and seemingly most relevant from an auditory point of view, are the vocal fundamental frequency and related parameters such as harmonic structure, and melody. These parameters are commonly used to illustrate differences in the cries of healthy and ill infants [Donzelli *et al.*, 1994].

Perhaps the true relevance of the fundamental frequency has been overshadowed by the lack of computerized extraction methods which can adequately deal with cry utterance vocalizations [Petroni *et al.*, 1994a, Petroni *et al.*, 1994b]. This would explain why the majority of researchers still use the spectrogram to determine both the fundamental frequency, and its evolution over the course of an utterance.

The following section will present some of the methods used to extract fundamental frequency from speech signals.

2.2 Fundamental Frequency Extraction

This section will present the background and the previous research undertaken on the extraction of vocal fundamental frequency. Since the overwhelming majority of work done to extract this parameter has been done on speech signals, the literature presented will focus mainly on extraction methods used for this particular class of vocalization.

Vocal fundamental frequency determination methods can be separated into two classes; time-domain methods and frequency-domain methods. Time-domain

techniques have the general advantage over frequency-domain methods in that time-domain methods require much simpler calculations than their frequency-domain counterparts. In addition, these methods allow the location and specification of the pitch epoch times which make these methods suitable for pitch-synchronous formant analysis [Hess, 1983, Medan and Yair, 1989].

On the other hand, frequency-domain fundamental frequency estimation methods segment the input signal into short blocks, also known as frames or windows, and use spectral transformations such as clipping or inverse filtering to extract the fundamental frequency. Typically, the extracted pitch values of these methods are then input to a preprocessor which then corrects for pitch halving or doubling errors due to the mis-labelling of the fundamental frequency with its first harmonic, as will be further discussed in section 3.1.2.

The advent of computers and computerized signal processing techniques in the late 1960s ushered in the start of the development of pitch extraction or pitch determination algorithms. In 1967, Noll published an article outlining a pitch extraction method based on the power spectrum of the logarithm of the power spectrum, called the "cepstrum" [Noll, 1967]. This transformation of the power spectrum effectively causes the source and filter components of the speech signal to be separated. This method was also used to extract the fundamental frequency from infant cry signals some years later [Tenold *et al.*, 1974].

Sondhi published a different pitch extraction method a few months after Noll's article [Sondhi, 1968]. The base methodology of the pitch extraction methods presented in his article was spectral flattening, achieved by clipping a portion of the signal contained within a window at a certain threshold value, and then performing an autocorrelation on the spectrally transformed signal. This method provided a simple, computationally inexpensive, and effective alternative to the cepstral pitch extractor. A real-time hardware implementation of this algorithm along with some other variations such as infinite peak clipping was presented by

Dubnowski, Schafer, and Rabiner [Dubnowski *et al.*, 1976].

In 1972, one of the more popular methods of pitch extraction, formant extraction, and speech coding was presented. John Markel published an article which described a digital inverse filtering method for formant estimation of an input signal window, where the characteristics of the input sequence would correspond, in a least square error sense, to a unit impulse train, with a period corresponding to the pitch period, presented to the filter [Markel, 1972a]. He later published an algorithm which would use this technique to model the spectrum of a low-bandwidth version of the input speech signal, inverse filter the signal window with the predicted filter, and then perform autocorrelation on the inverse filtered signal to determine the fundamental frequency [Markel, 1972b]. This method was called the simplified inverse tracking filter (SIFT) method, and still remains a popular method of pitch extraction from speech signals.

In mid-1973, John Makhoul, Joseph Maksym, and John Markel all published articles detailing different applications of this linear prediction technique [Makhoul, 1973, Maksym, 1973, Markel, 1973]. Makhoul presented an autocorrelation-based method of spectral analysis which approximated the short-time spectrum. Maksym published a pitch extraction method based on the adaptive prediction of the speech window where the prediction error was used as to determine the presence of voicing. The application of the digital inverse filter for both formant and fundamental frequency analysis was described by Markel, who also presented a post-processing method for the determining whether a given input signal window was voiced or unvoiced.

The theory and refined applications of the linear prediction technique for both fundamental frequency and formant extraction were later compiled in a book [Markel and Gray Jr., 1976].

With certain English vowels in speech, it was common to have two formant frequency values within a few hundred Hertz of each other. In the linear prediction

spectra for these particular vowels, these peaks would commonly be merged into one peak, making separation of the two formant frequencies practically impossible. This problem was addressed by McCandless [McCandless, 1974] using the chirp-z transform [Rabiner *et al.*, 1969] to resolve closely spaced formant peaks in the spectra of input signals, if accurate determination of formant values was desired, in addition to the extraction of F_0.

Another computationally simple, yet effective pitch extraction method for speech signals, with an associated decision logic system, was developed with the intention of having a method with characteristics similar to that of the autocorrelation method. This involved taking the absolute magnitude of the difference between the delayed input speech frame and the original at various delays [Ross *et al.*, 1974]. The original appeal behind this method was that it used no multiply operations and the nature of its operations made it suitable for a hardware implementation.

In 1976, the classic article by Rabiner, Cheng, Rosenberg, and McConegal presented the results of seven pitch detection algorithms which were tested on a number of different utterances spoken by adult male, adult female, and child speakers [Rabiner *et al.*, 1976]. A number of error measures were defined and computed in order to determine which algorithms were especially prone to particular errors, and which algorithm generated the best results.

Another method which examined the characteristics of the signal over a short time window was the maximum-likelihood pitch estimation method [Wise *et al.*, 1976]. This method involved sampling the autocorrelation of the input signal frame, and also offered improved resolution of the extracted fundamental frequency values.

The zero-crossing method of extracting pitch was presented by Geckinli and Yavuz [Geckinli and Yavuz, 1977]. This method involved low pass filtering the input speech signal to about 900 Hz so that each pitch boundary would be marked on a zero crossing. Twelve threshold values were used for the decision making process

with two specified threshold values set to the speaker's upper and lower values of a speaker's pitch range. A flowchart of the implementation of the algorithm was also presented by the authors of the above-cited article.

An improvement to the linear prediction method was proposed by Hermansky, Hanson, Wakita, and Fujisaki [Hermansky *et al.*, 1977]. In this article, the authors addressed the limitations of the linear prediction method for fundamental frequency extraction, especially for voices with high fundamental frequency values. The spectrum of the input signal would be transformed by taking the cube root of the power spectrum. Following this transformation, all-pole modeling of the the transformed spectrum would then be performed prior to the inverse filtering and pitch determination process.

Also in 1977, Rabiner published an article describing the use of the autocorrelation function for the purposes of pitch extraction, detailing its limitations and shortcomings in view of certain signal characteristics [Rabiner, 1977]. Despite its limitations this method was cited as having a reasonable performance under low noise conditions.

Friedman proposed a pseudo-maximum-likelihood method of pitch estimation which was based on a sequence of operations [Friedman, 1978]. First, the input signal window was subjected to linear-prediction inverse filtering. Then the inverse filtered signal was subjected to short-time spectral extraction using a bank of bandpass filters with envelope extraction performed on the filter outputs. The determination of pitch was then made using an algorithm which operated on these parallel envelopes, which were considered as a multi-component vector signal.

A real-time pitch detector was developed by Seneff using the spacing between the harmonics in a selective portion of the input spectrum to determine the fundamental frequency of the input signal window [Seneff, 1978]. The spectrum size was limited to an upper frequency value of about 1000 Hz since in the input spectrum, the higher frequency values become ragged and the harmonics become more

difficult to distinguish. The algorithm then used heuristics to extract the candidate harmonics from the spectrum at which point it then proceeded to calculate the fundamental frequency.

The following year, Ananthapadmanabha and Yegnanarayana published their attempts at extracting the epoch from the linear prediction residual [Ananthapadmanabha and Yegnanarayana, 1979]. It should be noted that epoch extraction is particularly useful for accurate pitch extraction since the start of an epoch signals the start of the pitch period. Due to some ambiguities present in the inverse-filtered signal regarding the exact start time of the pitch epoch, this signal is further filtered and processed in order to extract this information.

Matausek and Batalov [Matausek and Batalov, 1980] proposed another approach using the inverse-filtered signal following a covariance-based linear prediction stage to determine the glottal waveform. This process involved integrating the inverse filtered signal and then iteratively inverse filtering this signal in order to obtain a glottal model.

Duifuis, Willems, and Sluyter proposed a pitch extraction method based on a theory of hearing which states that the perception of fundamental frequency is assisted by the fundamental frequency of the spectrum which best fits the spectrum of perceived sounds [Duifhuis *et al.*, 1982]. The method presented by the authors involved subjecting the peaks of the spectrum of an input signal window to thresholding and component masking based on Goldstein's theory of pitch perception [Goldstein, 1973]. Once the spectral peaks were processed by this initial stage, the remaining peaks were "sieved" to determine the most likely fundamental frequency. Some improvements to this method were featured in a later article [Sluyter *et al.*, 1982].

Another method which used a similar method to process the spectral peaks, using a technique called the spectral comb, was proposed by Martin [Martin, 1982]. Here, a spectral correlation was performed on the power spectrum using a spectral

comb with "teeth" of decreasing amplitude and intervals, and the results from this technique were compared to that obtained using the cepstral pitch extraction technique.

Although the methods which had been published up to that time worked well on good quality adult male speech, few had been tested on speech which had unusual characteristics due to aperiodicities in the vibration of the vocal folds or from excessive fundamental frequency jitter and amplitude shimmer. One group published a brief comparison on the performance of a few of the more popular pitch extraction methods, such as the cepstral and SIFT methods, on several speakers with a variety of speech disorders [Laver *et al.*, 1982].

This issue gained some attention in the following years, as researchers determined that the fundamental frequency was an important parameter in determining the presence of laryngeal pathology. Kasuya, Kobayashi, and Kobayashi attempted to describe the pitch period perturbations present in patients with cancer of the vocal cords [Kasuya *et al.*, 1983]. Other attempts were made by Feijóo and Hernández [Feijóo and Hernández, 1985], and Imazumi [Imazumi, 1986] who examined a number of factors based on the characteristics of the pitch periods in utterances of speakers both with and without laryngeal pathology. In a related investigation, Veeneman and BeMent attempted to use inverse filtering to extract the glottal pulse and to determine whether there would be an abnormal glottal volume velocity, the latter being an indicator of pathology [Veeneman and BeMent, 1984].

Chung and Algazi first presented a crosscorrelation-based pitch extractor for the purposes of extracting pitch values from noisy speech, exploiting the high correlation between adjacent pitch segments [Chung and Algazi, 1985]. This method was quoted as performing well in the vicinity of voiced to unvoiced transitions where the local signal-to-noise ratio was low.

A method which is similar to the cepstrum pitch extraction method was proposed by Indefrey, Hess, and Seeser [Indefrey *et al.*, 1985]. In this article, the au-

thors proposed the application of a non-linear distortion in the frequency domain following the computation of the discrete Fourier transform of the input signal window. Then, prior to performing the pitch period determination, the non-linearly distorted spectrum was inverse transformed using the inverse discrete Fourier transform.

Charpentier implemented a method which used phase information from the discrete Fourier transform of an input signal window to extract the harmonic components from the input spectrum and then performed fundamental frequency determination based on the values of the extracted harmonics [Charpentier, 1986]. The use of the Fourier transform phase for extracting the fundamental frequency was also proposed by Brown and Puckette [Brown and Puckette, 1993].

Another linear prediction-based method which attempted to address the issue of short pitch period extraction was presented by Miyoshi, Yamato, Yanagida, and Kakusho [Miyoshi *et al.*, 1986]. These authors stated that the extraction of voiced sounds uttered by children or females could be accurately estimated using sample-selective linear prediction, which employed a two-step linear prediction process. Despite the appeal of the method, which attempted to deal with the issue of linear prediction and short pitch period sounds, the results presented in the paper did not provide a complete or convincing test of this method.

A novel, albeit complicated, pitch extraction method was proposed by Gong and Haton [Gong and Haton, 1987]. This method involved modeling speech as a sequence of a specified function type, referred to as a resemblance function, which allows amplitude and excitation of the signal to be time-varying. This resemblance function is statistically optimized from an energy function, and the pitch period estimate is achieved by the maximization of this function. From a frequency-domain perspective, this method is equivalent to a harmonic matching procedure with results being comparable to those achieved by other harmonic matching methods.

Another method presented in that same year used the spectral autocorrelation to extract pitch from noisy speech signals [Lahat *et al.*, 1987]. In this method, the spectrum of an input signal window was presented to a series of bandpass "lifters" covering the range of expected pitch periods, and then extracting the pitch from autocorrelation functions calculated at the output of the lifters. These extracted values were then presented to a median filter for the smoothing of extraneous pitch values.

Andrews, DeGroat, and Picone published a series of articles outlining their improvement to the classical cepstral-based pitch extraction method called the MUSIC method [Andrews *et al.*, 1989, Andrews *et al.*, 1990b, Andrews *et al.*, 1990a]. In these articles, the authors propose the use of singular value decomposition for estimating the power spectral density of a signal. In this method, singular value decomposition is also used in the place of the fast Fourier transform to estimate the cepstrum for the purposes of accurate pitch extraction in the presence of noise.

Cheng and O'Shaughnessy presented another method for the estimation of the glottal closure instant and period in an attempt to provide pitch estimation technique which could illustrate period-by-period changes in the pitch period [Cheng and O'Shaughnessy, 1989]. The method used a twelve-pole linear-prediction analysis of the input speech signal window, a crosscorrelation, and convolution to generate a non-stationary pitch period estimation.

The estimation of pitch using a set of harmonic sine waves to fit the input data using a mean-squared error criterion was proposed by McAulay and Quatieri [McAulay and Quatieri, 1990]. This method assumes, however, that the sinusoidal characteristics of the signal have already been analyzed by an analysis / synthesis method proposed in a previous paper [McAulay and Qualtieri, 1986]. The article, however, fails to present any comparison between the results achieved using the proposed method and other pitch extraction methods.

A method for the determination of pitch from aperiodic speech signals was pre-

sented by Hedelin and Huber [Hedelin and Huber, 1990]. In this article the authors identify four types of aperiodic voice excitation, and then proceed to present their decimated whitening autocorrelation pitch extractor, to deal with the irregularities present in these aperiodic speech signals, comparing their results to those of the classical pitch extraction methods, such as SIFT and autocorrelation.

A new method which uses a cochlear model coupled with a bank of auto-correlators in order to determine the pitch was presented by Slaney and Lyon [Slaney and Lyon, 1990]. The overall system was designed to mimic the human perceptual system using an auditory model whose outputs can be viewed in a three-dimensional graph of time versus frequency band versus intensity, called a correlogram. From these correlogram values, the subsequent post-processing stage determines the most likely pitch value for that given input signal window, without attempting or enforcing frame-to-frame continuity.

Another cepstral-based method which uses the one-sided autocorrelation of a input signal window was proposed by Nadeu, Pascual, and Hernando [Nadeu *et al.*, 1991]. Here, the cepstrum is taken from the one-sided, or causal part, of the autocorrelation sequence, allowing for an apparently sharper peak to be present in the resulting cepstrum at the lag corresponding to the pitch period.

With the emergence and application of neural networks in a number of different domains being attempted, it was only a matter of time before this paradigm would be called upon to solve the pitch extraction problem. In 1990, an attempt by Martínez-Alfaro and Contreras-Vidal for a neural network-based pitch detector, was presented [Martínez-Alfaro and Contreras-Vidal, 1991]. A multi-layer neural network, trained using backpropagation, was used, with speech signal windows consisting of 100 raw signal samples presented to the neural network inputs without prior preprocessing.

A method which appreciably improved both the resolution and accuracy of extracted pitch values over that of previous methods was the so-called super

resolution pitch determination method presented by Medan, Yair, and Chazon [Medan *et al.*, 1991]. This method is somewhat similar to that proposed by Chung and Alazi [Chung and Algazi, 1985], with the difference that Medan, Yair, and Chazon give a more detailed treatment on the use of the crosscorrelation for the purposes of pitch extraction. As well, they also present a method for interpolating between the sample values and achieving "super" or almost infinite resolution in the extracted pitch values. De Mori and Omologo present another method which is based on the super resolution method for the purposes of both visualization and pitch extraction [De Mori and Omologo, 1993].

Hanna recently presented a novel method for the extraction of pitch using the maximum likelihood [Hanna, 1992]. In this method, a frequency-domain maximum likelihood procedure is used for the estimation of the pitch frequency of voiced segments by maximizing a log-likelihood function over the range of possible pitch frequencies for the speech signal being analyzed.

Another method which has emerged over the past few years and is commonly used for the time-frequency analysis of signals uses wavelets. Although a number of articles present the use of wavelets to determine whether a segment is voiced or unvoiced or to detect the presence of pitch in an utterance [Kadambe and Boudreaux-Bartels, 1992], one group has used wavelets for pitch determination [Lunji *et al.*, 1993].

Although a number of different pitch extraction methods exist, as have been described in this section, the problem of accurate pitch extraction still remains an open problem, especially for speech which has short pitch periods, as it the case for some female speech, for children's speech, and of course, for infant cry utterances. As well, accurate determination and tracking of pitch on a period to period basis still eludes most published methods. In short, no real distinguishable trend or evolution in the development of fundamental frequency extraction methods has really emerged since the first digital signal processing techniques came into use

some 30 years ago. Although the complexity of the extraction methods has evolved from a better understanding of the underlying processes which produce speech, 100% accuracy still eludes most published extraction methods.

2.3 Neural Network-Based Classification Techniques

This section will initially present two examples of neural network-based classifiers in order to give a broad overview as to the wide variety of applications possible. However, the focus of this section is primarily on speech-related applications of artificial neural networks. Even with this particular focus, the amount of literature published on the applications of neural networks in the speech domain is quite extensive, and as such, it is impossible to mention all of the articles related to this domain. Thus, a survey of some of the more successful trials and architectures used in the speech domain will be presented.

Neural networks have been successfully applied to solve a wide variety of classification problems be it diagnosing HIV reverse transciptase inhibitors [Tetko *et al.*, 1994], to computing the likelihood of credit card fraud based on a pattern of user transactions [Ghosh and Reilly, 1994]. The broad appeal of neural networks lies in their ability to achieve good performance through the dense interconnection of simple computational elements. The potential benefits of the use of neural networks go far beyond the high computational rates, which are achieved from massive parallelism, by providing a greater degree of robustness, or fault tolerance, precisely due to the large number of node interconnections [Lippmann, 1987]. The training process of neural networks is still a major focus of the research undertaken in this domain, since it is the training process that adds robustness to the network by compensating for variabilities in the input patterns. Neural networks are non-parametric classification systems and thus make weaker assumptions about the shapes of the underlying distributions than do traditional statistical classifiers, and as such, may prove to be more robust when the distribu-

tions are generated by non-linear processes.

In speech applications their use first started with the article published by Kohonen, Mäkisara, and Saramäki which described the use of phonotopic maps for the visualization of speech signals [Kohonen *et al.*, 1984]. This mapping principle was originally used in image analysis and was later adapted for this speech application. Self-organizing networks [Kohonen, 1988] were used to form a two-dimensional map displaying the similarity relations between phonological units, obtaining more generality than the classical formant maps, since this method used the entire spectrum instead of the first three formant values to determine the spoken phoneme in a given word.

Until the first neural network tests on speech recognition emerged in the late 1980s, the method of choice for speech recognition applications was hidden Markov models (HMMs) [Rabiner, 1989]. However, with the first few recognition trials using neural networks giving reasonable results, the door was opened for further research into other network architectures and learning methods to be undertaken which could potentially be better suited to this task than the classical feedforward neural networks trained using back propagation.

The "IEEE First International Conference on Neural Networks" saw a number of presentations dealing with speech applications of neural networks. Shamma proposed a three-step view of the auditory processing and recognition process of speech which could be then emulated by a series of neural networks [Shamma, 1987]. One neural network could perform the initial transformation of the signal and then these transformations would be sent to the feature extraction stage, which would in turn send the extracted features to the learning and pattern recognition stage. Although no results or recognition rates were given in this paper, the method was indeed an intriguing way to process speech for recognition purposes, emulating the function of the human brain.

Gold, Lippmann, and Malpass [Gold *et al.*, 1987] also proposed the application

of neural networks to the recognition of steady state vowels in speech. In this paper, Hopfield neural networks [Hopfield, 1984] were modified and tested for the purposes of word recognition and this paper went on to describe an approach for dealing with time varying sequences, such as speech, for a Hopfield net which would have delay filters added in the activation functions.

The trial of several experiments on speech data using neural networks was also reported by Bourlard and Wellekens [Bourlard and Wellekens, 1987]. Here, 16 mel-cepstrum coefficients were generated for every 10 ms frame of input speech, which were then clustered using a k-means clustering algorithm and subsequently input into a neural network. Context information was also included at the inputs of the network in the form of previous and future input feature vector frames. Results from this proposed configuration, that of time-delay neural networks, and hidden Markov models, were compared in a later paper [Bourlard and Wellekens, 1989] where the relative strengths and weaknesses of these respective methods at capturing relevant speech features were discussed.

Lippmann and Gold presented a novel neural network architecture called a Viterbi-net and illustrated its use for the task of isolated word recognition [Lippmann and Gold, 1987]. This network performed temporal alignment of the input features derived from the speech signal to the input classification nodes, and used fixed delays and threshold logic to implement a modified version of the Viterbi algorithm, a method commonly used in many HMM speech recognizers. Recognition accuracy in this series of tests equalled that of HMMs. In a later article, Huang, Lippmann, and Gold suggested that the incorporation of these neural networks into HMMs might improve overall recognition performance [Huang *et al.*, 1988].

The use of "temporal flow" neural networks, or neural networks with recurrent connections, for dealing with time-dependent sequences was proposed by Watrous and Shastri [Watrous and Shastri, 1987]. In this article, the authors used time-

dependent input features, namely 16 filter-bank values, generated from the input signal, which was segmented by hand in order to avoid time-alignment problems between subsequent words. This set of experiments illustrated that this architecture was useful for word recognition.

Also, in 1987, Waibel, Hanazawa, Hinton, Shikano, and Lang presented a new architecture which was developed principally for capturing the acoustic features between subsequent segments of input signal windows for the purposes of phoneme recognition [Waibel *et al.*, 1987, Waibel *et al.*, 1989]. This architecture, called a time-delay neural network (TDNN), also had the feature that it could tolerate poorly aligned input frames of data, and still perform accurate recognition. The results achieved by this architecture were comparable or superior to those achieved using hidden Markov models [Waibel *et al.*, 1988]. For the phoneme recognition experiments, the input features consisted of 16 mel-scale coefficients computed every 5 ms from a 10 ms window of speech. These networks were later used in word recognition experiments with recognition rates exceeding 90% [Bodenhausen and Waibel, 1991].

Other applications of neural networks in speech related experiments included the recognition of place of articulation [Bengio and Mori, 1988]. In this particular article, the Boltzmann machine algorithm and back propagation algorithm were used to learn the front, center, or back, place of articulation for vowels. Coding the input spectral lines of the input window of speech used a scheme that employed the relative frequencies and amplitudes in a non-linear frequency scale representation similar to that of the human ear, giving results which exceeded that of an HMM on the same data set.

An alternate set of input features were used by Leung and Zue [Leung and Zue, 1988] who derived features from the input speech signal using Seneff's model of the human auditory system [Seneff, 1984] to classify phonemes cut from continuous speech. Here, contextual information was also supplied in

the form of cuts from the preceding and subsequent phonemes in the word to a feedforward neural network trained using back propagation. Recognition rates using this method were about 60% [Leung, 1989].

The idea of using Kohonen feature maps for the purposes of phoneme-based speech recognition was rekindled by Kepuska and Gowdy [Kepuska and Gowdy, 1989]. In this article the authors proposed a solution to the problem of feature vectors from other phonemic portions of an uttered word overlapping with the phonemic portion under consideration. For accurate recognition, the authors propose using the steady state portions of phonemes as input into the neural network recognizer.

A dynamic-programming-based matching system, coupled with a neural network trained using back propagation, was implemented and tested by Sakoe, Isotani, Yoshida, Iso, Watanabe [Sakoe *et al.*, 1989]. This network model, called a dynamic programming neural network (DPNN), could easily treat time-sequence patterns, using the popular dynamic time warping time alignment technique. Experimental results of 99.3% were achieved for isolated Japanese digit recognition.

Other research groups performed a comparison of different recognition techniques or different neural network architectures. One such group used both a static and a sequential presentation of input features consisting of 18 mel-cepstrum coefficients computed from every 10 ms of speech to determine the accuracy for a simple recognition task and for digit recognition as well [Demichelis *et al.*, 1989]. Another group performed a comparison of four neural network architectures which had been previously deemed as useful for speech recognition purposes [Fallside *et al.*, 1990], but their experiments either achieved very poor results or were incomplete. Other comparisons between neural networks and HMMs were performed by Bridle [Bridle, 1991] who also set out a series of comparative measures to aid in this determination.

As well, a comparison between a multi-layer feedforward neural network

and a competitive learning method, called a learning vector quantization [Kohonen, 1988] neural network, was presented by Ahalt, Jung, and Krishnamurthy [Ahalt and Jung, 1991]. In this article the authors use three feature sets derived from the input speech signal window: autocorrelation coefficients, weighted linear prediction cepstral coefficients, and formant frequency values. A comparison of different learning vector quantization-based neural network architectures for robust speech recognition was performed by Zhu, Li, Guan, and He [Zhu *et al.*, 1993].

A neural network hybrid was presented by Hataoka, Amano, Aritsuka, and Ichikawa [Hataoka *et al.*, 1990]. The authors presented an algorithm for large vocabulary speech recognition using two kinds of connectionist models. The first one was a phoneme recognition model which used a method combining neural nets and fuzzy inference called neural-fuzzy. This method used neural nets as acoustic feature detectors and fuzzy logic as a decision procedure. The other was a connected-word sequence selection method which used semantic information about conceptual relationships among vocabulary words.

Another neural network and hidden Markov model hybrid used for speech recognition was presented by Robinson [Robinson, 1992]. The recurrent neural network was used for the purposes of context modeling and also provided phoneme state occupancy probabilities for a simple context independent hidden Markov model. The description of the implementation of the entire recognition system was later described in another article [Robinson *et al.*, 1993].

Using neural networks as a postprocessor for HMMs was described by Jin and Chung [Jin and Chung, 1992]. The neural network was introduced to enhance the classification capability of hidden Markov modeling for speech recognition purposes. This postprocessor received stimuli from not one, but all word HMMs for each word in the input speech, and the input speech frames did not require prior segmentation.

As well, Bengio, De Mori, Flammia, and Kompe, presented the design and

evaluation of three neural networks in a composite neural network and HMM framework [Bengio *et al.*, 1992]. Of the three neural networks, one detects manner of articulation and the other two describe the signal in terms of place of articulation, all of which were inspired by acoustic-phonetic knowledge. The latter two networks were later merged when the hybrid system was implemented, with the HMM serving to model the neural network outputs.

The issues of using partial connections between nodes in adjacent layers was discussed by Ye, Wang, and Robert [Ye *et al.*, 1990]. In the experiment presented in the article, the neural network with partial connections was used to perform isolated word recognition. Results demonstrated the advantages of partial connections against that of full connections. Partial connections can introduce both temporal context constraints and some implicit knowledge into the network, and may also lead to efficient learning on a small data set size.

Another set of input features derived for a neural network in a speech recognition application was proposed by Nguyen, Lippmann, Gold, and Paul [Nguyen *et al.*, 1990]. In this article the authors propose the use of a front-end preprocessor based on the ensemble interval histogram model developed by Ghitza [Ghitza, 1986]. The network using the front-end preprocessor achieved results comparable to those using mel-scale filter-band inputs.

Neural networks were also investigated for the purposes of pitch detection and pitch determination. Barnard, Cole, Vea, and Alleva [Barnard *et al.*, 1991] presented two feedforward neural networks for pitch detection. One used the raw input signal, and the other neural network used features derived from the input signal, called peak descriptors, as inputs into the pitch detection network. In another experiment, Martínez-Alfaro and Contreras-Vidal used a feedforward neural network to perform pitch estimation [Martínez-Alfaro and Contreras-Vidal, 1991]. This neural network used 100 samples of the raw input signal, with the output consisting of 100 nodes, one for each of the possible pitch period lags.

Nakamura and Sawai [Nakamura and Sawai, 1992] proposed a modular time-delay neural network for the purposes of speaker-dependent speech recognition. Here the authors demonstrate that a modular network, with one network assigned to each speaker, performs as well as a single time-delay neural network with a larger hidden layer, given the larger storage capacity that is required for a larger collection of speakers by a single neural network.

An interesting, but brief, comparison of two auditory models and mel-cepstral coefficients as inputs to a phoneme recognition neural network, which employed an unsupervised learning method, was performed by Anderson [Anderson, 1993]. In this article it was noticed that different input patterns make different types of broad class recognition errors, but that auditory models offer some improvements over the mel-cepstrum coefficient inputs. Another similar comparison, performed using linear prediction coding coefficients, and features derived from an auditory model, this time for speaker identification purposes on two neural network architectures, also included Anderson as part of the investigating team [Colombi *et al.*, 1993].

Another novel neural network architecture was proposed by Li, Fang, and Li [Li *et al.*, 1993]. In this article, the authors propose a self-organizing neural tree, which is suitable for hierarchical classification and vector quantization. This network promises to provide good results for speech recognition and image coding, and has the advantage that the training time for the neural tree is much shorter than for other competitive networks.

Wu and Chan presented an neural network for the purpose of speaker independent word recognition [Wu and Chan, 1993]. The network was composed of three concatenated subnetworks. One subnet converts the information contained within the features extracted from a speech signal frame into a probability vector whose components correspond to the estimated probability of the feature vectors belonging to the phonetic classes that constitute the words in the vocabulary. These outputs are then crosscorrelated by the second neural network and then presented

to the decision making classification subnetwork for final classification.

Another neural network model which used inputs derived from neurophysiological findings in the auditory system was presented by Yamauchi, Fukuda, and Fukushima [Yamauchi *et al.*, 1993]. The system used two separate modules, one to extract auditory features from the input signal, and the second to perform recognition based on the extracted auditory features which accumulates features over time and in three separate neural network blocks, accommodating different speaking rates without affecting recognition.

Hadjitodorov, Boyanov, Ivanov, and Dalakchieva presented a system for speaker identification which employed two neural network architectures [Hadjitodorov *et al.*, 1994]. This method used one neural network based on self-organizing maps, and another network using the autoregressive neural network model, with the final classification decision obtained through a voting principle using the decisions of the two classifiers.

Another attempt at speaker recognition was performed by Kuah, Bodruzzaman, and Zein-Sabatto [Kuah *et al.*, 1994]. Twelve feature parameters were obtained from the mel-cepstrum coefficients and from linear prediction coding coefficients which then served as input into a feedforward neural network. Three different speakers uttering 13 different words were used to train and test the system achieving good results.

A novel use of wavelets and neural networks for the purpose of both speaker identification and the classification of unvoiced sounds was recently proposed by Kadambe and Srinivasan [Kadambe and Srinivasan, 1994]. Wavelets are used to help in the assistance of classifying unvoiced sounds, and in the identification of speakers based on only one pitch period of speech data. These features are then input into a feedforward neural network for classification.

Since the initial application of artificial neural networks to speech-related domains over 10 years ago, the architectures used to address different classification or

recognition issues have progressed from simple feedforward architectures, to the use of specially designed architectures, to multiple neural network configurations, to hybrid configurations combining neural networks with other classification methods. Despite the successes of more complex configurations or hybrid systems, good results using relatively simple neural network architectures have nevertheless been achieved for a number of speech and phoneme recognition experiments. These results motivate our simple initial attempt using neural networks for the purposes of cry classification before more complex classification systems are investigated.

Chapter 3 Improved Fundamental Frequency Extraction for Infant Cry Vocalizations

This chapter presents an improved fundamental frequency determination method, called the improved crosscorrelation vector-based fundamental frequency extractor, which is capable of tracking rapid changes in pitch due to double-harmonic break episodes in the utterance signal, and capable of dealing with pitch values which are within the large range of allowable F_0 values for infant cry vocalizations. The method, which will be described in section 3.1, allows the determination of F_0 on a period-by-period basis and can be used for both improved visualization of the utterance, and for further pitch-synchronous processing of these vocalizations. The results of this improved fundamental frequency extraction method will also be compared to the results obtained on several cry recordings using six other methods adopted from the speech processing domain described in section 3.2. They are the linear predictive coding (LPC) method, its variant, the simplified inverse filter tracking (SIFT) method, the cepstral extraction method, the harmonic sieve, the spectral flattening autocorrelation method, and the super-resolution pitch extraction method. These various methods will be tested on five different cry recordings from the data set described in section 3.3 and the results will be presented and discussed in sections 3.4 and 3.5 respectively. The chapter concludes with an illustration of the improved visualization of utterances achievable using the improved crosscorrelation vector-based fundamental frequency extraction method.

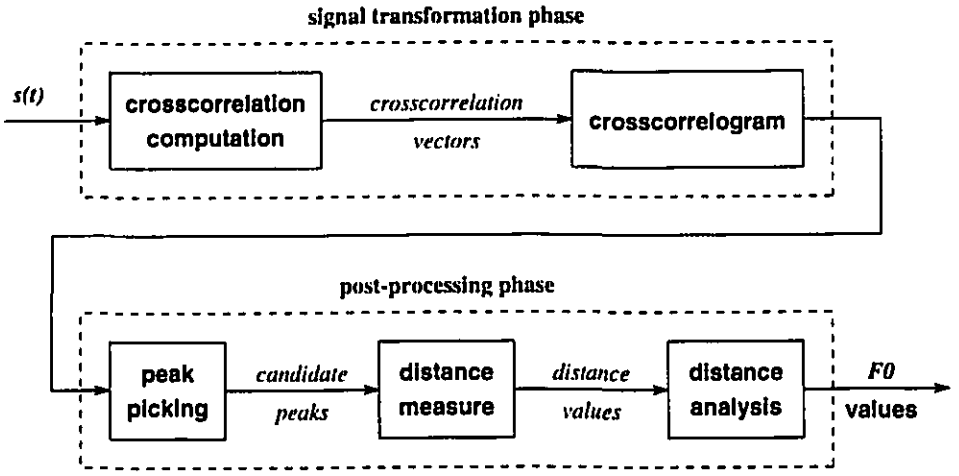


Figure 3.1: Block Diagram of the Improved Crosscorrelation Vector-Based Fundamental Frequency Extraction Process

3.1 Improved Crosscorrelation Vector-Based Fundamental Frequency Extraction

This section describes the different parts of the improved F_0 extraction method which is based on the processing of the sequences of crosscorrelation vectors generated in an earlier processing phase of the signal. An overview of the method is first given, followed by an indepth description of the individual stages of processing performed by this method.

3.1.1 Overview of the Improved Fundamental Frequency Extraction Method

Figure 3.1 presents the method in a block diagram format consisting of a number of steps which are grouped into two different stages, namely, the **signal transformation** phase and the **post-processing** phase.

The first phase transforms the input signal using the normalized crosscorrelation into a feature space which is then used in the post-processing phase, for the purpose

of accurately determining and tracking the value of F_0. Ideally, the sampled input signal $s(n)$ should consist of at least one complete utterance, that is, from the start of a vocalization to a return to silence, in order for the post-processing phase to function accurately. If only a portion of an utterance is available, the method still performs properly, but obviously the progression of F_0 cannot be monitored or tracked for the missing portions of the signal.

Although this is not a real-time method, the multiple steps required by this algorithm are necessary to ensure that correct F_0 values will be produced for infant cry vocalizations, and especially troublesome class of speech signal. This method is useful not only for infant cry signals, but for speech in general, and could be of particular interest for cases where the F_0 values of a speaker needs to be subjected to a very detailed analysis, as could be the case if the emotional state, or laryngeal pathology of a speaker is to be determined. The algorithm is not restricted to a particular range of F_0 values; any range can be accommodated. Subsequent post-processing could yield F_0 values for every pitch period in the recording with so-called "infinite" or "super" resolution [Medan *et al.*, 1991].

The following sections give a detailed description of each portion, or stage, of the improved crosscorrelation vector-based fundamental frequency extractor.

3.1.2 Crosscorrelation-based Pitch Extraction

As was mentioned in section 2.1, time-domain pitch extraction methods are both computationally simpler than their frequency-domain counterparts, and also yield more accurate pitch values. This class of F_0 determination methods also allows the possibility of locating the pitch epoch, which is the time at which the vocal folds close, and which is denoted by the abrupt increase in the signal waveform, as illustrated at the locations labeled A, C, and E in figure 3.2.

Closure of the vocal cords initiates the pitch period in a voiced signal. The

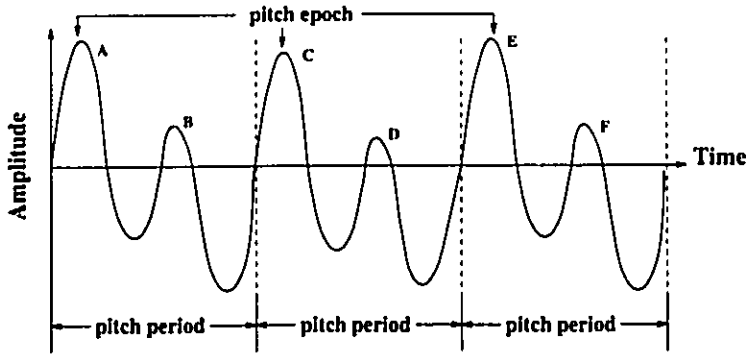


Figure 3.2: Sample Speech Signal Waveform

abrupt increase in the signal amplitude marking this event is then followed by a decaying amplitude envelope, as can be seen in figure 3.2 by the decrease of the amplitude value of the pitch epoch peak, and the amplitude value of the oscillation following the pitch epoch, labeled B, D, and F. The rate of decay of the amplitude envelope and the period of the intermediate oscillations between subsequent pitch epochs are proportional to the bandwidth and equal to the period of the highest energy formant frequency of the vocal tract. Typically, this corresponds to the lowest resonant frequency of the vocal tract, namely, the first formant frequency (F_1) [O'Shaughnessy, 1987].

Consequently, these methods are useful for for pitch-synchronous processing of the signal, allowing not only the fundamental frequency to be determined on a period-by-period basis, but other parameters as well, such as the values of the formants, for example. Formant frequency values correspond to the resonant frequencies of the vocal tract, allowing one to gain some insight into the shape of the vocal tract for cries uttered in different situations.

Since the accurate extraction and tracking of vocal fundamental frequency is what is desired in order to overcome the deficiencies of other F_0 extraction methods, it is no coincidence that the core method upon which the F_0 extraction method presented here is based, is a time-domain method.

The normalized crosscorrelation is at the heart of the improved crosscorrela-

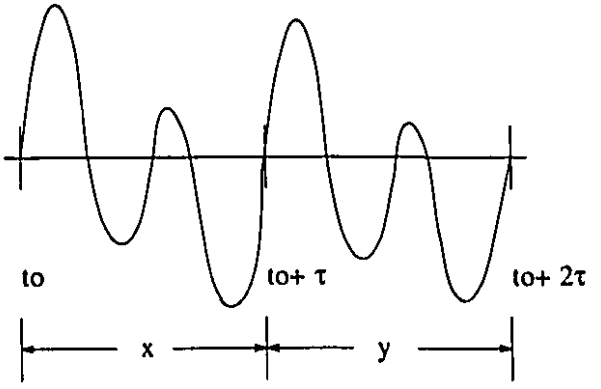


Figure 3.3: Adjacent Segments of Voiced Speech Signal

tion vector-based fundamental frequency extraction method as it is for the super-resolution extraction method originally described in a journal article by Medan, Yair, and Chazon [Medan *et al.*, 1991]. The benefits of using the normalized cross-correlation for fundamental extraction were presented in a paper by Chung and Alazi [Chung and Algazi, 1985]. The details of the normalized crosscorrelation, and its usefulness for F_0 extraction, are described below. Although this method leads to a highly accurate method of F_0 extraction for speech signals, it requires further refinements in order for it to be accurate for infant cry vocalizations as well.

Consider a periodic portion of a voiced speech segment $s(t)$ and two adjacent segments $x_\tau(t, t_0)$ and $y_\tau(t, t_0)$, as are shown in figure 3.3. Each of these segments are of length τ and both $x_\tau(t, t_0)$ and $y_\tau(t, t_0)$ span a segment of the signal $s(t)$ in the interval $[t_0, t_0 + 2\tau]$. If, starting at $t = t_0$, there is a portion of the signal 2τ which contains exactly two pitch periods so that $\tau = T_0$, corresponding to the fundamental frequency period, and where $x_{T_0}(t, t_0)$ is the first pitch period of the signal segment and $y_{T_0}(t, t_0)$ is the second pitch period of the signal segment, it can be said that the two segments will differ only in amplitude and from other distortions resulting from dissimilarities between the two signals.

The difference in amplitude between these two adjacent segments can be expressed in terms of an amplitude modulation factor, denoted as $a(t_0)$, and the distortion and dissimilarity factor between $x_{T_0}(t, t_0)$ and $y_{T_0}(t, t_0)$ can be expressed

as $e(t, t_0)$.

Consequently, the two adjacent signal segments can be represented in terms of each other as follows:

$$x_{T_0}(t, t_0) = a(t_0) y_{T_0}(t, t_0) + e(t, t_0) \tag{3.1}$$

with the condition that the two successive pitch periods are sufficiently similar, so that the aforementioned assumptions hold. The time interval $\tau = T_0$ for which the error term $e(t, t_0)$ is minimized over the time interval $[t_0, t_0 + \tau]$ according to an error norm, say the normalized square error, is defined as the pitch period for the time instant $t = t_0$.

Minimizing equation 3.1, using the normalized square error norm leads to the following equation:

$$E_\tau(t_0) = \frac{\int_{t_0}^{t_0+\tau} [x_\tau(t, t_0) - a(t_0) y_\tau(t, t_0)]^2 dt}{\int_{t_0}^{t_0+\tau} [x_\tau(t, t_0) y_\tau(t, t_0)] dt} \tag{3.2}$$

where the denominator of equation 3.2, serves as a normalization term, compensating for the occurrence of non-zero mean segments, which are common when the segments do not include complete signal cycles. The argument of the integral in the denominator can be replaced by either $[x_\tau(t, t_0)]^2$ or $[y_\tau(t, t_0)]^2$ if one considers the energy contained in $x_\tau(t, t_0)$ to be similar to that contained $y_\tau(t, t_0)$ for the purposes of normalization. In a practical implementation, the value of τ should be restricted to the range of expected pitch period values, which in the domain of infant cry vocalizations range from 0.4 ms to 6.6 ms, corresponding to frequencies between 2500 Hz and 150 Hz respectively.

Next, equation 3.2 is differentiated with respect to $a(t_0)$ in order determine for which τ the dissimilarity or error measure between the two adjacent signal segments is minimized, and to find an optimal value for the amplitude modula-

tion term. The latter turns out to be $a(t_0) = \int_{t_0}^{t_0+\tau} x_\tau(t,t_0)y_\tau(t,t_0)dt / \int_{t_0}^{\tau}[y_\tau(t,t_0)]^2dt$, where the numerator represents the inner product $(x,y)_\tau$ from t_0 to $t_0+\tau$, and the denominator is the energy of the segment $y_\tau(t,t_0)$.

Using this result, the minimization of equation 3.2 can be expressed in the following manner:

$$E_\tau(t_0) = 1 - \left[\frac{(x,y)_\tau^2}{|x|_\tau^2|y|_\tau^2}\right]. \tag{3.3}$$

The second term on the right-hand side of equation 3.3 is immediately identified as a normalized crosscorrelation term $\rho_\tau(x,y)$. Consequently, minimizing the above equation, is analogous to maximizing $\rho_\tau(x,y)$ to find the pitch period. It should be noted, however, that if adjacent segments of length corresponding to subsequent multiples of the pitch period T_0 are sufficiently similar to each other, that is, for $\tau = 2T_0, 3T_0$, then $\rho_\tau(x,y)$ will also produce maxima, as is expected.

The normalized crosscorrelation approach calculates the instantaneous value of the pitch period which usually corresponds to the point where this value is greatest, as the index τ sweeps over the range of expected pitch values.

When dealing with a sampled signal, as is the case for this application, one can replace the time indexes t and t_0 by the sample indexes n and n_0. Moreover, in the implementation of this first portion of the method, the adjacent segments of the sampled signal, $s(n)$, namely $x_n(n_0)$ and $y_n(n_0)$, were taken to be of length $2n$, instead of being of length n. This was done in order to reduce the effects of strong formant peaks in the calculation of the normalized crosscorrelation values, to improve its immunity to noise, and to remove the effects of short periodic noise bursts during silent portions of the signal from being incorrectly identified as a very short cry vocalization.

The above method, as implemented, processes the sampled cry vocalization signal $s(n)$ as shown in figure 3.4.

First, the crc sscorrelation values are generated for all adjacent segments sep-

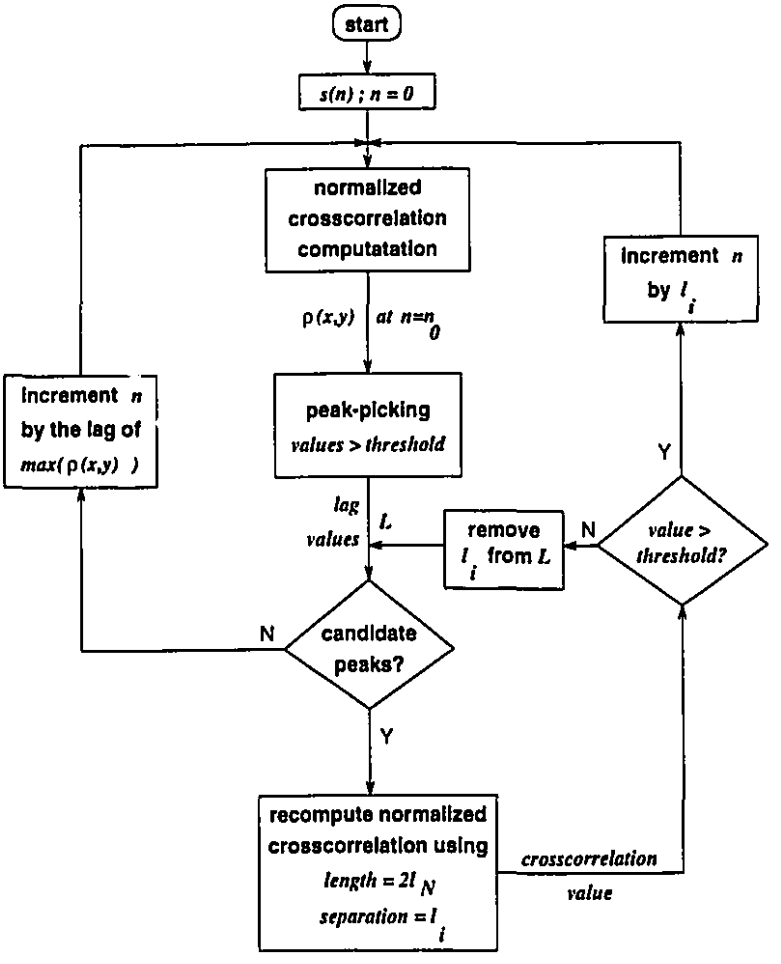


Figure 3.4: Flow Chart of the Crosscorrelation Block of the Signal Transformation Phase

arated by n of length $2n$. Since the cry recordings are sampled at 16 kHz, this corresponds to traversing lag values from 6 to 110, which correspond to periods of 0.375 ms to 6.875 ms or frequency values of 2667 Hz to 145.5 Hz. Figure 3.5 shows a signal segment and a plot of its corresponding crosscorrelation vector. The crosscorrelation is evaluated for each lag sample n at n_0.

Once the crosscorrelation values have been computed for all the lag values at a given time index n_0, the values which make up the crosscorrelation vector for $n = n_0$ are searched for peaks above a certain threshold. All lag values that have corresponding maxima values greater than this threshold value are saved. These lags represent the locations in $\rho(x, y)_n$ of possible pitch period values which will be

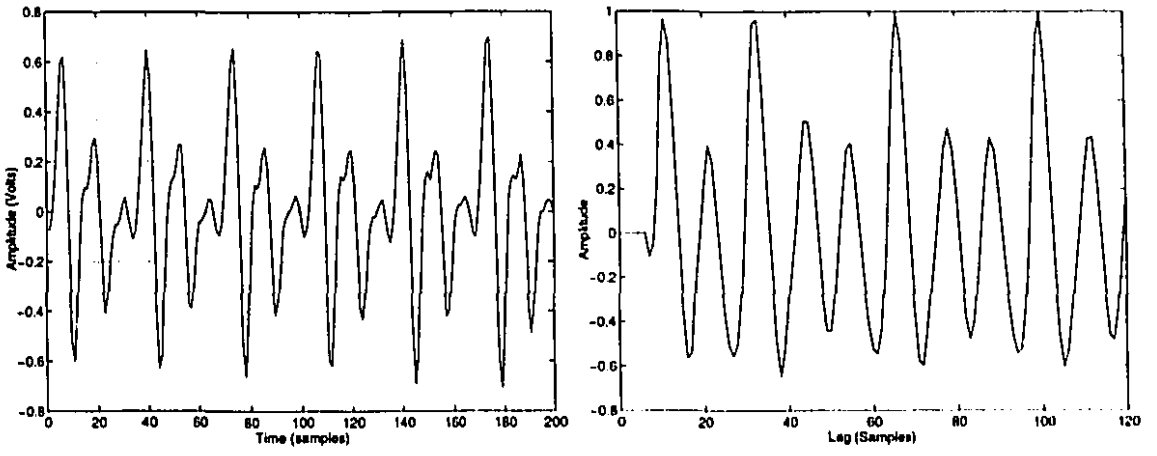


Figure 3.5: Cry Utterance Segment and its Corresponding Crosscorrelation Vector

subject to further processing in order to determine in an approximate fashion which of the lags most likely corresponds to the pitch period. In the simplest cases, the lag where the crosscorrelation vector has its highest value could be considered to be the pitch period, or, alternatively, the lag of the first crosscorrelation maxima to exceed a certain threshold could be considered to be the pitch period, as a number of classical F_0 extraction methods do [Ross *et al.*, 1974, Rabiner, 1977].

However, the more successful F_0 extraction methods use some form of post-processing on the pitch period candidates extracted from a speech signal in order to improve overall accuracy of the extraction algorithm. This is important when infant cry signals are processed as well. Identifying a possible pitch candidate correctly saves computation time when the crosscorrelation vectors are generated in this initial phase, since the time, or sample, index n_0 will be incremented by the value of the most likely pitch period lag in preparation for the calculation of the next sequence of crosscorrelation values, also called a crosscorrelation vector. In the case where the true F_0 lag is greater than the identified lag, extra crosscorrelation computations will be performed, thus increasing the time required for the algorithm to move through a given signal.

In order to minimize the likelihood of performing these extra crosscorrelation computations, the following post-processing is done on the crosscorrelation vector

peak candidate lag values [Medan *et al.*, 1991]. First these lag values are ordered from smaller to larger lag values $L = [l_1, l_2, \ldots, l_N]$, where l_N corresponds to the largest lag value whose crosscorrelation value exceeds the prespecified threshold. For all the l_i's in L, starting with $i = 1$, two adjacent segments of sampled signal $s(n)$ at $n = n_0$ are re-processed, where n_0 is the current time index in $s(n)$ from which the sequence of crosscorrelation values were first calculated, leading to the set of lag values L for $n = n_0$. Beginning with $i = 1$, the adjacent segments of $s(n_0)$ are **each of length** $2l_N$, rather than the previous length $2l_i$, and are each separated by l_i, that is one segment is given by $s(n_0)_{2l_N}$, and the other segment is given by $s(n_0 + l_i)_{2l_N}$. The normalized crosscorrelation between these two segments is then calculated. The first l_i whose crosscorrelation value exceeds another prespecified threshold is the value by which the time index n_0 is incremented for the computation of the next series of crosscorrelation values for the crosscorrelation vector.

The above re-computation of the normalized crosscorrelation for the lag values in L is done in order to minimize the occurrence of lags corresponding to strong, narrow bandwidth, first formant (F_1) frequency values, which occur at frequencies in the vicinity of $2F_0$, being selected as the time index increment. In a voiced utterance, the value of the first formant can be identified as the inverse of the period of dominant oscillation between two successive pitch epochs [Rabiner and Schafer, 1975].

Cases where the period of the first formant frequency is chosen as the increment value, increase the computation time of the crosscorrelation vectors for a given voiced utterance, since the time index is incremented in smaller steps in these cases. This method is based on the experimental results that if the lag l_i corresponds to a formant peak, then the two adjacent signal segments $s(n_0, 0)_{2L_N}$ and $s(n_0, l_i)_{2L_N}$ will be dissimilar, and correspondingly, the normalized crosscorrelation of these segments will be small. If, on the other hand, the lag l_i corresponds to a pitch period, then these two adjacent segments will be very similar which will in turn yield a high normalized crosscorrelation value. So, in effect, l_i can be considered

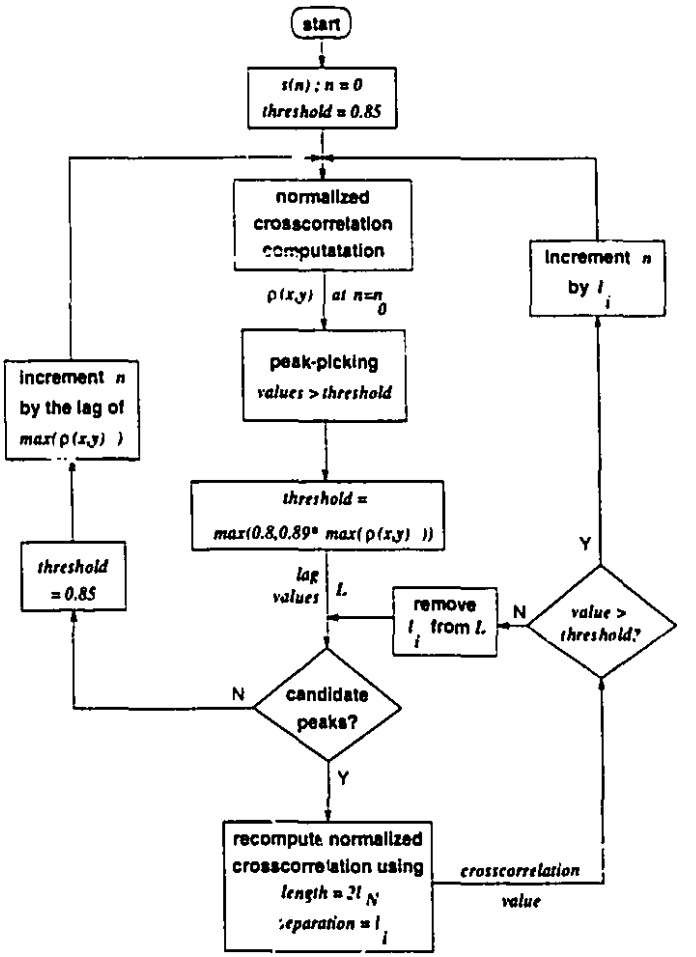


Figure 3.6: Flow Chart of the Crosscorrelation Block of the Signal Transformation Phase Augmented with Adaptive Threshold Setting

to be an approximate value for the pitch period, or a *pseudo* pitch period value.

Using fixed threshold values for the aforementioned thresholds has the disadvantage that this value does not change as the characteristics of the signal or the periodicity in the signal either becomes stronger and more prominent, or weakens. A solution to this would be the introduction of a crosscorrelation vector maxima threshold value; a dynamic variable which is set by augmenting the flow diagram shown in figure 3.4 to that shown in figure 3.6. When the crosscorrelation vector generation algorithm begins, the threshold is initially set to 0.85 so that relatively high crosscorrelation values, due to brief noise bursts or locally periodic disturbances in the signal, are not picked up and tracked. Once a maxima in the

crosscorrelation vector exceeds this initial threshold value of 0.85, the threshold for both the crosscorrelation values of L and for the subsequent time index value are set to the greater of 0.8 and the largest value in the crosscorrelation vector multiplied by 0.89.

Having an adaptive threshold value permits low crosscorrelation values to be considered as candidates in portions of the signal where there may be a formant change or a lower signal to noise ratio which causes the similarity between two adjacent segments to be reduced. As well, this adaptive threshold has the benefit of reducing the aforementioned effects of relatively high crosscorrelation values on the time index due to the presence of a very narrow bandwidth F_1, by setting the threshold to be a percentage of the maximum crosscorrelation value when two adjacent segments are very similar.

Despite the use of adaptive thresholds which attempt to minimize the occurrence of incorrect pseudo pitch period values, the characteristics of some cry signals are such that there is usually very little decay between the pitch epoch and the subsequent peak within the pitch period, unlike speech signals, where this decay is appreciable. The rate of decay is inversely proportional to the bandwidth of the highest energy formant [O'Shaughnessy, 1987], which for cries is almost always F_1. Furthermore, the method of proclaiming the first l_i whose crosscorrelation value exceeds the threshold as the pseudo pitch period does not completely eliminate the occurrence of errors. In fact, for very narrow bandwidth F_1 values in cry utterances, this method still leads to numerous gross pitch errors, that is, the pseudo pitch period is either twice or half the true value. This is clearly unacceptable if accurate F_0 extraction is required.

A number of journal papers published over the years have attempted to address the problem of gross pitch errors by beginning to track a certain pitch period value in the vicinity of the previous pitch period value, once others have been found for a number of consecutive time indexes or frames [Hess, 1976, Markel, 1972a,

Medan *et al.*, 1991]. This works well for the majority of speech signals, since a given F_0 value is usually within ±25% of the previous value. This is not the case for infant cry vocalizations which can abruptly change or have double harmonic break episodes present [Wasz-Höckert *et al.*, 1968]. Consequently, further processing for these crosscorrelation vectors considered as a group is necessary in order to obtain the desired accuracy.

During episodes where the crosscorrelation maxima for a given vector or series of vectors are all below the threshold, be it due to a the occurrence of a dysphonic, silence, or ambient noise interval, the time index will be advanced according to the lag value corresponding to the crosscorrelation maxima with the largest value for that vector. This procedure can be identified on the left-hand side of figure 3.6.

3.1.3 Grouping of the Crosscorrelation Vectors

The crosscorrelation vectors generated from the signal $s(n)$ over the length of the recording can be placed together in a 2-dimensional manner similar to that in which fast Fourier transform vectors of successive signal segments are placed together for the generation of a spectrogram. This concatenation of crosscorrelation vectors yields a matrix of lag versus time where the entries in this matrix represent the crosscorrelation value of a specific lag value at a specific time index. This allows the progression of the crosscorrelation vector maxima, which can be considered as being pitch period candidates, to be tracked over time.

By using the observations regarding the crosscorrelation maxima made in the previous section, it is expected that the actual pitch period lag value will be contained in one of the first few lag values in L, whose crosscorrelation maxima exceed the threshold, as the lag values are traversed from low values (high frequency) to high values (low frequency). The lag values of subsequent maxima which exceed the threshold, and which follow the true pitch period lag, represent sub-harmonics

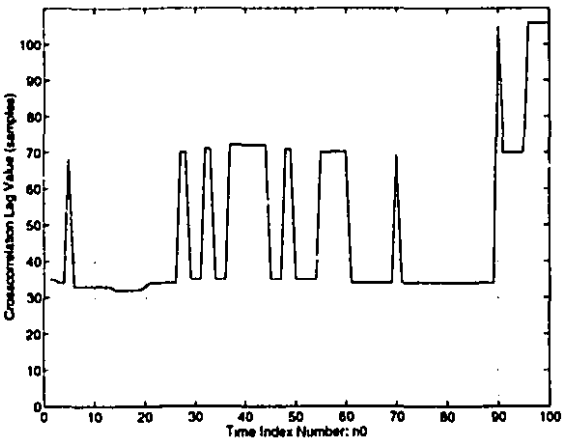


Figure 3.7: Plot of the Lag Values with the Largest Crosscorrelation Values

of either the fundamental frequency period or the first formant period.

Identifying the lag where the first maxima exceeds the threshold in the crosscorrelation vector as being the pitch period, as has been suggested by some researchers, yields good results for speech signals, since high-energy formant values typically have large bandwidths. Hence crosscorrelation vector maxima due to intra-pitch period oscillations fall below threshold values and are eliminated from further considerations [De Mori and Omologo, 1993]. This observation is not true for some cry signals however, and so this heuristic is not useful for accurate pitch period determination for this class of signals.

One observation made over the course of examining numerous crosscorrelation vectors for a large number of cry utterances, was that the crosscorrelation value of the pitch period lag will be the largest of all the other maxima values for the majority of time indexes in a given section of an utterance, where the pitch period values will be within $\pm 25\%$ of the previous pitch period value at the previous time index. Although this observation does not preclude rapid and abrupt changes in the pitch period values of an utterance, it does imply that once the pitch period changes in a cry, it does so for a number of periods, not only for one or two pitch periods.

An example of this observation is shown in figure 3.7. The lag values corre-

sponding to the largest of the crosscorrelation maxima values in the vectors change abruptly and unpredictably, leading to numerous pitch halving or doubling errors, or gross pitch errors, if this technique were used to extract the pitch period, as was described in the previous section. Figure 3.7 plots the progression of pitch period values for a portion of a voiced utterance with an actual pitch period varying between 32 to 34 samples, which corresponds to F_0 values between 500 Hz and 470.6 Hz. The post-processing of the set of crosscorrelation maxima L in order to eliminate the effects of narrow bandwidth F_1 values which occur at values of $2F_0$, as done in the super-resolution pitch extraction method, also yields inconsistent results, as will be shown in section 3.4. This necessitates the following level of post-processing of the crosscorrelation vectors in order to extract F_0 accurately from cry utterances.

3.1.4 Post-Processing Phase

The observation that, in a given utterance, the majority of lag values corresponding to the largest of the the crosscorrelation maxima in the sequence of crosscorrelation vectors corresponds to the true pitch period lag can be exploited for the post-processing phase. Once again, figure 3.7 illustrates that using the heuristic of selecting the lag value where the crosscorrelation value is greatest as the value for the pitch period leads to very inconsistent results. Despite these inconsistencies, however, it is readily observable that the majority of the lag values in this voiced section do indeed correspond to the true pitch period lag value of between 32 and 34 samples. In fact, of the 100 time indexes in figure 3.7, only 33 of these values are incorrect, and actually correspond to sub-harmonics of F_0 which are integer multiples of the pitch period.

Let us first define a pitch contour derived from a series of contiguous cross-correlation maxima lag values, which are above the specified threshold, and have subsequent lag values of the maxima lying within ±25% of the current lag value. If

a distance measure such as a simple sum of the crosscorrelation maxima values over the length of the pitch period contour would be used, then at the end of the utterance, this "contour" would have the highest distance value of all the other maxima lag contours. Empirically, for a contour to be considered as a valid vocalization, this should occur for an interval lasting at least 8 time indexes. Shorter interval lengths are usually due to locally periodic noise bursts and are not considered as valid markers of voiced events in the utterance.

In order to achieve this distance measure and distance analysis previously shown in the post-processing block of figure 3.1, the following steps, explained in flow chart form in figure 3.8, must be performed.

First, the crosscorrelation vector matrix is thresholded twice using two different thresholds: a high threshold value, t_H, and a low threshold value t_L. In the first pass, all crosscorrelation vector maxima that have values greater than $t_H = 0.8$ are kept and stored in a matrix labeled M_H. Then in the subsequent pass, all crosscorrelation peak values greater than $t_L = 0.6$ are kept and stored in a matrix labeled M_L. The lag values where crosscorrelation maxima lie above the respective thresholds in the M_H and M_L matrices, are set to a value of 1. At all other lag values in these matrices, the entries are set to 0, leading to very sparse M_H and M_L matrices.

A number of different thresholds were tested for both stages of this peak extraction process but this combination yielded the best results. These thresholds allow crosscorrelation maxima lag values to be accepted when either a formant change occurs, causing a brief drop in the crosscorrelation maxima values, but still allowing the method to continue the pitch period track, during periods in the signal where the cry utterance is particularly weak, or when there is ambient noise picked up by the recording, all while successfully excluding "false starts" in a contour due to a brief and locally periodic noise signal.

Once this thresholding has been performed, the values in the matrix thresholded

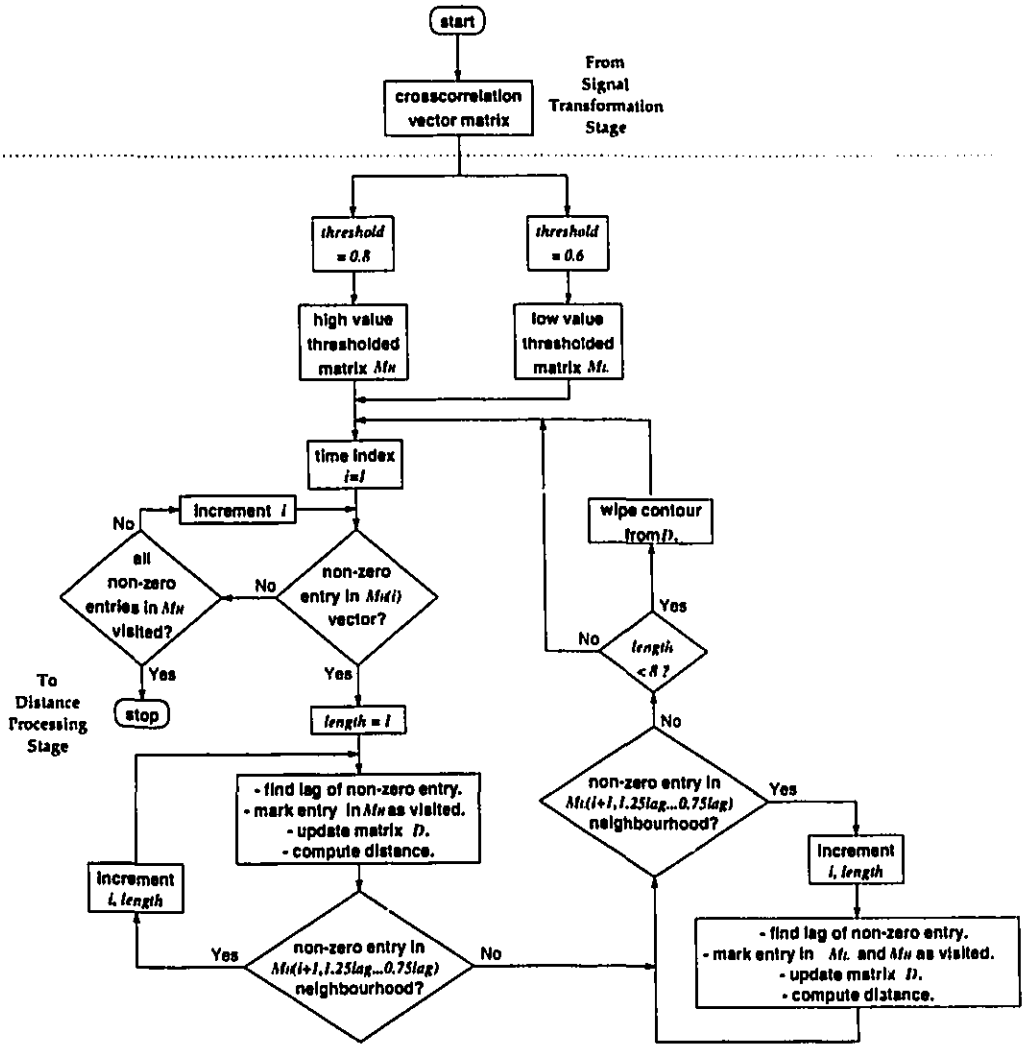


Figure 3.8: Flow Chart of the Peak Picking and Distance Computation Stages of the Post-Processing Phase of the Pitch Period Extractor

with the higher value, M_H, are examined. Starting at the first time index, $i = 1$, the algorithm increments the time index i until a non-zero entry is found for a lag l at time i in M_H. This indicates that the maxima in the crosscorrelation vector matrix lies above the high threshold value, t_H. Once one such lag is found, the algorithm then begins to look for non-zero entries in M_H at subsequent time indexes in a window whose limits are set as being ±25% of the previous lag value. This range of ±25% represents the limit of possible period-to-period changes for human vocal cords [Hess, 1983].

The tracking in the neighbourhood of a given lag value represents that tracking

of a candidate pitch contour and this continues in M_H until there are no peaks within the range of allowable pitch period lag values when it switches over to test, M_L, the matrix thresholded with the low threshold, t_L. As these peaks are visited by the algorithm over the course of a contour, they are removed from both the high and low thresholded matrices, M_H and M_L, so that these same peaks will not be considered in future passes by the algorithm. The tracked lag values in the contour are placed in a third matrix, referred to as the *contour matrix*, denoted as D.

This process is repeated until all the non-zero entries in M_H have been visited, and in this process, all contours lasting less than 8 consecutive time indexes are discarded for the reasons described earlier, namely because these short contours are usually due to short, locally periodic noise bursts.

As a given contour is tracked, a cumulative distance measure is computed for each time index according to the following formula:

$$D(l \pm 25\%, i+1) = (cm(l \pm 25\%, i+1) + 1)^2 + D(l, i) \tag{3.4}$$

where $cm(\cdot)$ is the value of the crosscorrelation maxima occurring at a lag in a neighbourhood of $\pm 25\%$ of the current lag value l at time instant i, for the following time instant $i+1$. After all the non-zero entries in M_H have been visited, the algorithm moves to the next and final portion of the post-processing phase.

3.1.5 Distance Processing

As alluded to in section 3.1.4, the calculation of a distance measure for accurate F_0 extraction is necessary due to the fact that the lowest lag crosscorrelation maxima which falls above a certain threshold is not necessarily the pitch period. Nor is the pitch period the lag with the maximum crosscorrelation value. However, the correct pitch period lag will have the majority of crosscorrelation value maximums for the majority of time frames in a given contour. With this in mind, the distance

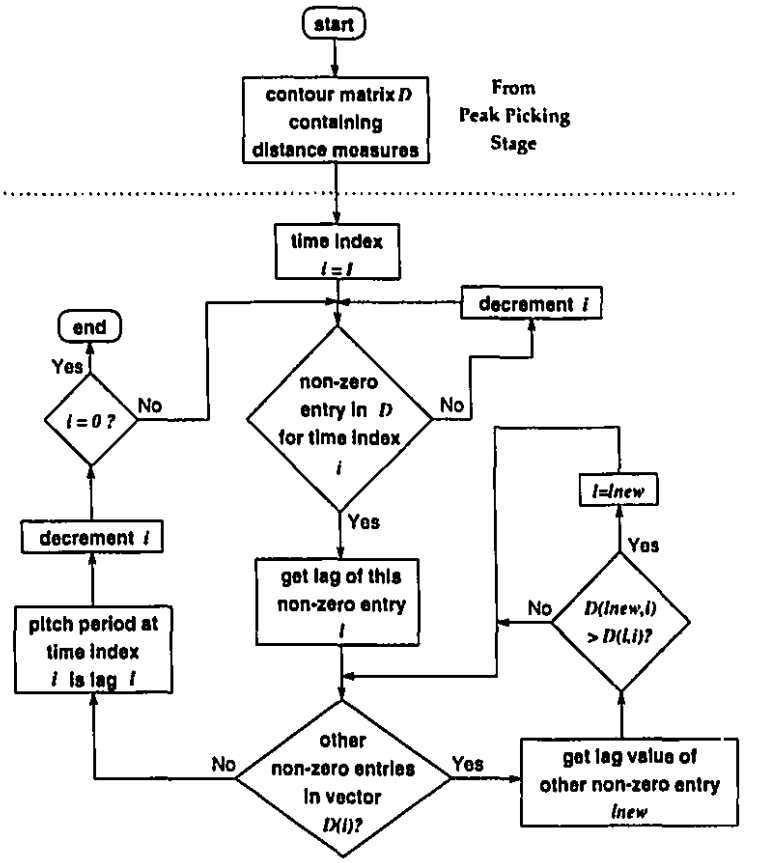


Figure 3.9: Flow Chart of the Distance Analysis Stage of the Post-Processing Phase of the Pitch Extractor

measure of equation 3.4 was formulated. At the end of a contour in a voiced section of a recording, the lag value corresponding to the contour with the largest distance value will correspond to the true pitch period contour. Using this information coupled with additional heuristics, the distance processing algorithm proceeds with the distance analysis shown in figure 3.9.

The distance measures calculated for every lag of every non-zero entry at every time index are stored in matrix D, referred to as the *contour matrix*. The algorithm takes this matrix and begins from the last time index of the contour matrix and proceeds backwards to the start of the matrix or the initial time index. Thus, it should be noted that the following description of the algorithm is described from the perspective of moving along the decreasing time index. The algorithm searches for the presence of a crosscorrelation distance maxima, indicated by the occurrence

of a non-zero valued entry in a given lag value for a particular time index. If none exist for a particular time index, the algorithm proceeds backwards in time, decrementing the time index counter until one such entry is found in the contour matrix. If at any given time index, more than one non-zero entry is present, the scores of the contours are checked, and the one with the highest distance score is picked as the pitch period contour.

It should be noted that these contour "ends" are treated as a particular case. For the case of extremely weak signals, periodicity only appears at lags corresponding to multiples of the true pitch period. Typically, these non-zero entries in the contour matrix occurring at the end of a voiced section due to sub-harmonic period values, only last for a few indexes and never for more than 8 time indexes. So, once the first non-zero lag entry is found for the first time at a given time index, the algorithm then proceeds to examine the 8 earlier time indexes to observe if there are any other contour peaks at lower lag values (higher frequency). If there are, the scores of the contours are checked for sections of the contour for the length of the shorter of the two contours being compared. The one with the maximum score is chosen to be the winning contour. Tracking of that lag value initiates from this point, and the other contours with lower scores are discarded. The lag which is tracked from this point represents the true pitch period contour.

The algorithm then proceeds by moving backwards in time, by decreasing the time index value in order to traverse the matrix D. While tracking a given contour, one of the following three events may occur: the current contour being tracked ends, a new contour appears at a lower lag value (higher frequency), or a new contour appears at a larger lag value (lower frequency).

The first event represents the start of the utterance and the point from which the algorithm proceeds, looking for other peak contours as it does when the algorithm first begins. The second event represents the appearance of a contour at a lower lag value, which could be the result of a return to the "true" pitch period value after a

double harmonic break episode, or due to the occurrence of a narrow bandwidth first formant, F_1, which occurs at multiple of the true F_0, in that particular time instant in the utterance. The score of the new contour is then checked against that of the current contour for the length of the shorter of the two contours. If the score of the new contour never exceeds the score of the current contour for the length of the shorter of the two contours, then the new contour is due to a strong, narrow bandwidth F_1 effect, and is discarded. If the score of the new contour goes above that of the current contour, then, the new contour represents the end of the "true" pitch period lag and the current contour represents a double harmonic break episode. Tracking then resumes about the new contour.

The third event occurs during a double harmonic break episode, where the period of the vocal cord vibrations essentially doubles, which corresponds to a halving of F_0. This represents a return to the true pitch value after one of the aforementioned episodes. These types of episodes are common in certain types of cries, including pain and some other physiological disorders [Wasz-Höckert *et al.*, 1968, Wasz-Höckert *et al.*, 1985]. For adults, the occurrence of these types of events are not very common, but their occurrence may be due to the presence of abnormal growths on the vocal cords [Kasuya *et al.*, 1983]. For this type of event, when the current contour ends and another is present at a greater lag value (lower frequency), the algorithm begins tracking about the new contour, looking once again for the occurrence of one of these three events.

Dysphonic episodes, where an energy smearing occurs across the entire frequency band and no clear harmonics are present in this portion of the vocalization, commonly occur in pain cries as well. A spectrogram of an utterance containing a dysphonic episode can be seen in figure 3.10. In this figure, dysphonia can been in the interval from time 0.03 seconds to 0.15 seconds where the clear harmonic peaks, denoted by the dark bands occurring prior to and following the dysphonic episode disappear, and are replaced by a noise-like spectrum.

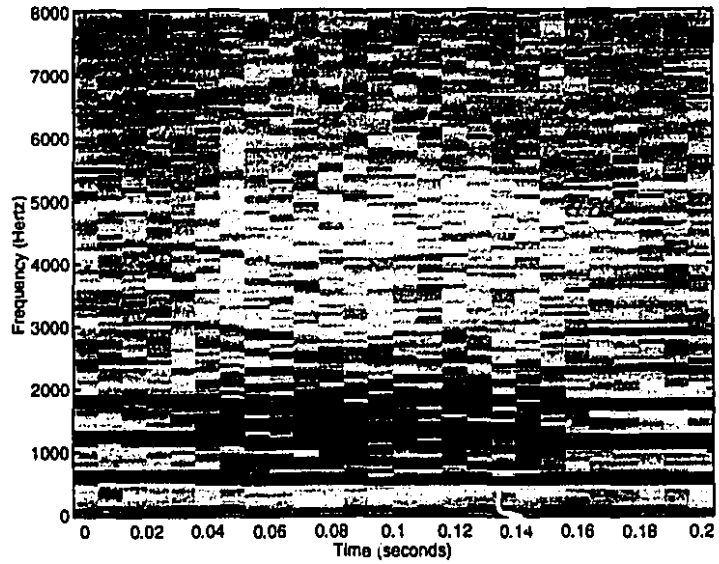


Figure 3.10: Spectrogram of a Cry Utterance Containing a Dysphonic Episode

During such episodes, the algorithm would proceed as follows. Since there is no periodicity present during dysphonic episodes, the crosscorrelation values will, for the most part, fall below the low maxima threshold value of $t_L = 0.6$, and, with a few erratic exceptions, will remain below the high maxima threshold of $t_H = 0.8$. In the event that there are a few maxima in this type of episode which are within the same neighbourhood of a particular lag value, the duration of these contours will be very short in duration, lasting only a few time indexes, but which will always last less than 8 time indexes, as previously mentioned. Consequently, these contours will be removed during the crosscorrelation maxima post-processing stage described earlier in section 3.1.4.

What results following this distance analysis part of the post-processing stage are non-zero values in the contour matrix D at particular lags and time indexes, which correspond to the true pitch period values. This implies that for any given time index, there will be at most, only one non-zero entry. The lag at which the non-zero entry occurs, corresponds to the pitch period for that time index. The results of this processing applied to some test cry utterances will be illustrated in section 3.4, and compared with other popular pitch extraction routines in section 3.2.

3.1.6 Implementation and Computational Considerations

The code for the implementation of the pitch extraction algorithm described in the previous section was implemented using versions 4.0a to 4.2a of the high-performance numeric computation and visualization software MATLAB, developed by the Mathworks Incorporated [Mat, 1992]. MATLAB stands for MATrix LABoratory and supplies a number of numerical analysis, signal processing, and graphics rendering routines in an interactive environment. It is also possible to call MATLAB from inside "C" or FORTRAN programs using a series of function calls, thus allowing a fast computational engine to be incorporated as part of an an external application program. Once a MATLAB routine has been fully tested and debugged, it can be compiled in a pseudo-C format providing links to external functions for even faster execution.

The fundamental frequency extraction routines described in the previous sections were implemented using three separate programs. One routine performs the signal transformation phase of sections 3.1.2 and 3.1.3, calculating the cross-correlation vectors, and grouping them into a matrix which is indexed in time. This routine was implemented using approximately 300 lines of MATLAB code. Another routine performs the post-processing task of the peak-picking and distance contour calculation described in section 3.1.4. The third routine performs the distance analysis and the final pitch period determination process described in section 3.1.5. The latter two routines are implemented in 150 and 135 lines of MATLAB code respectively.

Computationally speaking, the method, especially the signal transformation phase, is particularly intensive. The post-processing stages of the crosscorrelation vector matrix are quite fast relative to the signal transformation portion of the method. Intuitively, this can be understood by the fact that for a given time index, the computation of the normalized crosscorrelation requires $O(N)$ computations, where N corresponds to the number of lags in the expected range of pitch period

values, which for infant cries, is especially large, as was previously mentioned.

As well, the length of time required to generate the sequence of crosscorrelation vectors depends not only on the length of the utterance, but also depends on the duration of the voiced portions in a particular recording, and what the pitch periods of the voiced portions are. Since the time index for the crosscorrelation vector calculation advances in time increments related to the pitch period of that portion of the cry signal, an utterance with a low fundamental frequency will be traversed much more rapidly than one with a much higher pitch, since one pitch period represents a larger increment in the time index for the latter case than for the former.

This same is also true when considering portions of recordings where there are long segments of silence or noise. In these cases, the time index is incremented by the lag value for which the normalized crosscorrelation function was the largest, irrespective of whether or not this value exceeded the threshold value. For portions where there may be locally periodic noise bursts with relatively high frequency values, the method will move more slowly across these potions of the recording than it will along silent portions, where the larger maxima values in the crosscorrelation vector occur at larger lag values.

The subsequent post-processing handles the crosscorrelation vector matrix, and thus the time required to complete the pitch period extraction is dependent upon the number of crosscorrelation vectors in the matrix, I. The peak-picking process requires $O(NI)$, as does the distance analysis process, which finally yields the pitch period values for a given recording.

With all the above in mind, it is still be useful to discuss some typical computational times in order to illustrate the typical time required to process a cry recording. Using the experimental set-up described in section 3.3 and the aforementioned routines, a cry recording lasting 3 seconds with voiced portions with an average pitch period of approximately 32 samples (500 Hz) lasting for about 2.5 seconds requires

approximately 3 minutes and 30 seconds in the signal transformation phase, and 40 seconds in the post-processing phase, of which 30 seconds are spent thresholding and calculating the distance values for the contour matrix, and the final 10 seconds are spent in the distance analysis portion of this final phase.

The current implementation, and the very nature of this processing method, precludes a real-time implementation, although some improvements in processing speed are suggested in section 5.1. Despite the non real-time nature of this method, and the seemingly intensive nature of the computations performed during its execution, even without the proposed improvements for increased speed mentioned in section 5.1.1, still requires less time, provides better resolution, and more importantly, more accuracy than the more popular F_0 extraction techniques which have been borrowed from the speech domain, and applied to infant cry utterances. This will be illustrated in section 3.4 and discussed in section 3.5.

3.2 Comparison with Other Methods

Prior to the design and development of the improved crosscorrelation vector-based fundamental frequency extraction method described in section 3.1, a number of the classical more popular, and more successful of the methods used for fundamental frequency extraction for adult speech signals, which operate only on the input signal, were implemented and tested. In this section, the results of these methods applied to infant cry vocalizations are compared to the improved crosscorrelation vector-based method. First, however, some background on the methods implemented and tested for comparative purposes will be given. These methods are the linear predictive (LPC) residual for F_0 estimation [Maksym, 1973] and it's popular variant, the spectral inverse filter tracking (SIFT) algorithm [Markel, 1973], cepstral pitch extraction [Noll, 1967], the harmonic sieve F_0 extraction routine [Sluyter *et al.*, 1982], spectral flattening by clipping the speech signal [Sondhi, 1968], correlogram-based pitch extraction

[Slaney and Lyon, 1990], and the crosscorrelation-based super-resolution pitch determination method [Medan *et al.*, 1991]. The following sub-sections briefly describe each of these methods, and outlining the problems that these methods encounter when processing infant cry signals, as these issues have not been previously discussed in any great depth or published in the literature.

3.2.1 Linear Predictive Coding (LPC) and the Simplified Inverse Filter Tracking (SIFT) Algorithms

Overview of LPC

Linear predictive coding (LPC) is one of the most popular speech analysis and fundamental frequency extraction methods and it has also been used in speech coding applications as well [Reddy and Swamy, 1984]. The underlying reason behind the popularity of LPC is due to its accuracy in representing the spectral characteristics of the input signal, and to the relatively simple computations required to accomplish this task. LPC assumes that the speech signal to be modelled is generated by an all pole filter, which represents the vocal tract, excited by a periodic pulse train, which represents the glottal pulses produced from the vocal cords. The all-pole assumption does not hold if there are zeros in the spectrum due to nasal phonemes in speech or from to unvoiced sounds [O'Shaughnessy, 1987]. Nevertheless, this all-pole simplification is not a major source of errors for the majority of speech signals.

The presumption behind the speech production process in LPC is that an excitation source $U(z)$ excites an all pole shaping filter

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}}, \tag{3.5}$$

yielding an output speech signal $\hat{S}(z)$, which is similar to the actual speech output $S(z)$ in a least-square error sense. For voiced sounds, excitation source $U(z)$ is viewed as being a uniform sample pulse train. In order to obtain an estimate of $H(z)$, for a given input frame of data, the speech signal is considered to be stationary within a given analysis frame. In order to determine the LPC coefficients, a_k, one of two methods can be used; the least-squares autocorrelation method, or the least-squares covariance method [Markel and Gray Jr., 1976]. Because of its greater simplicity, the least squares autocorrelation method is the one which is most commonly used to determine the LPC coefficients.

When determining the order of the poles, p, used to model the spectrum of the input frame, the following is taken into consideration. Typically, two poles are required to model each formant resonance of the vocal tract, with two to four additional poles used to model the zeros in the spectrum, so that voiced signals can be matched with reasonably good accuracy. Unvoiced sections and silence, however, result in a very poor spectral match, and this fact can be used to determine which frames are voiced, and which are not [Atal and Rabiner, 1976]. To assist in the modelling of the input speech frame, the spectrum of the input speech signal is flattened using pre-emphasis prior to LPC analysis. This enables the higher frequency formants to be modelled by effectively reducing the dynamic range of the input spectrum, countering the attenuating effect of the vocal tract at higher frequencies.

After the spectrum of an input speech frame has been modelled and the LPC coefficients, a_k, determined, the input speech samples can then be inverse filtered through the all-pole filter. What emerges from this process is referred to a the residual signal and contains either a periodic pulse train, if the signal is voiced, or a noisy signal, otherwise. An autocorrelation can be performed on the residual to determine if periodicity is present, and if so, what the period of the pulses is, effectively determining the pitch period of the input frame.

In order to save both computation time and to reduce predictor order, p, for the purposes of pitch extraction, the input speech signal can be low-pass filtered and decimated at a frequency value above the frequency of the first formant resonance peak, so that for speech signals, a low-pass filter with a cut-off of about 1000 Hz is usually used.

Decimation refers to the process of converting a signal from a given sampling rate to a lower sampling rate. This process can be achieved using one of two general methods. One way is to pass the signal through a digital to analogue converter, filter the signal if necessary, and then resample the analogue signal at the desired sampling rate. A second method is to perform the sampling rate conversion entirely in the digital domain [Proakis and Manolakis, 1988].

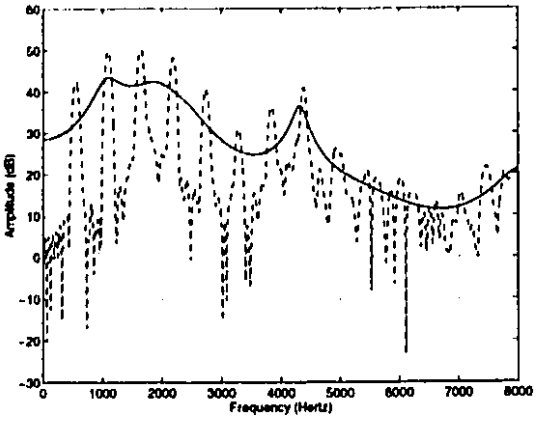
The LPC modelling of the reduced bandwidth spectrum is then performed on the reduced bandwidth signal. This process requires less poles to model this sole resonance in the input spectrum. The low-passed input signal is then inverse filtered and then an autocorrelation is taken of this residual. This method is commonly know as the simplified inverse filter tracking (SIFT) algorithm [Markel, 1972b].

Problems with LPC and Infant Cry Signals

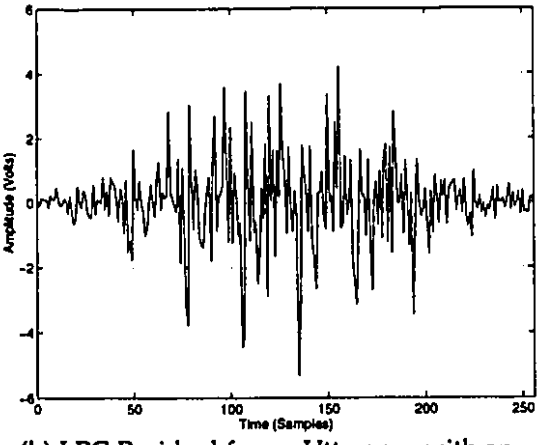
As has already been mentioned, this method is an especially popular and very successful method for pitch extraction of adult speech. As was mentioned in section 2.1 of chapter 2, many improvements to this method have been proposed since this method was first introduced in 1972. Despite these improvements, its use on infant cry signals is not as successful as it is for speech, and this is due to a number of reasons. First, this method assumes and requires that the input signal be stationary within a given frame, and that there be approximately three or four pitch periods per input frame. Hence a frame size must be chosen which can accommodate the range of expected pitch periods while not containing too high a number of pitch periods so that the stationarity assumption no longer holds.

For adult speech, this is not a problem since the fundamental frequency range is mostly limited to values between 60 Hz and 300 Hz (16 ms to 3 ms). This range of F_0 can be accommodated by a 48 ms window without violating stationarity. For infant cry signals, however, the range of F_0 can go from values as low as 100 Hz to values in excess of 2500 Hz (10 ms to 0.4 ms). A window which would contain 3 pitch periods of an utterance with an F_0 of 100 Hz would also contain 75 periods of an utterance with and F_0 of 2200 Hz. In the latter case, it is clear that stationarity cannot be assumed, and because of this, the method would yield an average value of the pitch periods contained within the window or a grossly incorrect value due to amplitude variations across such a large range of values. Any small variations in the pitch periods contained within the window or frame would be lost to these undesirable effects. Consequently, no one window or frame size can accommodate the wide range of expected F_0 values without either violating stationarity, or risk having too few samples inside a given window when the fundamental frequency is low. Nevertheless, a fixed window size based on the frequency range of the F_0 values of a particular utterance was used for experimental purposes, in order to evaluate the operation of this method on infant cries, even if, in an automated system, this sort of "adaptive window sizing" could not be performed.

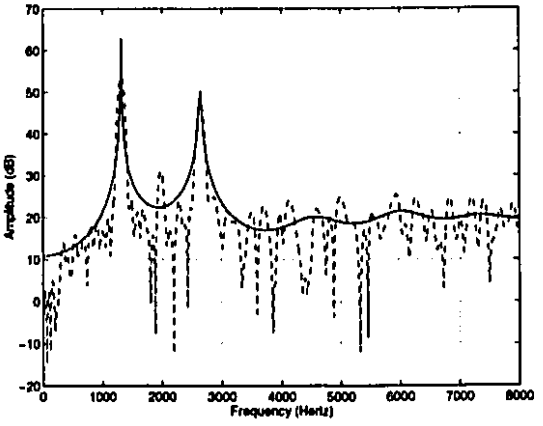
Another problem with LPC when used on infant cry signals is the determination of the number of poles required to model the cry spectrum. Based on infant vocal tract size, the first, second, and third formant values are expected to occur at about 1100 Hz, 3300 Hz, and 5500 Hz respectively [Golub and Corwin, 1985]. Consequently, given the bandwidth of the cry recordings described in section 3.3, ten poles were used to model the vocal tract shape; six for the three expected formant peaks, with the remaining four poles used to model any zeros occurring in the input spectrum. If the F_0 in a given analysis frame is in the vicinity of 400 Hz, then there are a sufficient number of harmonic peaks under each formant so that the least-square error modelling of the spectrum will indeed track the spectral envelope, and not the harmonic peaks [O'Shaughnessy, 1987]. However, if the



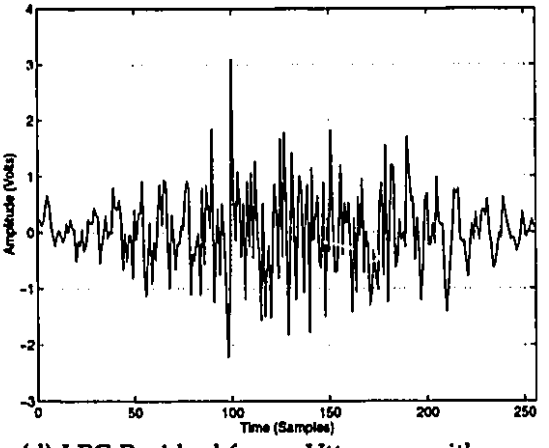
(a) LPC (solid) and Original Spectrum (dashed) for an Utterance with an approximate F_0 of 500 Hz



(b) LPC Residual for an Utterance with an approximate F_0 of 500 Hz



(c) LPC (solid) and Original Spectrum (dashed) for an Utterance with an approximate F_0 of 1300 Hz



(d) LPC Residual for an Utterance with an approximate F_0 of 1300 Hz

Figure 3.11: LPC (solid line) with Original (dashed) Spectra and Residuals for Cries with an F_0 value of about 500 Hz and 1300 Hz

value of F_0 increases much beyond this value, this will no longer be the case, and the poles will model the harmonic peaks since the number of poles p will be approximately equal to the number of harmonics in the spectrum. Since the effect of these harmonic peaks will be removed from the input spectrum during the inverse filtering process, the residual will not show the presence of a periodic pulse train, as is expected for a voiced signal.

Figure 3.11 illustrates this point. Although both inverse filtered residual signals are rather "noisy", the one for the utterance with the higher F_0 value, shown in

figure 3.11(d), shows no clear periodicity since the poles in the LPC spectrum match the two harmonics of the pre-emphasized spectrum. For the utterance with the lower F_0 value, however, the poles model the resonant frequencies in the spectrum of the input signal frame, as shown in figure 3.11(a). Consequently, the inverse filtered residual of figure 3.11(b) shows sharp peaks corresponding to the pitch period starting at approximately sample 50.

Decreasing the number of poles only postpones the fundamental frequency value at which this problem will occur. For signals with F_0 values greater than 1000 Hz, there will only be either three or four harmonics present in the spectrum, due to the attenuating effects of the vocal tract. In these cases, even if the number of poles are reduced from ten, as mentioned above, to four, to accommodate these higher frequency signals, the spectral modelling will follow the harmonics, and not the spectral envelope, and, in turn, will remove all traces of a periodic pulse train from the LPC residual.

3.2.2 Cepstral Pitch Extraction

Overview of the Cepstral Method

Cepstral analysis is another way of deconvolving the filter and excitation components of a speech signal [Noll, 1967]. This process transforms the product of two signals into a sum of two signals. If the two signals are very different spectrally, then it is possible to separate them using a simple linear filtering operation.

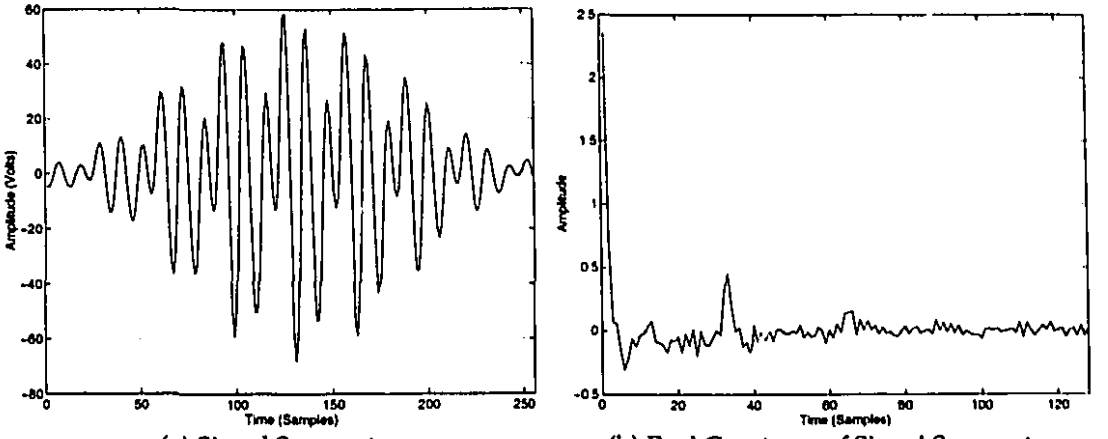
For speech signals the two components of interest are the excitation and vocal tract response, for which the speech signal $s(n)$ can be viewed as being the convolution of the excitation $e(n)$, and the vocal tract tract response $v(n)$. In the frequency domain $S(z) = V(z)E(z)$, so if we take the logarithm of $S(z)$ we get $\log(S(z)) = \log(V(z)) + \log(E(z))$. The inverse transform of $\log(S(z))$ is defined

as the cepstrum, also referred to as $\hat{s}(n)$, where the caret denotes the cepstra. So $\hat{s}(n) = \hat{v}(n) + \hat{e}(n)$, with $\hat{v}(n)$ decaying to zero over the first few milliseconds, inversely proportional to the resonant frequencies, or formants, of the vocal tract, and $\hat{e}(n)$ appearing as a periodic pulse train at multiples of the pitch period. The cepstrum $\hat{s}(n)$ is a complex-valued quantity but information regarding the periodicity in the input signal can be derived from using only the real portion of the inverse transform of the log spectrum, or $real(\hat{s}(b))$.

For adult male speech, for example, the first excitation peak would be expected to appear in the range of 5 ms to 16 ms of the cepstrum, which corresponds to F_0 values from 200 Hz to 60 Hz respectively. The vocal tract excitation would end in the cepstrum at a time approximately equal to the inverse of the first formant, which at its lowest value occurs at about 3.5 ms (285 Hz). In this case, then, the vocal tract and excitation sources can be separated by considering values occurring in the cepstrum from 0 ms to 4.5 ms, and attributing these contributions to the vocal tract response, and considering the occurrence of the first sharp peak occurring after 4.5 ms as being the pitch period. This linear separation process is called *liftering*.

Problems with Cepstral Processing and Infant Cry Signals

Since the cepstrum also uses a fixed window of signal samples on which it performs the required processing, it is subject to the same considerations regarding window size as were mentioned in the preceding section. As is the case with speech which contains F_0 values higher, than, say 400 Hz, as is the case for female speech or children's speech, the separation between the excitation and and vocal tract response contributions in the cepstrum is not so neat or clear cut. The lowest of the formant frequency values, namely F_1, has an approximate value of about 1100 Hz for infant cries, which would correspond to a peak occurring at 0.9 ms in the cepstrum. Consequently, based on the discussion in the previous subsubsection, values of the cepstrum between 0 ms and 1 ms would be considered



(a) Signal Segment (b) Real Cepstrum of Signal Segment

Figure 3.12: Signal Segment and Its Corresponding Real Cepstrum

as containing the vocal tract components. Thus, in cases where the F_0 in the cry being analyzed falls below 1000 Hz, the liftering operation can easily separate the vocal tract and excitation components. If the F_0 were to go above 1000 Hz, as commonly occurs with cries, the two components cannot be separated through liftering since the excitation contribution will be contained within the range where values corresponding to the vocal tract contribution are expected to be found.

If the lifter separating the two components in the cepstrum was set at 0.45 ms, to handle the expected range of pitch period values for infant cry utterances, then the first sharp peak occurring above this time threshold would be considered as corresponding to the pitch period. Since, however, the area between 0.45 ms and 0.9 ms falls in the range of the inverse of the first formant frequency, it is possible that for certain cries where F_1 has a very narrow bandwidth and contains a relatively large amount energy, that this component will have a very sharp peak in the cepstrum, and will consequently be tagged as corresponding to the pitch period or the excitation component of the signal. In similar cases, it is very likely that vocal tract responses be incorrectly chosen as the pitch period, and as such, this method only works reliably for cry utterances with F_0 values less than 1000 Hz, making it unsuitable for pitch period extraction from infant cries in general. Figure 3.12 shows a signal segment and its corresponding real cepstrum, which clearly shows

the occurrence of a sharp peak at sample 34, corresponding to a period of 2.125 ms or an F_0 of 470.6 Hz.

Also, some confusion exists as to what exactly constitutes a sharp cepstral peak, since there is no clear correlation between the strength of a cepstral peak and whether the segment under consideration is voiced or not.

3.2.3 The Harmonic Sieve

Overview of Method

This method, originally proposed by Goldstein [Goldstein, 1973], but implemented by Duifhuis, Willems and Sluyter [Duifhuis *et al.*, 1982], uses a harmonic sieve preceded by an implementation of Goldstein's theory of hearing [Goldstein *et al.*, 1978], in order determine the best fit for an input stimuli containing only a few spectral components, using a maximum likelihood criterion. Basically, the pitch determination method consists of two elements: a spectral analyzer that detects and measures the frequency of the harmonic components, and a harmonic pattern recognizer.

First, the fast Fourier transform (FFT) of a given input frame of signal samples is presented for the subsequent spectral analysis. Next, the effect of frequency masking of the spectral components on each other is determined based on the Goldstein theory of hearing. In this portion of the method, two thresholds are used for the purpose of determining if and which frequency components are masked by other components in the spectrum, or if certain components are too weak to be considered altogether. The latter threshold is an absolute threshold, which reflects the limit of audibility, and is set to a value of 26 dB below the highest peak level in the spectrum of the input frame. The former, however, is a relative threshold, which takes effect with respect to the amplitude of the other spectral components,

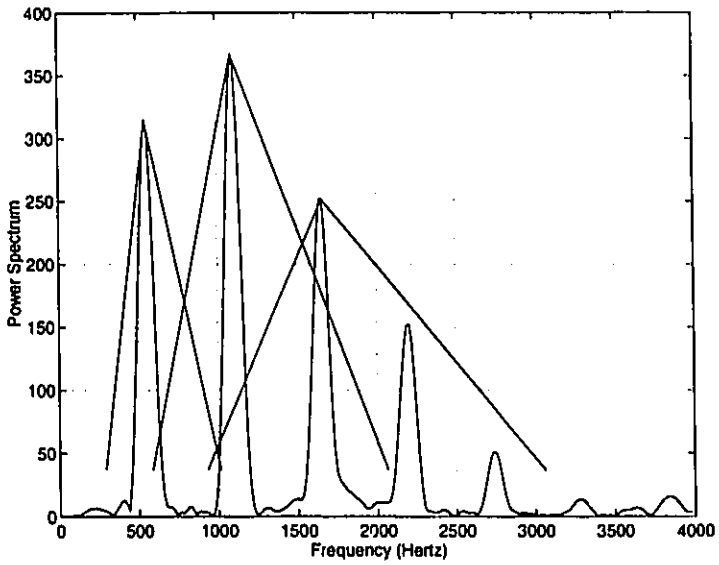


Figure 3.13: Power Spectrum with Goldstein's Theory of Hearing Masking Thresholds

and is based on the psychophysical masking threshold. This threshold is set as follows; for a given spectral component, the masking threshold is set to a value of 90 dB/octave on the low frequency side of the component, and to 45 dB/octave on the high frequency side. All spectral components falling under these thresholds are considered as being masked by the component under consideration and are thus removed from the component set. These threshold values are considered to roughly correspond to the critical band filter characteristics of the human ear.

Figure 3.13 shows a power spectrum of a cry utterance signal segment with the relative masking threshold lines derived from the given spectral peaks. Note that the fourth and fifth harmonics in the spectrum fall under the 45 dB/octave masking threshold on the high frequency side of the third harmonic and would thus be removed from the component set which would be presented to the sieve. The sixth and seventh harmonic fall under the absolute threshold representing the limit of audibility.

Once all the spectral components have been processed by this initial stage, the remaining components are sent to the harmonic sieve, in order to determine which of these components are the true harmonics for a given fundamental frequency can-

didate value in the sieve, and which are spurious. The sieve contains "meshes" of bandwidth proportional to the value of their center frequencies, and, the "meshes", correspondingly, get wider as their center frequency increases, allowing for slight variations in the fundamental frequency of the signal contained within a given input frame to be tolerated. The lowest center frequency values of the meshes in the sieve span the range of expected frequency values for the input signal, with step sizes being less than the value of the mesh width so that no portion of the frequency scale is missed during this sieving process.

For each of the sieves, which are characterized by their fundamental frequency value, the number and location of components which fall through the sieve are checked, and are labelled according to their candidate harmonic number. Based on this set of candidate harmonic numbers for all the values in the expected F_0 range, it is then decided which of the candidate harmonics correspond to the optimum set.

This is determined by taking a normalized distance measure for all the harmonic sets which is in turn calculated by taking the number of the highest candidate harmonic, or spectral component, adding the number of input harmonics, and dividing by the number of classified harmonic components in the spectrum. The number of unclassified spectral components, or components which do not "fall through" a given sieve, increases the distance value for sieves centered at frequency multiples of the true F_0.

Problems with Harmonic Sieving and Infant Cry Signals

The problem with this method when applied to infant infant cry signals is, first and foremost, the number of computations involved for a given input signal frame for the range of expected F_0 values. Also a number of frequency errors arise from the Goldstein theory of hearing pre-processing phase of this method when the masking of the fundamental frequency component occurs, due to particularly strong

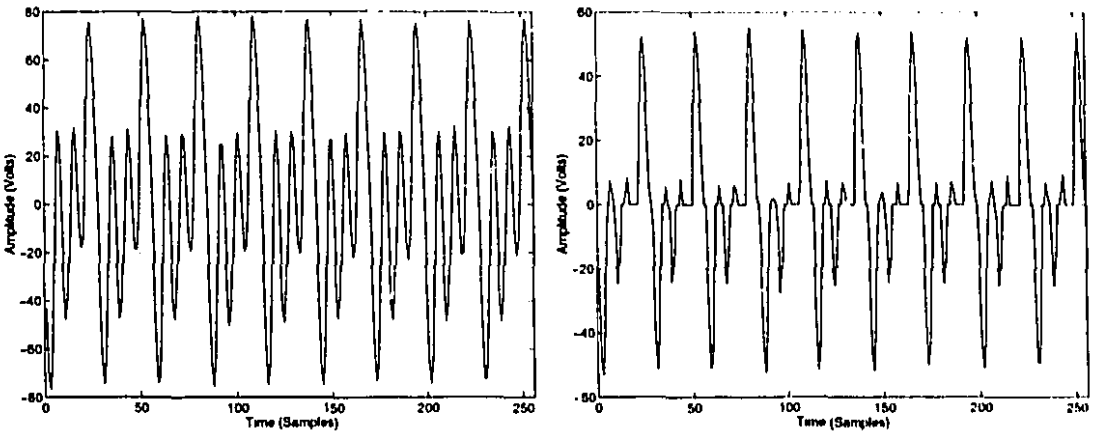
formant values, which in turn increase the amplitude of the harmonic peak located at this particular frequency. This has the effect of decreasing the score for the sieve centered at the correct F_0 value, whereas sieves centered at multiples of F_0, especially $2F_0$, being chosen as the correct F_0 value. This anomaly was also observed by Duifhuis, Willems and Sluyter, the researchers who originally implemented this method, when dealing with speech signals with similar characteristics to those described above. Their solution to this problem was to use tracking to follow a certain pitch value, as long as its distance score remained below a certain threshold. In addition to decreasing the number of gross pitch errors, it also decreases the number of calculations required by limiting the range of frequencies to be sieved while a specific F_0 is tracked.

This solution works for speech since the values of F_0 normally remain around the value of the preceding F_0 value, obtained from the previous input signal frame. Such an assumption cannot be made for cry utterances, however, especially since for certain types of cry utterances double harmonic break episodes commonly occur. Although this method is novel in the way that it uses the theory of hearing formulated by Goldstein to improve the "standard" harmonic sieve or matching process, it is still not an optimal solution for infant cry signals.

3.2.4 Pitch Extraction by Spectral Flattening

Overview of Method

Another method popular due to the relative simplicity of the computations required and its ability to enhance the periodicity in the signal is the clipping autocorrelation method, and the variations of this technique, which were originally proposed by Sondhi [Sondhi, 1968]. In his paper, Sondhi describes three different spectral flattening methods, namely spectrum flattening followed by a minimum phase phase correction for synchronization of the harmonics, spectrum flattening fol-



(a) Signal Segment (b) Clipped Signal Segment $k = 0.3$

Figure 3.14: Signal Section and Clipped Signal Section

lowed by autocorrelation, and non-linear distortion followed by autocorrelation. Of these three methods, the latter two are described as being the best suited for the fundamental frequency extraction of speech due to the simplicity of the required computations, relative to the first method, as well as the superior ability of this method to distinguish between formant peaks and pitch epoch peaks following the autocorrelation of the non-linearly distorted signal.

Aside from the simplicity of this method which makes it especially appealing, it is also suitable for a real-time implementation. In Sondhi's description of the method, a 30 ms segment of speech is taken, and in every 5 ms portion of the signal, the maximum absolute value of the signal, a_0, is found, and all values between $\pm ka_0$ are removed from the signal. A typical value of k is 0.3. Figure 3.14 shows a signal segment and the signal segment subsequent to the clipping operation. Following this "clipping" operation, the autocorrelation of the clipped signal is performed. The lag of the first crosscorrelation maxima found which exceeds a certain threshold is chosen as corresponding to the pitch period. As subsequent frames are processed, this threshold is progressively reduced for maxima values in the vicinity of lag values of the previous pitch period value. The original threshold is restored if voicing ends, of if the pitch changes abruptly.

Problems with Spectral Flattening and Infant Cry Utterances

As is the case with the other frame-based methods mentioned in the previous subsections, the issue of frame size is once again an important consideration. Here, however, there is an additional issue regarding the length of time within a given frame in which the maximum absolute signal value, a_0, will be determined. As well, the value of the clipping threshold k is also an issue that requires careful consideration. Some infant cries have important variations in amplitude between pitch periods, so that a certain threshold value would include these all the peaks for a particular value of k, but exclude other lower amplitude peaks, for example, in other sections of the input frame, leading to pitch halving errors.

Moreover, as was mentioned in section 3.1.2, some cry signals have a very small decay in the amplitude between the pitch epoch and the subsequent periodic peak due to a high energy, narrow bandwidth formant occurring at a frequency approximately equal to $2F_0$, something which does not normally appear in adult speech signals. Using a clipping value threshold, k, of 0.3 would still include the these F_1 peaks in the clipped signal. Thus, the use of the threshold and the segment range in which it is to be applied over, should be adaptive, which implies having a priori knowledge of the signal characteristics, which is not possible.

Consequently, although it is a simple and effective method for speech signals, it can only be used on cry signals with a limited F_0 range. Using one method with one frame size and threshold for the expected range of F_0 values works with some F_0 values but compromises performance of other values.

3.2.5 Correlogram-Based Pitch Extraction

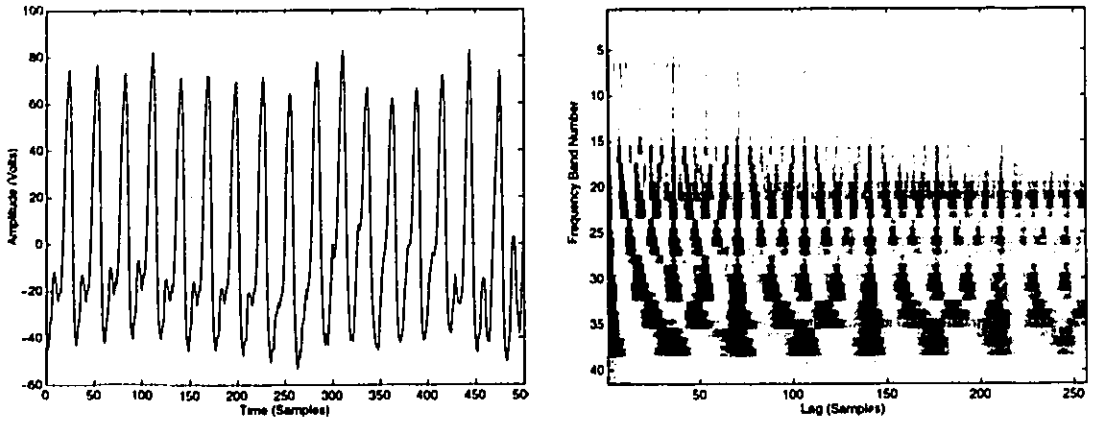
Overview of Method

This method was developed by Malcolm Slaney in the research lab of Apple Computer in California [Slaney and Lyon, 1990]. The implementation of the correlogram-based pitch extraction algorithm, available through an anonymous `ftp` site on the Internet, consists of a number of different MATLAB files, each of which performs a different function in the algorithm [Slaney, 1994]. Basically, Slaney's pitch detector is based on Licklider's "Duplex Theory" of pitch perception, which is believed to accurately model how humans perceive pitch [Seneff, 1978]. This pitch detector combines a cochlear model, which separates the signal into different frequency bands, which is then followed by a bank of autocorrelators, which perform independent autocorrelation for each channel. The outputs of the individual channels are combined in a visual manner, called a *correlogram*, which is then subsequently filtered, non-linearly enhanced, and summed across all channels before a pitch estimate is formed from this information.

An example of a cry utterance segment with pitch period of approximately 35 samples (640 Hz) and its corresponding correlogram is shown in figure 3.15.

Problems with Correlogram-Based Pitch Extraction and Cry Utterances

Once again, as has been the case for all the other methods discussed in this section, this method is also faced with appropriate frame size concerns, and since no one size can accommodate the entire range of expected frequency values, the frame size must be tailored to the characteristics of the cry recording being analyzed. The problem with this method lies in its computational complexity, due to the complex implementation of the cochlear processing stage. This portion of the processing requires a significant amount of time for the calculation of F_0 on a given segment or frame of data. For example, processing of a 16 ms window of data (256



(a) Cry Utterance Segment (b) Correlogram of Cry Utterance Segment

Figure 3.15: Cry Utterance Segment and Corresponding Correlogram

samples of a signal sampled at 16 kHz), this algorithm requires over 30 seconds on a SPARC 10+ to calculate a pitch candidate value. If consecutive frames of 16 ms each overlapping by 50% with the previous frame samples are taken for a 5 second utterance, the algorithm requires approximately 2 hours to generate pitch values for all the frames, compared to a time of just under 5 minutes for the improved crosscorrelation vector-based method described in section 3.1. Due to these prohibitive computation times, the algorithm by default uses no overlap, and actually spaces subsequent frames by 1000 samples, which risks losing the occurrence of some important transitions or events for cry utterances. As well, this method is not free from the pitch halving or doubling errors present in other methods, as will be shown when results are presented in section 3.4.

3.2.6 Super-Resolution Pitch Extraction

Brief Overview of Method and Problems with Processing Cry Utterances

This method originally proposed by Medan, Yair, and Chazon [Medan *et al.*, 1991], and the signal transformation phase of the improved crosscorrelation vector based

fundamental frequency extraction method presented in section 3.1.3, both use the normalized crosscorrelation method as a means of computing a set of pitch candidates [Chung and Algazi, 1985]. The algorithm presented by Medan, Yair, and Chazon tracks the F_0 contour, however it does not make use of the crosscorrelation vectors generated by the crosscorrelation computation. Tracking a specific pitch period value begins after pitch period values for the previous four or five time indexes are within $\pm 25\%$ of the value of the previous candidate. This heuristic performs well for speech, but fails for cry utterances, and yields a number of gross pitch errors during diplophonic or double harmonic break episodes, as will be shown in section 3.4. These events, as has previously been mentioned, are common in certain types of cry vocalizations, and it is important that these events be properly handled by a given F_0 extraction method.

Another method by De Mori and Omologo proposed a variation of Medan, Yair, and Chazon's algorithm by making use of the crosscorrelation vectors resulting from the normalized crosscorrelation computations to calculate a cumulative distance measure for the duration of the recording, using the observation that F_0 contours for adult speech remain within the same neighbourhood, once a given pitch period is found [De Mori and Omologo, 1993]. This algorithm, however, increments the time index in fixed steps, not in increments corresponding to the most likely pitch period candidate lag for a given time index. Once again, the algorithm works well for speech; specifically for speech that does not have the occurrence of diplophonic episodes, and this, in turn, leads to inconsistent results for diplophonic or double harmonic break episodes in cry signals as well.

3.3 Data Set and Experimental Set-up

The data set used for the testing and validation of the pitch period extraction methodology proposed in section 3.1 above consisted of 230 cry episodes recorded at the Nôtre-Dame-de-Grâce CLSC (Community Health Clinic) from sixteen two

to six month old infants and 329 cry episodes recorded at the Royal Victoria Hospital in Montreal from premature infants ranging in gestational age from 24 to 36 weeks. None of the infants involved in the study had a history of perinatal or postnatal complications. All the parents of the infants gave their informed consent to participate in this study.

For the data set consisting of infants of normal gestational age, the cry episodes recorded were the results of one of three stimulus events: **pain / distress** from routine immunization procedure; **fear / startle** from a jack-in-the-box; and **anger / frustration** from a head restraint. For the premature infant data set, the cry episodes recorded were also the result of one of three stimulus events: a needle stick in the infant's heel as part of a routine immunization procedure, a washing and disinfecting of the heel with a cotton pad prior to the hell stick, and a gentle squeeze of the heel.

All the cry vocalization recordings were made on a Sony TCM-500DEV audio cassette recorder with an omni-directional Senheiser MKE2 microphone placed 10 centimeters away from the infant's mouth. Once recorded, the signals were then low-pass filtered to 8 kHz, prior to digitization using a 16 kHz sampling rate and a 12-bit analogue-to-digital converter. These digitized recordings were then transferred to a SPARC 10+ for further processing and analysis.

Prior to their use in the various F_0 extraction routines tested, the recordings were high-pass filtered using a 301 tap finite impulse response (FIR) filter with a cutoff of 240 Hz designed using the Remez Exchange Algorithm provided by MATLAB [Oppenheim and Schafer, 1975]. The motivation behind the use of an FIR filter with numerous taps was to achieve zero frequency distortion, due to the inherent linear phase characteristics of FIR filters, while achieving a stop-band attenuation of approximately 30 dB [Proakis and Manolakis, 1988]. The frequency response of the FIR high-pass filter is shown in figure 3.16 for the frequency range from 0 Hz (DC) to 400 Hz. From 400 Hz to 8 kHz, the frequency response is flat at 0 dB.

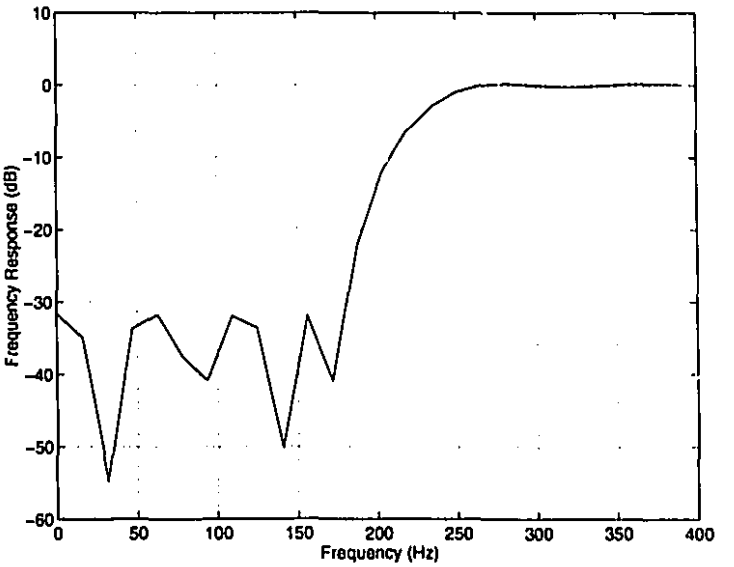


Figure 3.16: Frequency Response of FIR High-Pass Filter

Other high-pass filters were also investigated, namely Chebyshev type *I* and type *II* infinite impulse response (IIR) filters, because of the large stop band attenuation and sharp filter roll-off achievable from these filters using a small filter order. Due to the phase distortion inherent in IIR filters, their use was not pursued, in the interest of minimizing signal distortion subsequent to filtering, and at the expense of increasing the number of filter taps required by an order of magnitude.

Section 3.4.1 will present in more detail the files which will be used to compare the results obtained by using the method proposed in section 3.1 with those of section 3.2.

3.4 Results

This section compares the results of the methods presented in sections 3.2 with the improved crosscorrelation vector-based pitch period extraction method of section 3.1 on five different utterances by presenting both the extracted pitch tracks and a table of error rates, broken down by error type, as generated by the re-

spective methods. Since a comparison of pitch extraction methods for infant cry vocalization has not previously been discussed or presented in the literature, a corresponding treatment as to the limitations, shortcomings, and desired improvements in existing methods have not been reported. Consequently, this section and section 3.2 will attempt to rectify this fact while illustrating the improvements of the improved crosscorrelation vector-based method presented in section 3.1.

This section begins with a description of the individual recordings used in the pitch period extraction tests, displaying the spectrograms of these recordings and providing a verbal description regarding the relevant features in these particular files. Next, the implementation of the methods of section 3.2 is briefly discussed. This is then followed by a presentation of the pitch contours extracted by the methods which were presented in section 3.2 and the improved crosscorrelation vector-based pitch extractor described in section 3.1. Error rates will then be presented in tabular form for comparative purposes.

3.4.1 Recordings Used in the Evaluation

This section gives both an illustrative and descriptive treatment of the infant cry recordings used for testing the various pitch extraction algorithms. Although the set of five recordings presented here is by no means an exhaustive set, it is representative of the type of vocalizations which are commonly found in the cries of both premature infants and full-term neonates.

The files used for testing were:

1. A02004: An anger/frustration cry from a full-term infant,
2. A07104: A second anger/frustration cry from another full-term infant,
3. B056ST: A pain/distress cry from a premature infant,
4. C1213SQ3: A second, less painful, cry from a premature infant, and
5. P09102: A pain/distress cry from a full-term infant.

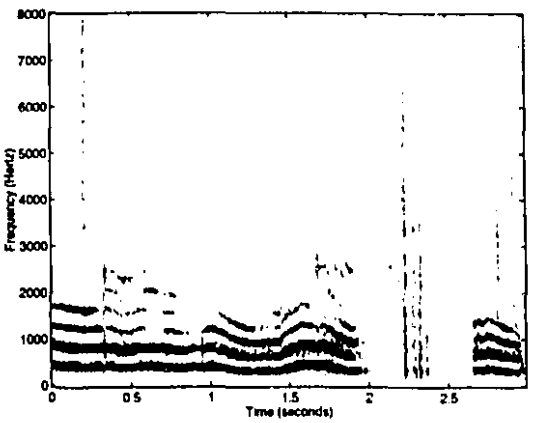


Figure 3.17: Spectrogram of A02004 (An Anger Cry from a Full-Term Infant)

File A02004

The spectrogram of file A02004 is shown in figure 3.17. This file represents the cry of a full-term infant uttered when its head was restrained, and is labelled as being an anger or frustration cry. This recording contains two voiced utterances or episodes. The first voiced utterance is characterized by a rather flat fundamental frequency contour, with an initial F_0 value of approximately 485 Hz , with some episodes of noise occurring at a number of points between the start and the end of the utterance two seconds later. From the spectrogram, it can be seen that just after the one second mark, the pitch decreases for about half a second before beginning to increase once again. Approximately 0.25 seconds later, the pitch begins to decrease until the end of the utterance. The second voiced utterance in this recording is relatively brief, and begins with an initial F_0 value of approximately 400 Hz. This episode also follows a short rising and falling fundamental frequency pattern. This file was selected to illustrate how the different methods would perform on relatively smooth contours, punctuated with some episodes of ambient noise.

File A07104

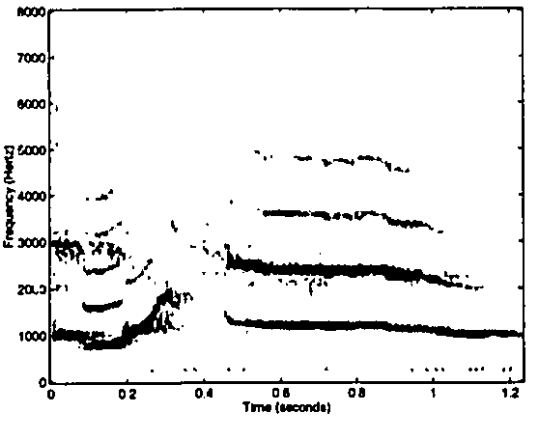


Figure 3.18: Spectrogram of A07104 (An Anger Cry from a Another Full-Term Infant)

File A07104, whose spectrogram is shown in figure 3.18, corresponds to another recording of an anger or frustration cry uttered by a full-term neonate. The characteristics of this utterance are appreciably different from those of the previous anger / frustration recording. This file contains two voiced utterances with high fundamental frequency values. The first voiced episode features a rather rapid and abrupt change from a value of about 1000 Hz to a value of about 800 Hz in pitch, which is then followed by a section with a rapidly increasing F_0, which lasts until the end of the utterance at the 0.3 second mark. The second utterance begins at around 0.5 seconds with an F_0 of about 1250 Hz, and features a narrow bandwidth F_1 occurring at a frequency approximately equal to twice that of the fundamental frequency, as indicated by the significantly darker colour of the second harmonic, which lasts until the one second mark of the spectrogram. For the remainder of the recording, the F_1 value decreases in both value and bandwidth. This file was selected to illustrate how the different pitch extraction methods would perform on an utterance with an unusually high pitch and on utterances that have very little decay between the pitch epoch peak and the subsequent peak in the signal due to a narrow bandwidth formant with a frequency approximately equal to that of the second harmonic.

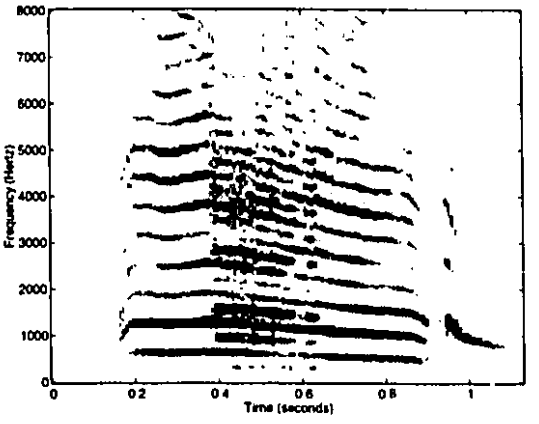


Figure 3.19: Spectrogram of B056ST (A Pain Cry from a Premature Infant)

File B056ST

The next recording used in tests was file B056ST, and its spectrogram can be seen in figure 3.19. This cry was recorded after a premature infant received an immunization needle in its heel. As was the case for the previous two files, this file also contains two voiced utterances. The first one begins at approximately the 0.15 second mark with a fundamental frequency of about 530 Hz which rapidly increases to a value of about 640 Hz. For this initial portion of the first utterance, the cry has a narrow bandwidth first formant occurring at a frequency of $2F_0$, or about 1300 Hz. Just prior to the 0.4 second mark, a double-harmonic break episode begins, which is almost immediately followed at about the 0.45 second mark by a brief dysphonic episode, where there is no periodicity present in the signal as the vocal folds vibrate in a chaotic manner leading to a smear in the energy values across the spectrum. Following this brief dysphonic episode, at about the 0.5 second mark, the signal resumes its double harmonic break episode, where the F_0 is essentially halved, until about the 0.65 second mark, with a brief return to the original F_0 value just before the 0.6 second mark. Following this double harmonic break episode, F_0 returns to a value in the vicinity of 600 Hz, slowly decreasing in value until the end of the episode at the 0.9 second mark. The second voiced episode beginning at the 0.95 second mark consists of an inspiratory phonation starting with an initial pitch of about 1500 Hz which decreases in value until the 1.1

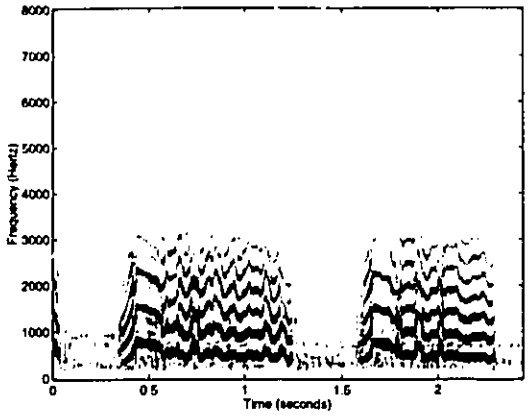


Figure 3.20: Spectrogram of C1213SQ3 (A Pain Cry from Another Premature Infant)

second mark. This file was selected to illustrate how the different pitch extraction methods would behave during and after a double-harmonic break episode, and on an inspiratory phonation.

File C1213SQ3

The fourth file in the set is C1213SQ3 and its spectrogram is shown in figure 3.20. This file contains the tail end of a vocalization followed by two complete episodes uttered by a premature infant after its heel was squeezed. The recording begins by catching the end of a phonation which has a fundamental frequency of about 640 Hz which decreases during its short duration from the start of the recording to the 0.05 second mark. The second voiced utterance begins 0.3 seconds into the recording with an F_0 of about 400 Hz and rapidly increases to a value of about 750 Hz. The F_0 contour then varies rapidly until the end of the voiced episode, about 1.25 seconds into the recording. The last voiced utterance begins shortly after the 1.5 second mark and has similar F_0 characteristics to that of the previous utterance, ending 2.3 seconds into the recording. This recording was selected to illustrate how the different pitch extraction methods would behave on a file that has very rapidly varying pitch values.

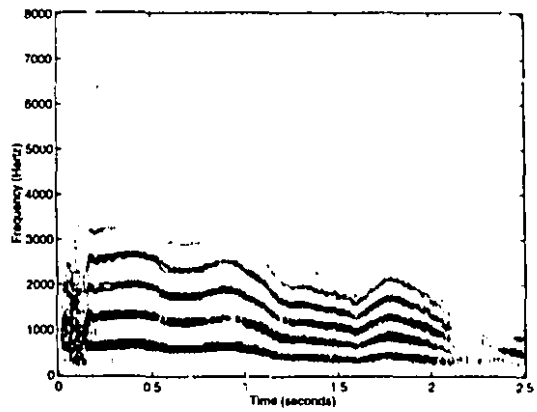


Figure 3.21: Spectrogram of P09102 (A Pain Cry from a Full-Term Infant)

File P09102

The last file in the test set is P09102 has its spectrogram shown in figure 3.21. This recording was made following the immunization of a full-term infant, and thus represents a pain or distress cry. This recording begins with a brief portion of phonation with an F_0 of about 640 Hz. At about 0.125 seconds into the recording, a long voiced utterance begins with a brief section containing a high fundamental frequency, which then quickly drops to a value of about 727 Hz. The F_0 contour follows a slightly increasing slope until 0.5 seconds into the recording, at which point the values start to decrease. This decrease lasts for about 0.25 seconds before F_0 begins to increase once again for another 0.25 seconds. After this, the values decrease rapidly until the 1.6 second mark. Then, there is a final portion with rising-falling F_0 pattern which ends this voiced utterance shortly after the 2 second mark. Shortly before the end of the recording another brief phonation occurs with a double harmonic break episode at the start of the contour which is then followed by a sharp increase to a value of twice the initial F_0. This recording was selected to illustrate how the different pitch extraction routines would track a relative smooth and stable F_0 contour preceded by a brief high F_0 burst, and a weak double harmonic break episode.

3.4.2 Implementation of Pitch Extraction Methods

This subsection briefly presents the characteristics of the various elements used in the pitch extraction routines of section 3.2, stating information such as the window size, whether the signal was pre-emphasized or decimated prior to analysis, and what method was used for voiced/unvoiced determination. All of the routines were implemented using MATLAB.

Linear Predictive Coding

This method was implemented using about 115 lines of MATLAB code. The signal was subject to pre-emphasis prior to analysis, and was segmented into fixed windows each containing 256 samples, corresponding to a duration of 16 ms, each of which was tapered using a Hamming window. Subsequent frames overlapped by 66%. The spectrum of the Hamming windowed sections were modelled using 12 poles and the autocorrelation method. A particular frame was labelled as voiced if the predictor error was less than 0.4, and unvoiced otherwise.

Simplified Inverse Filter Tracking (SIFT)

This method was implemented using about 240 lines of MATLAB code. The original signal was low-pass filtered using a linear-phase finite impulse response filter with a cut-off of just under 4 kHz and then decimated by 2. The decimated signal was subject to pre-emphasis prior to analysis, and was segmented into fixed windows each containing 128 samples, corresponding to a duration of 16 ms. Each of these windows were tapered using a Hamming window, with subsequent frames overlapping by 66%. The spectrum of the Hamming windowed sections were modelled using 6 poles and the autocorrelation method. The method described by Markel, which uses a combination of the voicing flags from previous segments and the correlation values of the current frame, was implemented for voiced-unvoiced

determination [Markel, 1973].

Cepstral Pitch Extraction

This method was implemented using about 90 lines of MATLAB code. The signal was segmented into consecutive windows of 256 samples, corresponding to 16 ms portions of the signal, which were also tapered using a Hamming window, with subsequent windows overlapping by 50%. The real portion of the inverse Fourier transformed log magnitude spectrum was used for the pitch determination process. If the energy of a specific frame exceeded a value of 3.3 dB, the segment was labelled as voiced. Otherwise, the segment was labeled as being unvoiced. This simple voiced-unvoiced determination method was used since there is no clear correlation between the value of a cepstral peak and whether or not a specific segment is voiced or unvoiced [Noll, 1967].

Harmonic Sieve

The implementation of the harmonic sieve, preceded by an implementation of Goldstein's theory of hearing, was performed using 300 lines of MATLAB code. The original 16 kHz sampled signal was low-pass filtered using a linear-phase finite impulse response filter with a cut-off of just under 4 kHz and then decimated by 2. The decimated signal was then divided into consecutive frames each containing 128 samples, corresponding to a 16 ms duration, with subsequent frames overlapping by 50%. All windows were tapered using a Hamming window. The method was implemented as described by Duifhuis, Willems, and Sluyter and included an implementation of their voiced-unvoiced determination method, which is based on the score of the sieve with the lowest value representing the most-likely pitch candidate [Duifhuis *et al.*, 1982]. If the score is below a particular value, the segment is labelled as voiced and is labelled as being unvoiced otherwise.

Spectral Flattening Autocorrelation Method (SFAC)

The implementation of this method was done using 150 lines of MATLAB code. The signal was segmented into windows each with a length of 256 samples, or 16 ms, with subsequent windows overlapping by 66%. Since the voiced-to-unvoiced determination method mentioned in section 3.2.4 led to numerous errors, an alternate method of making this determination was adopted. If the energy of a particular window was less than 3.0 dB, the window was labelled as unvoiced and was labelled as voiced otherwise.

Correlogram-Based Pitch Extraction

This was implemented using the MATLAB routines in the "Auditory Toolbox" developed by Slaney to perform the various processing stages of this method as outlined in section 3.2.5 [Slaney, 1994]. The signal was segmented into 256 sample windows which overlapped by 50%.

Super-Resolution Pitch Extraction

This method was implemented using 490 lines of MATLAB code as was described in the paper by Medan, Yair, and Chazon [Medan *et al.*, 1991]. Once the cross-correlation value at a specific time index exceeded the adaptive threshold, it was labelled as voiced, and was otherwise labelled as being unvoiced. To improve the resolution of the pitch values extracted by this method, the pitch period was interpolated between sample values, with subsequent frames being advanced by the extracted pitch period, in samples.

3.4.3 Error Analysis Results

This subsection illustrates the pitch contours extracted by the various pitch extraction methods on the five test files.

Figures 3.22 to 3.26 show the pitch contours extracted from the test files described in section 3.4.1 using six methods borrowed from the speech domain which were described in section 3.2 and whose implementation was presented in section 3.4.2. The results for these methods are shown in sub-figures (a) to (f). The last sub-figure in each of these figures, labelled (g), displays the F_0 contour obtained from the improved crosscorrelation vector-based F_0 extraction method presented in section 3.1, whose implementation was described in section 3.1.6.

The figures illustrate the extracted fundamental frequency versus time, where the fundamental frequency is simply the extracted pitch period divided by the sampling rate of 16 kHz.

As was done in the classical fundamental frequency extraction review paper of Rabiner, Cheng, Rosenberg, and McConegal [Rabiner *et al.*, 1976], six different error parameters were computed. For every utterance in the test set a reference pitch contour which is denoted by $p_r(m)$, determined by inspection of the pitch value for every pitch period, and averaging the F_0 values within a window for the frame-based methods. The extracted pitch contour is denoted by, $p_e(m)$, where $e = 1 \ldots E$ and E denotes the number of pitch detectors used in these tests. Here, seven pitch detectors are compared, so that $E = 7$. By comparing the reference pitch contour $p_r(m)$ with the extracted pitch contour $p_e(m)$ for every e, that is, for each of the seven pitch extraction methods tested, and for every m, that is, for each interval or section, either voiced or unvoiced in the recording, one of the four following events can occur.

1. $p_r(m) = 0, p_e(m) = 0$, in which case both the reference and extracted contours have classified the interval m as unvoiced.

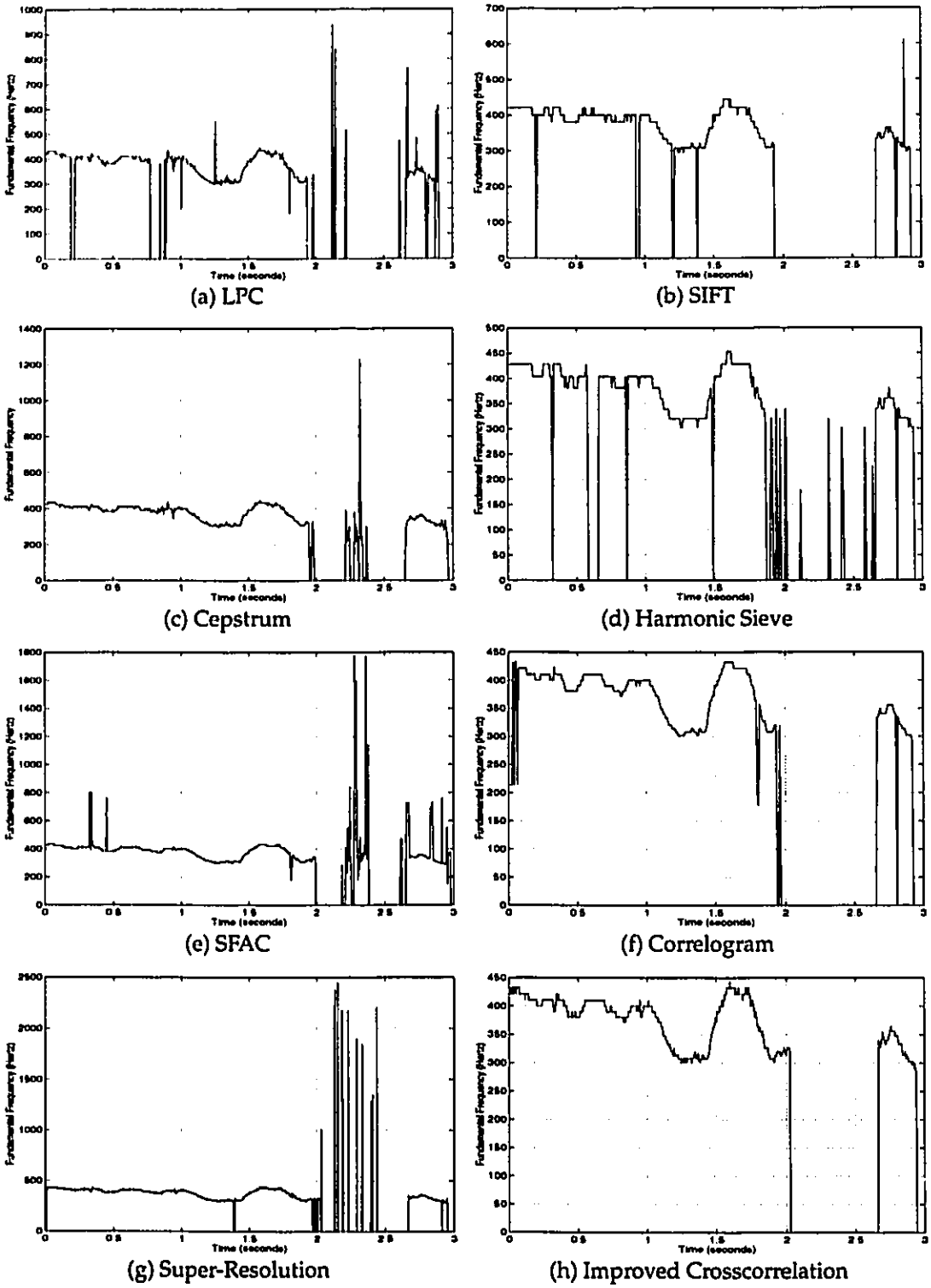


Figure 3.22: Pitch Contours for Recording A02004

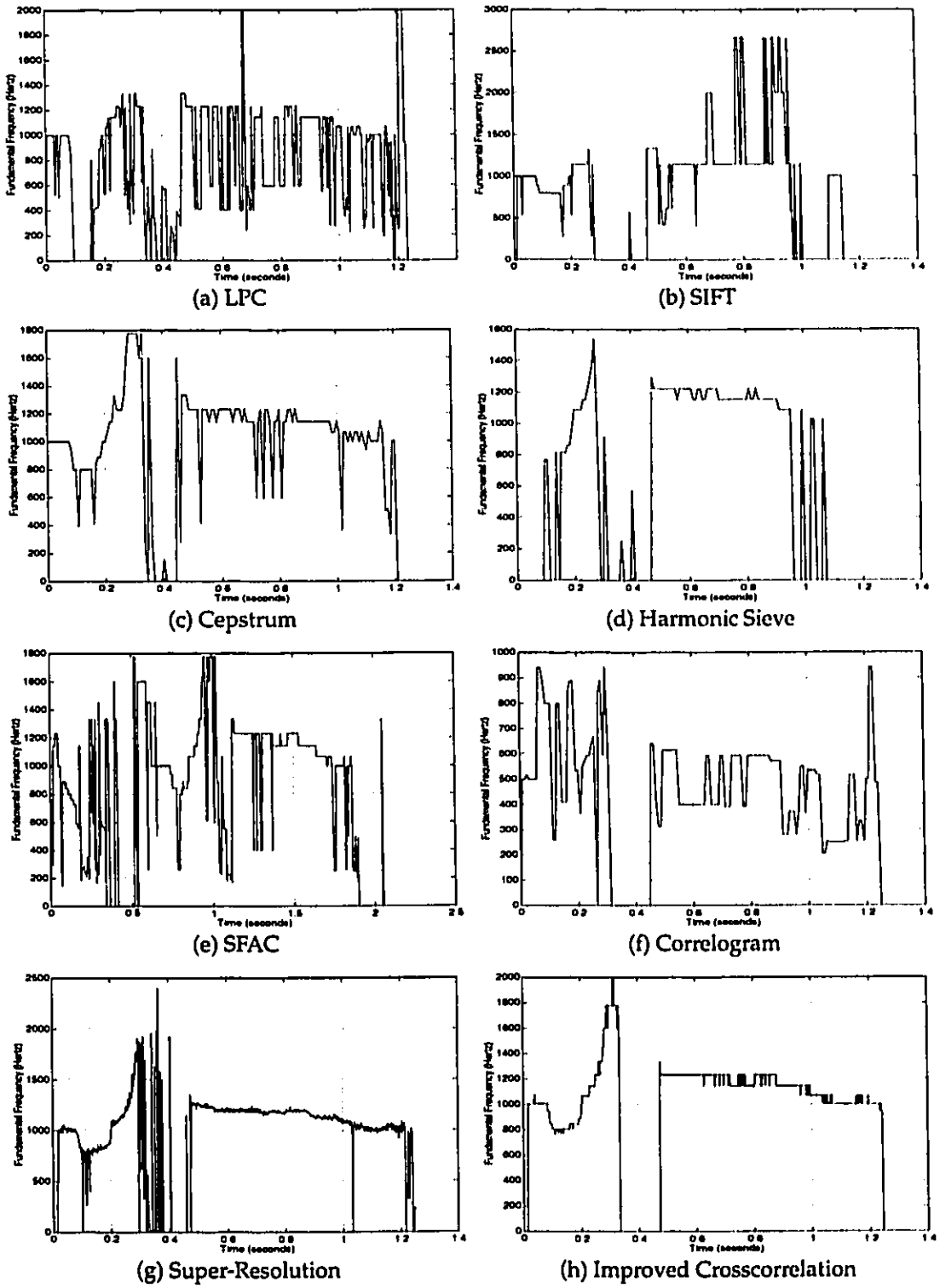


Figure 3.23: Pitch Contours for Recording A07104

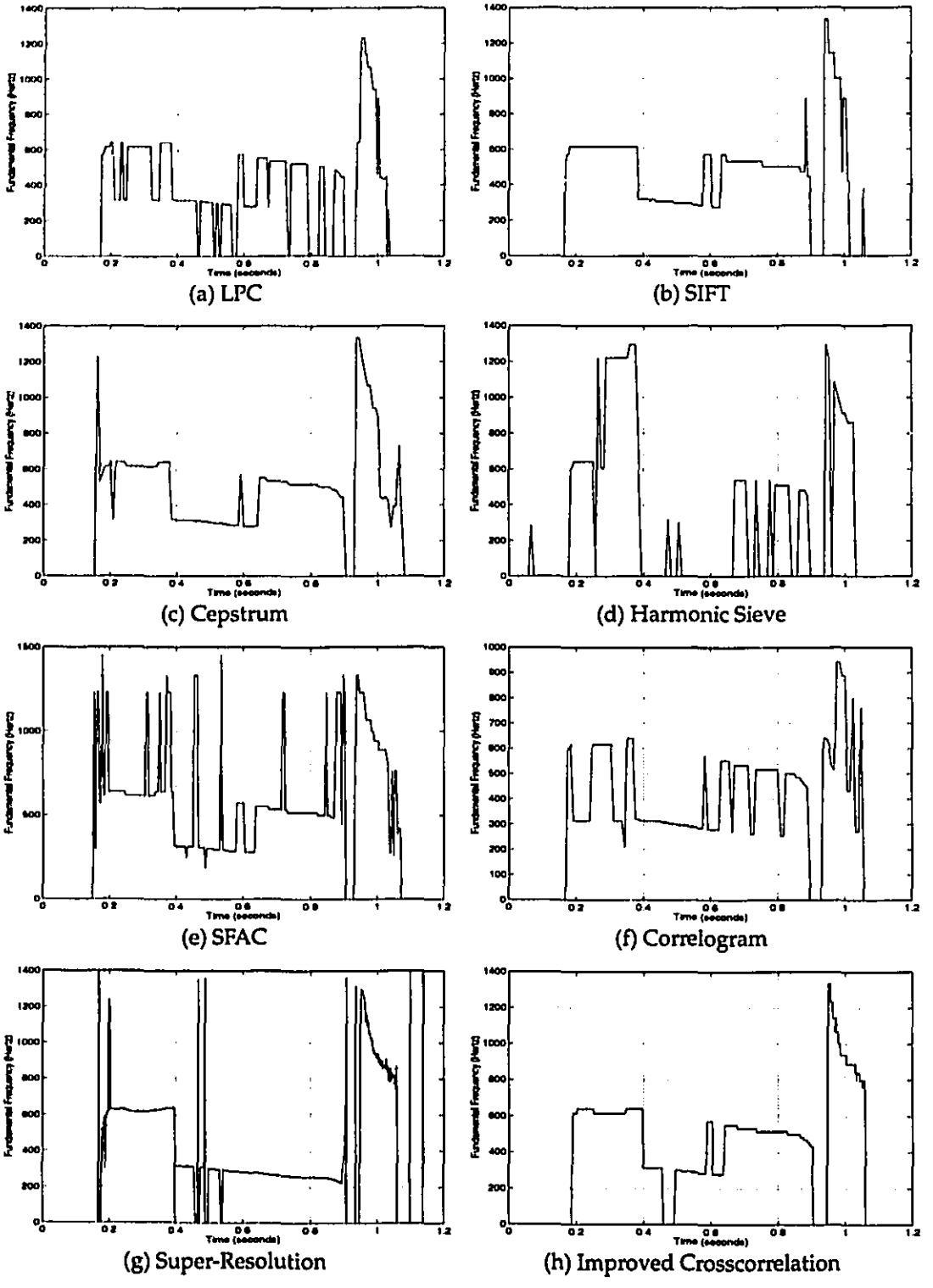


Figure 3.24: Pitch Contours for Recording B056ST

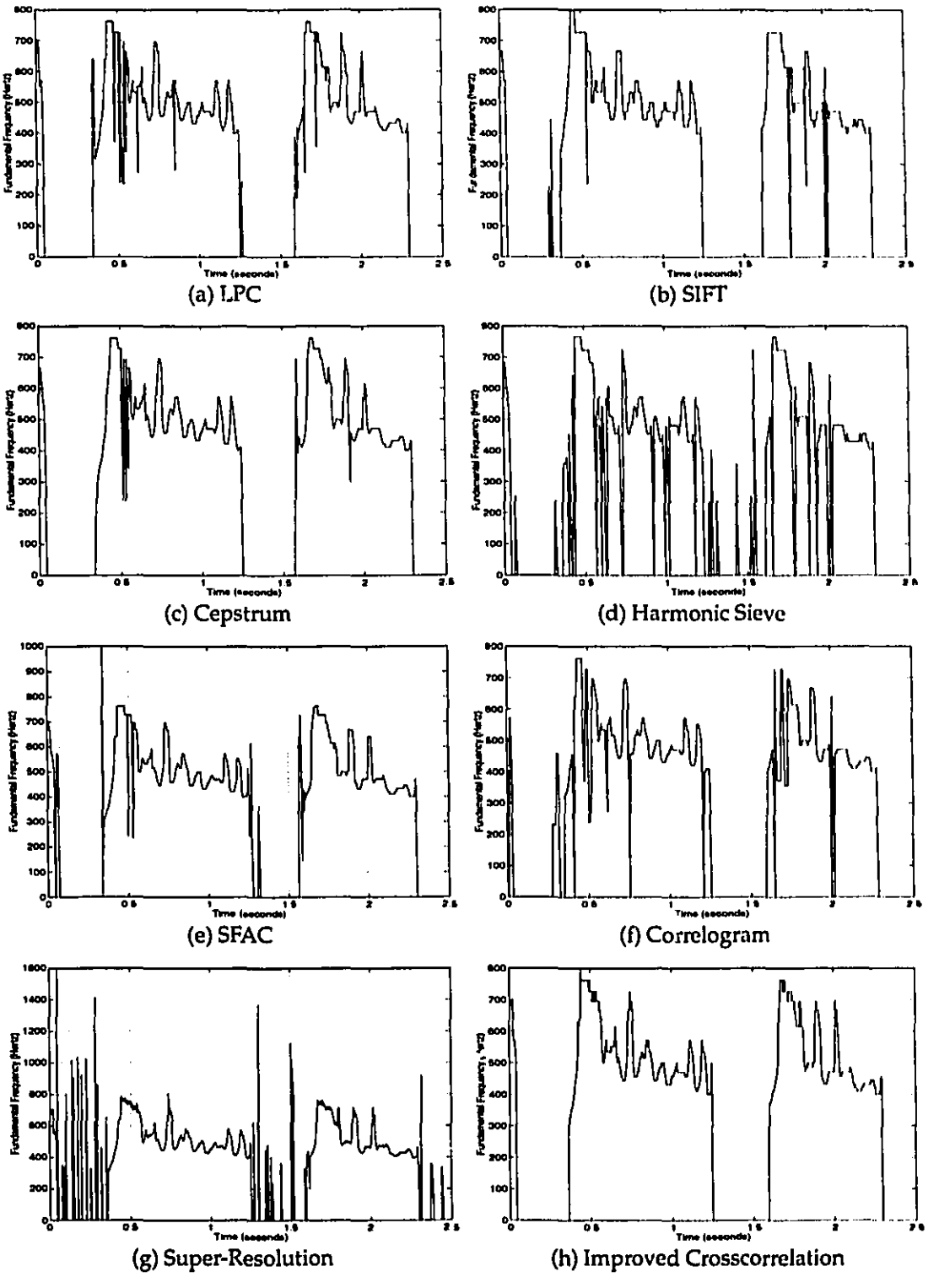


Figure 3.25: Pitch Contours for Recording C1213SQ3

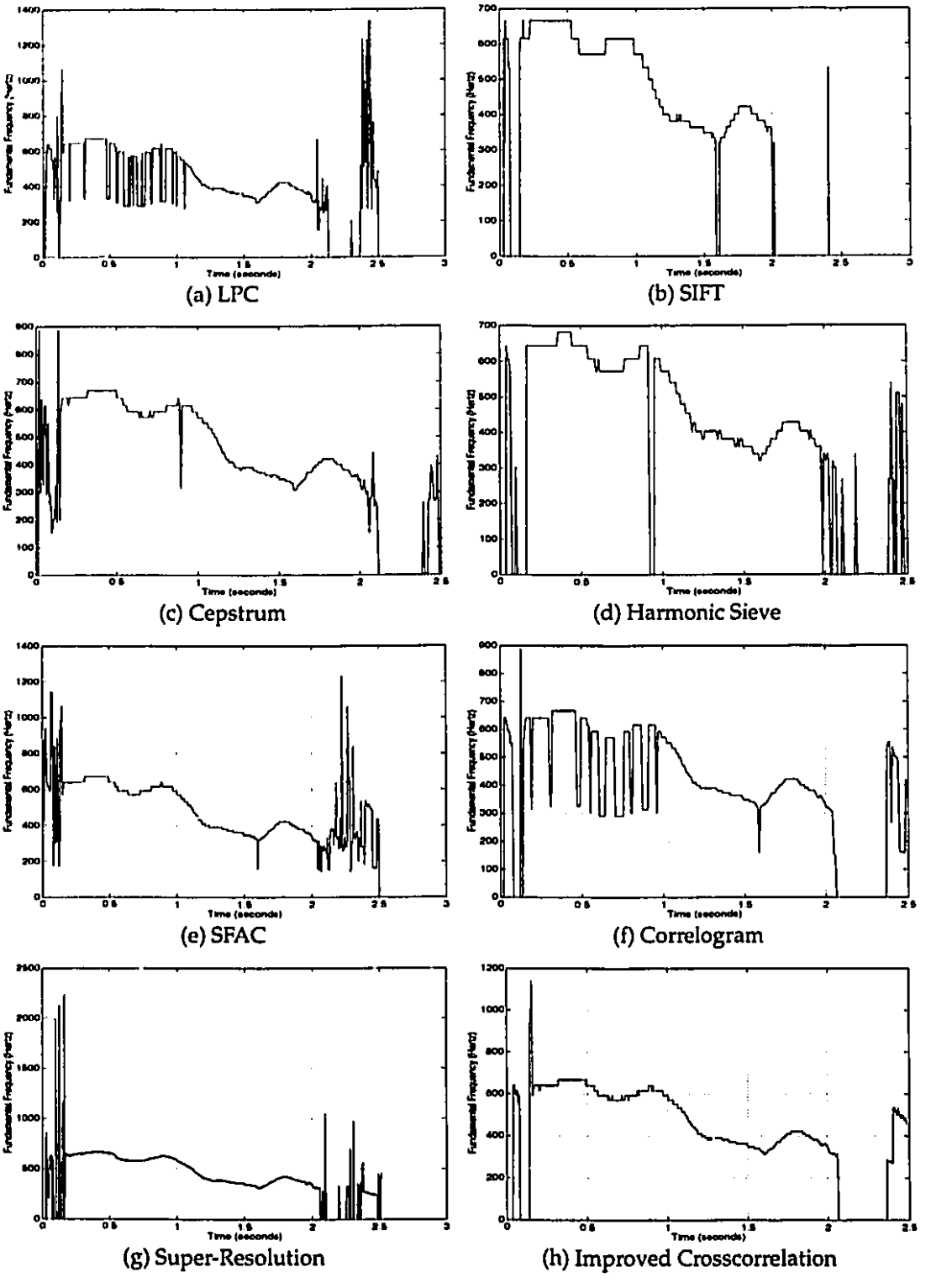


Figure 3.26: Pitch Contours for Recording P09102

2. $p_r(m) = 0$, $p_e(m) \neq 0$, in which case the reference contour has denoted that interval m is unvoiced but the extraction method e has denoted this interval as voiced. This event is tagged as a unvoiced-to-voiced error.
3. $p_r(m) \neq 0, p_e(m) = 0$, in which case the reference contour has labelled interval m as voiced, but the extraction method e has identified the same interval as being unvoiced. This event is tagged as a voiced-to-unvoiced error.
4. $p_r(m) = P_1 \neq 0$, $p_e(m) = P_2 \neq 0$, in which case both the reference and extracted contours label interval m as being voiced, but the values of pitch periods P_1 and P_2 differ. In this event, two types of errors can occur. If the difference between the two extracted pitch periods is small, then a fine pitch error is said to have occurred, otherwise, a gross pitch error has occurred. The former denoted a difference of a few samples whereas the latter typically denotes errors such as pitch halving or doubling.

Defining the error as the difference between the reference and extracted pitch period samples as

$$e(m) = P_1 - P_2 \tag{3.6}$$

then if $|e(m)| \geq 5$ samples, which represents an error of 0.3125 ms in estimating the pitch period, the error is classified as a **gross pitch error**. Given the range of F_0 values in infant cry signals, is a reasonable measure for the cutoff between both fine and gross errors. Consequently, if $|e(m)| \leq 5$ samples a **fine pitch error** is said to have occurred.

Following the above discussion, we can now present the six error measures used to compare the performance results:

1. **Gross Error Count:** For this measurement, the number of gross pitch errors was counted. Also, in order to normalize for the different frame rates and for the different granularity of results offered by different methods, the gross error count for a given method was divided by the number of voiced intervals in order compute the percentage of intervals classified as voiced by both the

reference and the pitch extractor in question, for which gross errors occurred.

2. **Number of Pitch Errors:** Here, the number of intervals N_i in an utterance in which fine pitch errors occur, were counted. As was the case for the gross pitch error count, this value were also divided by the number of intervals classified as voiced by both the reference and the pitch extractor in question, in the recording in order to compute the percentage of voiced intervals in which a fine pitch error occurred.
3. **Mean of the Fine Pitch Errors:** The mean, $\overline{e}$, of the fine pitch errors is defined as

$$\overline{e} = \frac{1}{N_i} \sum_{j=1}^{N_i} e(m_j) \tag{3.7}$$

where m_j is the j^{th} interval in the utterance in which there occurs a fine pitch error and N_i is the number of fine pitch errors occurring in the utterance.

4. **Standard Deviation of Fine Pitch Errors:** This measure is defined as

$$\sigma_e = \sqrt{\frac{1}{N_i} \sum_{j=1}^{N_i} [e(m_j)]^2 - \overline{e}^2}. \tag{3.8}$$

This represents a measure of accuracy in measuring the pitch period during voiced intervals.

5. **Voiced-to-Unvoiced Errors:** This measurement is taken by counting the number of frames where this error occurred and, as well, dividing by the number of voiced intervals in order to compute a percentage value. This measure denotes the accuracy of classifying voiced intervals.
6. **Unvoiced-to-Voiced Errors:** This measurement was taken by counting the number of frames in which this event occurs during unvoiced intervals, and, as well, dividing this value by the number of unvoiced intervals in order to compute a percentage value. This denotes the accuracy of classifying unvoiced intervals.

These measures give a good description of the strengths and weaknesses of the

	Pitch Detector								
File	LPC	SIFT	CEPS	HSIEV	SFAC	CORR	SPR	ICVBM	Sum
A02004	7/505	1/519	0/280	0/259	15/568	7/277	0/848	0/862	30/4118
	1.39%	0.19%	0.00%	0.00%	2.64%	2.53%	0.00%	0.00%	0.73%
A07104	93/252	42/204	18/129	0/86	30/273	123/134	22/1157	0/1191	328/3426
	36.9%	20.6%	13.9%	0.00%	11.0%	91.8%	1.90%	0.00%	9.57%
B056ST	28/171	12/191	13/101	12/54	29/202	30/101	78/364	0/427	202/1611
	16.4%	6.28%	12.9%	22.2%	14.3%	29.7%	21.4%	0.00%	12.5%
C1213SQ3	15/404	6/393	10/202	0/177	3/404	14/196	1/852	0/828	49/3456
	3.71%	1.53%	4.95%	0.00%	0.74%	7.14%	0.12%	0.00%	1.42%
P09102	77/525	4/473	10/256	1/242	19/525	39/259	27/1014	0/1023	177/4317
	14.7%	0.85%	3.91%	0.41%	3.62%	15.1%	2.66%	0.00%	4.10%
Sum	220/1857	65/1780	51/968	13/818	96/1972	213/967	128/4235	0/4331	
	11.8%	3.65%	5.27%	1.59%	4.87%	22.0%	3.02%	0.00%	

Table 3.1: Gross Pitch Errors

various extraction methods, and as well, serve to demonstrate the improvements obtained in using the pitch extraction method proposed in section 3.1. Section 3.5 will elaborate on these results.

These error measures were computed and are presented in tables 3.2 to 3.6. These tables illustrate the improvements achieved with the new crosscorrelation vector-based fundamental frequency extraction method versus the methods borrowed from the speech domain. In all of the tables, the "Sum" column computes the total number of the type of error, indicated in the caption of the respective table, across all the pitch extraction methods tested. On the other hand, the row labeled "Sum" computes the total number of the type of error, indicated in the caption of the respective table, for that particular pitch extraction method across all the test files. The former illustrates if any of the test files are particularly prone to one type of error over another. The latter, however, illustrates which of the extraction methods, if any, generate a higher number or percentage of that type of error when compared to the other extraction methods. Table 3.6 sums across all the error types presented in tables 3.2 to 3.5 for each of the test files and pitch detection methods illustrating which of the F_0 extraction methods yield the best results across all the error measures computed.

	Pitch Detector								
File	LPC	SIFT	CEPS	HSIEV	SFAC	CORR	SPR	ICVBM	Sum
A02004	15/505	17/519	0/280	0/259	19/568	16/277	0/848	0/862	67/4118
	2.97%	3.28%	0.00%	0.00%	3.35%	5.78%	0.00%	0.00%	1.63%
A07104	19/252	4/204	17/129	0/86	8/273	1/134	0/1157	0/1191	49/3426
	7.94%	1.96%	13.2%	0.00%	2.93%	0.75%	0.00%	0.00%	1.43%
B056ST	8/171	11/191	0/101	0/54	5/202	0/101	0/364	0/427	24/1611
	4.68%	5.76%	0.00%	0.00%	2.48%	0.00%	0.00%	0.00%	1.49%
C1213SQ3	28/404	7/393	14/202	1/177	19/404	0/196	0/852	0/828	69/3456
	6.93%	1.78%	6.93%	0.56%	4.70%	0.00%	0.00%	0.00%	2.00%
P09102	18/525	12/473	14/256	0/242	16/525	2/259	0/1014	0/1023	62/4317
	3.43%	2.54%	5.47%	0.00%	3.05%	0.77%	0.00%	0.00%	1.44%
Sum	88/1857	51/1780	45/968	11/818	67/1972	19/967	0/4235	0/4331	
	4.74%	2.87%	4.65%	1.34%	3.40%	1.96%	0.00%	0.00%	

Table 3.2: Fine Pitch Errors

3.5 Discussion of Experimental Results

This section reviews the results of the tests performed in section 3.4. First, the pitch contour plots of the individual test files will be discussed, noting where and why certain methods fail, and how the improved crosscorrelation vector-based fundamental frequency extraction performs comparatively to the other methods tested on the recordings. Following this, the results shown in tables 3.1 to 3.5 will be discussed and the reason behind the failure of certain F_0 extraction methods on recordings with certain characteristics will be addressed. As well, the substantial improvement achieved using the improved crosscorrelation vector-based fundamental frequency extraction method will be illustrated.

3.5.1 Fundamental Frequency Contours

File A02004

Figure 3.22 shows the fundamental frequency contours extracted from file A02004 using the respective methods indicated in the captions of the sub-figures. As was mentioned in section 3.4.1, this file contains a relatively well-behaved contour with slow and smooth variations in its progression, punctuated with some brief

	Pitch Detector								
File	LPC	SIFT	CEPS	HSIEV	SFAC	CORR	SPR	ICVBM	Sum
A02004	1.22	0.90	0.00	0.58	0.95	0.97	0.00	0.00	4.62
A07104	2.09	1.00	1.10	0.00	1.03	0.00	0.00	0.00	5.22
B056ST	0.71	0.00	0.00	0.00	1.09	0.00	0.00	0.00	1.80
C1213SQ3	1.65	0.38	1.46	0.00	2.37	0.00	0.00	0.00	5.86
P09102	1.39	1.03	1.70	0.00	1.39	1.41	0.00	0.00	6.92
Sum	7.06	3.31	4.26	0.58	6.83	2.38	0.00	0.00	

Table 3.3: Standard Deviation of Fine Pitch Errors

episodes of ambient noise occurring during the course of the first episode. Looking at the results of the different contours extracted from the various methods, it can be immediately seen that the cepstrum-based, super-resolution, and the improved crosscorrelation vector-based pitch extraction methods yield the smoothest and most accurate pitch values. Both the cepstrum and the super-resolution method fail to ignore the locally periodic noise burst occurring between time 2 and time 2.5 in the signal.

The contour of the first episode extracted by the super-resolution method has a brief interruption occurring at about time 1.35 seconds, which, in turn, is due to a sudden change in the characteristics of the first formant frequency, which changes the characteristics of the signal. Consequently, adjacent pitch periods in the signal are significantly different and the crosscorrelation value drops below the voiced/unvoiced threshold at that point in the utterance.

All the methods, except for the improved crosscorrelation-vector based fundamental frequency extraction method, have difficulty tracking the pitch at end of the first episode due to the weakness of the signal which either causes the energy contained within a given frame and the autocorrelation peak values to fall below the voicing threshold for the cepstrum in the former case, and for the LPC, SIFT, spectral flattening autocorrelation, and correlogram methods in the latter case. For the super-resolution method, the oscillation between the pitch and no-pitch values occurs due to the weakness of the signal which causes the crosscorrelation values to oscillate above and below the voicing threshold. For the improved crosscorrelation

	Pitch Detector								
File	LPC	SIFT	CEPS	HSIEV	SFAC	CORR	SPR	ICVBM	Sum
A02004	69/574	55/574	7/287	28/287	6/574	10/287	12/860	0/862	187/4305
	12.0%	9.58%	2.44%	9.76%	1.05%	3.48%	1.39%	0.00%	4.34%
A07104	21/273	69/273	7/136	50/136	0/273	2/136	4/1161	0/1191	153/3579
	7.69%	25.1%	5.15%	36.8%	0.00%	1.47%	0.34%	0.00%	4.27%
B056ST	31/202	11/202	0/101	47/101	0/202	0/101	6/370	0/427	95/1706
	15.3%	5.45%	0.00%	46.5%	0.00%	0.00%	1.62%	0.00%	5.57%
C1213SQ3	0/404	11/404	0/202	25/202	0/404	6/202	0/852	0/828	42/3498
	0.00%	2.72%	0.00%	12.4%	0.00%	2.97%	0.00%	0.00%	1.20%
P09102	1/526	53/526	7/263	21/263	1/526	4/263	7/1021	0/1023	94/4168
	0.19%	10.1%	2.66%	8.01%	3.04%	1.52%	0.69%	0.00%	2.25%
Sum	122/1979	199/1979	21/989	171/989	7/1979	22/989	29/4264	0/4331	
	6.61%	10.1%	2.12%	17.2%	0.35%	2.22%	0.68%	0.00%	

Table 3.4: Voiced-to-Unvoiced Errors

vector-based method, the pitch is tracked until voicing stops due the fact that the method uses a lower voiced-to-unvoiced threshold and a distance measure, based on the crosscorrelation maxima values, allowing the method to track a specific pitch contour even after the signal weakens considerably provided that periodicity is maintained still present in the portions of the signal.

The LPC, SIFT, and harmonic sieve methods all degrade under noisy conditions, as can be seen in the interruptions occurring in the contour of the first voiced episode. The SIFT method performs better than the LPC method, due to the reduced bandwidth of the input signal used in SIFT, which allows it to remove some of the effects of the noise bursts from the input spectrum. The harmonic sieve is also adversely affected by the presence of noise bursts, and by the presence of frames in the signal with weak or low amplitudes, when the number of peaks falling through the sieve is sharply reduced, causing those portions to be flagged as unvoiced.

File A07104

A cry utterance which is typical of the high fundamental frequency values found in these signals, and which also provides one example of the rapid F_0 variations sometimes found in these signals, is contained in file A07104. The pitch contours

	Pitch Detector								
File	LPC	SIFT	CEPS	HSIEV	SFAC	CORR	SPR	ICVBM	Sum
A02004	7/170	0/170	12/84	6/84	50/170	0/84	15/482	0/458	90/1702
	4.12%	0.00%	14.3%	7.14%	29.4%	0.00%	3.11%	0.00%	5.29%
A07104	24/39	0/39	3/20	2/20	38/39	0/20	19/144	0/137	86/341
	61.5%	0.00%	15.0%	10.0%	97.4%	0.00%	13.2%	0.00%	25.2%
B056ST	10/78	4/78	10/38	1/38	21/78	3/38	23/145	0/110	72/603
	12.8%	5.13%	26.3%	2.63%	26.9%	7.89%	15.9%	0.00%	11.9%
C1213SQ3	8/204	2/204	4/101	8/101	22/204	6/101	55/315	0/235	101/1465
	3.92%	0.98%	3.96%	7.92%	10.8%	5.94%	17.5%	0.00%	6.89%
P09102	35/98	2/98	18/46	5/46	98/98	0/46	27/180	0/135	185/747
	35.7%	2.05%	39.1%	10.9%	100%	0.00%	15.0%	0.00%	24.8%
Sum	84/589	8/589	47/289	22/289	229/589	9/289	139/1266	0/1075	
	14.3%	1.36%	16.3%	7.61%	38.9%	3.11%	11.0%	0.00%	

Table 3.5: Unvoiced-to-Voiced Errors

extracted from this recording are shown in figure 3.23. As well, the second episode in this recording features a very narrow bandwidth first formant which occurs at a frequency equal to twice the value of the fundamental frequency, at least for the first 0.6 seconds of this episode.

Of the methods applied to this signal, the improved crosscorrelation vector-based method gives the best results, providing smooth pitch contours for both episodes in the utterance. In the first contour, the the rapid change in F_0 from about 1000 Hz to 800 Hz is successfully tracked as is the subsequent rapid rise in F_0 from about 800 Hz to 2000 Hz. Also, the evolution of the second contour is successfully tracked even though the occurrence of the narrow bandwidth F_1 seems to affect some of the other methods quite adversely.

The super-resolution method behaves reasonably well. Interruptions occur in the first contour, mainly due to a change in formant values the middle portion of that episode, and at the end of the episode because of a weakening signal. Interruptions in the second episode occur because of some noise occurring at the beginning and shortly after the 1 second mark, and as the signal weakens at the end of the episode.

All of the other methods yield extremely poor results due to inconsistencies in the pitch values extracted, the most common error being pitch halving errors.

	Pitch Detector								
File	LPC	SIFT	CEPS	HSIEV	SFAC	CORR	SPR	ICVBM	Sum
A02004	98/744	74/744	19/371	34/371	90/744	33/371	27/1342	0/1320	374/6007
	13.2%	9.81%	5.12%	9.16%	12.1%	8.89%	2.01%	0.00%	6.23%
A07104	157/312	115/312	45/156	52/156	76/312	126/155	45/1305	0/1328	616/4037
	50.3%	36.9%	28.8%	33.3%	24.3%	81.3%	3.45%	0.00%	15.3%
B056ST	77/280	44/280	23/139	60/139	55/280	33/139	107/515	0/537	399/2309
	27.5%	15.7%	16.5%	43.2%	19.6%	23.7%	20.8%	0.00%	17.2%
C1213SQ3	51/608	26/608	28/303	34/303	44/608	26/303	56/1167	0/1063	265/4963
	8.39%	4.28%	9.24%	11.2%	7.24%	8.58%	4.80%	0.00%	5.34%
P09102	131/624	71/624	49/309	27/309	134/624	45/309	61/1201	0/1158	518/5158
	21.0%	11.4%	15.9%	8.74%	21.5%	14.6%	5.08%	0.00%	10.0%
Sum	514/2658	329/2658	164/1278	207/1278	399/2658	263/1278	296/5530	0/5406	
	19.3%	12.4%	12.8%	16.2%	15.0%	20.6%	5.35%	0.00%	

Table 3.6: Total Errors

Starting with figure 3.23(a), it can be seen that the LPC method produces a number of errors, but the most common of these errors are indeed pitch halving errors. As was mentioned in section 3.2.1, this occurs because the spectrum of an utterance containing a very high F_0 will contain very few harmonics. In these cases, the the poles of the LPC spectrum will model the harmonic peaks and not the spectral peaks. Consequently, the effects of F_1 and F_0 will be removed from the input signal when it is inverse filtered, leaving only weak periodicity present at a period equal to twice that of the pitch period, resulting in pitch halving errors. In the first episode of the signal, following the drop in F_0 from 1000 Hz to 800 Hz occurring at about time 0.1 seconds, there are so few harmonics present in the input signal frame that the poles model the harmonics so well that the inverse filtered signal shows no periodicity at all, and these frames are tagged as being unvoiced. In short, this file illustrates the limitations of LPC on this type of cry utterance.

The SIFT method yields better results, as it uses both less poles and a version of the signal which has been decimated by 2, reducing the bandwidth of the spectrum that needs to be modelled by 2 as well. Here, the number of pitch doubling errors is considerable. This occurs because the F_0 peak in the spectrum is stronger than that of its harmonics. Inverse filtering the signal frame removes this effect, but retains the effect of the harmonics in the inverse filtered spectrum. Consequently, this appears as peaks occurring at twice the fundamental frequency value of the

input signal frame. Also, SIFT has problems successfully tracking the F_0 contour at the end of the second episode due to the poor harmonic content in the signal. In these frames, the LPC coefficients, or poles, model the fundamental and its first harmonic in the input spectrum, removing all traces of periodicity from the input signal frame in the inverse filtered residual.

Next, the cepstrum pitch extraction method performs reasonably well, save for a number of pitch halving errors. This is mainly caused by variations in the amplitude of the input signal, or shimmer, and due to some brief bursts of additive noise, which may cause every other period to look more like the fundamental than the true pitch period does, due to these effects.

The harmonic sieve method misses approximately the first half of the first pitch contour and the final portion of the second pitch contour due to poor harmonic content, which tends to flag a particular interval as being unvoiced.

The spectral flattening autocorrelation method also produces a number of pitch halving errors which are caused by amplitude shimmer in a given frame of the input signal which causes every other period in the input frame to look more like the pitch period than the actual pitch period. In addition, the frames between the two voiced episodes in the signal are incorrectly labeled as being voiced, since the energy in these frames exceeds the voiced-unvoiced threshold.

Cochlear filtering and the subsequent correlation of these channels leads to inconsistent results for the correlogram-based pitch extractor. In all but a few frames the extracted pitch value is incorrect and the pitch extracted is either $\frac{F_0}{2}$ or $\frac{F_0}{3}$, making this method practically useless for this vocalization.

File B056ST

The pitch contours extracted by the various methods on file B056ST are displayed in figure 3.24. As stated in section 3.4.1, this file corresponds to the cry of a prema-

ture infant following a heel stick and contains portions of sustained phonations, dysphonation, double harmonic break episodes, and an inspiratory phonation located at the end of the utterance. Also, there are places at the beginning and at the end of the first vocalization where a narrow bandwidth F_1 occurs at a frequency equal to twice that of the fundamental frequency.

For this recording, the improved crosscorrelation vector-based method yields perfect results for both episodes. It successfully tracks the start of the double harmonic break episode beginning at time 0.4 seconds, and lasting until about 0.65 seconds, which is punctuated by a dysphonic episode shortly before time 0.5, and a brief return to the true F_0 in the neighbourhood of the 0.6 second mark. In addition, the inspiratory phonation is successfully tracked from start to finish, despite the rapid drop in the fundamental frequency values over the course of the inspiration.

The next best result is achieved by the SIFT method which successfully handles the double harmonic break episode, save for the brief dysphonic episode which it classified as voiced. This error is due to the occurrence of some periodic portions contained within a frame that also contains some dysphonic signal portions, with the periodic portions causing the frame to be labeled as periodic. In one frame at the end of the first episode, this method incorrectly classifies the pitch as being the formant frequency, due to the occurrence of the narrow bandwidth F_1 which occurs at a frequency equal to $2F_0$. Due to the poor harmonic content in the inspiratory phonation, the final portion of this utterance is not tracked. For the initial portion of this utterance, a number of pitch halving errors occur.

The cepstrum-based F_0 extraction method also does reasonably well, but in the first episode it misses the dysphonic portion due to the same effect as was mentioned for the SIFT method. As well, tracking the return to the original F_0 during the course of the double harmonic break episode occurs for only one frame, which is too short. At the beginning of the first episode, the method also labels the F_1 frequency as being F_0, since the sharpest peak in the cepstrum occurs for

F_1 instead of F_0, and during a brief portion of noise shortly after, produces a pitch halving error. The inspiratory phonation is correctly tracked at first, but soon falls victim to pitch halving errors due to the shimmer present in the pitch periods in this portion of the signal.

As far as the other methods are concerned, they all do rather poorly for a number of reasons. The LPC method over-models the input spectrum causing the periodicity due to the fundamental frequency to be removed when the signal is inverse filtered, causing a number of pitch halving errors, as can be clearly seen in figure 3.24(a). As well, there are a number of voiced-to-unvoiced errors in the course of the first episode as well. The initial portion of the inspiratory phonation is tracked but soon terminates due to the erratic nature of this portion of the signal.

The results from the harmonic sieve are extremely poor as Goldstein's theory of pitch perception causes a number of harmonic peaks to be "masked" or eliminated prior to sieving. For a number of frames, both prior to, and during the double harmonic break episode in the first utterance, the strength of the first formant peak masks the effect of F_0 and the second harmonic, so that the sieve determined the F_0 to be F_1. In other portions of the first episode, where the voice-to-unvoiced errors occur, there are so few harmonic peaks remaining in the signal spectrum subsequent to the Goldstein pitch perception stage that the sieve cannot make a definite conclusion regarding the true value of F_0.

The spectral flattening autocorrelation method has a number of pitch doubling errors arising from the clipping of the signal which includes formant peaks in the clipped signal. This is due to the narrow bandwidth of F_1 for some portions of this signal, and due to the small decay in amplitude between the pitch epoch and subsequent intermediate oscillation peak due to the strong and narrow bandwidth formant, which occurs at a frequency of $2F_0$, resulting in the pitch period of the clipped signal to be twice the true F_0.

The correlogram based method once again leads to a number pitch halving

errors, presumably because the periodicity across all the cochlear bands at half the true F_0 is stronger than it is for the true F_0, which is caused by some noise or amplitude shimmer in these portions of the signal.

Lastly, the super-resolution pitch extraction method correctly tracks the beginning of the double harmonic break episode, since the change in the signal causes the crosscorrelation between adjacent segments to fall below the voicing threshold, causing the method to stop tracking F_0. The drawback of this method is that once it begins tracking a certain F_0, it does not stop unless the crosscorrelation peak for the pitch period lag falls below this voicing threshold. This is precisely what happens for this particular recording during the double harmonic break episode. Also, the beginning of the first episode displays another drawback of this method. If the method has not settled into tracking a specific F_0 value, the method is subject to incorrectly classifying F_0 if there are perturbations in the signal which cause periodicity at other values to be briefly stronger than those at the true pitch period lag.

File C1213SQ3

As was described in section 3.4.1, this recording, whose results are shown figure 3.25, is an example of a cry that has two very erratic and quickly varying F_0 contours, and also contains the tail end of one contour, with periods of ambient noise episodes occurring between these episodes.

Once again, the improved crosscorrelation vector-based method gives perfect results, successfully tracking the progression of the rapidly changing F_0 values without interruption for the two episodes. None of the other methods tested extract smooth, uninterrupted F_0 contours, so the individual contours will be reviewed individually.

The LPC generates more interruptions in the contours than its reduced band-

width counterpart, the SIFT method. For LPC, the occurrence of rapidly varying pitch periods located within a given frame result in pitch halving errors, especially if some portions of the frame are corrupted by noise. This effect also plagues the cepstral and spectral flattening autocorrelation methods as well. The SIFT method, however, appears to be more robust to these effects, which appear only when lower frequency noise is present in the signal.

Next, the harmonic sieve produces numerous voiced-to-unvoiced errors, as a result of the effects that quickly varying periods contained within a given window have on the resulting spectrum. In these cases, the harmonic peaks in the spectrum do not occur at exact multiples of the fundamental frequency peak. When a sieve with an initial frequency in the neighbourhood F_0 is used, not all the peaks in the spectrum will fall through the sieve, and the segment will be classified as unvoiced.

As was the case for the three test files described previously, the correlogram based method performs poorly for this recording as well. Once again the periodicity across all the cochlear bands at half the true F_0 appears to be stronger than that for the true F_0, which is most likely caused by some noise or amplitude shimmer in these particular portions of the signal.

The super-resolution pitch extraction method performs reasonably well on the contours, save for a few erratic portions at the start of the second contour before the algorithm settles and begins to track the pitch. The major disadvantage of this method is in how it handles the non-voiced, noisy portions of the signal. From the contour shown in figure 3.25(g), it can be immediately observed that this method tracks the locally periodic portions of noise bursts, leading to a number of unvoiced-to-voiced errors.

File P09102

The extracted pitch contours for recording P09102, the last of the 5 test files, is shown in figure 3.26. This recording features a long voiced episode with a slowly decreasing F_0 with little variations, preceded and followed by short voiced episodes, with silence separating the three contours.

Once again, the improved crosscorrelation vector-based pitch extraction method, gives the best results. It tracks all three contours perfectly, including the high F_0 portion at the beginning of the second voiced episode, and the double harmonic break episode at the beginning of the third contour.

The LPC pitch extractor produces to a number of pitch halving errors, especially in the second utterance when the signal amplitude weakens somewhat, causing less harmonics to be present in the signal spectrum. This causes the poles to model the harmonic peaks precisely, leaving little apparent periodicity at the true pitch period. In the end portion of the second utterance, the pitch values extracted become erratic, as they also do for the third utterance, due both to the weakness of the input signal and to the presence of some additive noise in the signal.

The SIFT method provides a smoother contour than LPC does, but misses the initial high F_0 burst at the start of the second contour. As well, this pitch extraction method gets interrupted at about the 1.6 second mark when the signal weakens as it reaches the lowest F_0 value for this utterance. Note that the third episode consists of only one peak; all other pitch contours are ignored as the signal is weak, and noise is present in the signal, so that the autocorrelation of the inverse filtered residual does not indicate periodicity.

The cepstrum-based and spectral flattening autocorrelation pitch extractors both provide smooth contours, save for a few pitch halving errors occurring in some places. Note, however, that these methods successfully extract one correct pitch period value from the initial high F_0 portion of the second utterance, but fail to

classify the portion of the recording between the first and second episodes as unvoiced. Both the contours for the first and third episodes are inconsistent due to weak signal values.

The harmonic sieve performs well in the second utterance until the signal weakens, at which point the low number of harmonics present in the spectrum to be sieved causes the segment to be flagged as unvoiced. The break in this contour prior to the one second mark is due to the Goldstein theory of hearing pre-processor eliminating some harmonic peaks due to masking effects by other neighbouring, higher amplitude peaks so that the remaining components, when sieved, yield inconclusive results, causing this series of input frames to be labelled as unvoiced. The first and last F_0 contours are also erratic due to weak signal values.

As is the case for the other test files presented above, the correlogram generates a number of gross pitch errors for both the second and third utterances. This is due to the stronger periodicity present across the cochlear filtered bands at twice the pitch period than at the actual pitch period. This method does, however, identify one frame of the initial high F_0 portion of the second utterance, even if it is subsequently followed by voiced-to-unvoiced errors.

The F_0 contours extracted from the super-resolution method, shown in figure 3.26(g), once again display erratic behaviour during the first few time indexes of the utterance before the pitch tracking about a certain pitch value begins, localizing the search for pitch candidates in subsequent frames. Due to the short and erratic nature of the first utterance and the high F_0 portion at the start of the second, the extracted pitch values at these points varies enormously. This occurs as the method is thrown off by locally strong periodicities present in the signal due to noise or narrow bandwidth F_1 values. Some spurious and locally periodic noise bursts are picked up between the second and third utterances. Note once again that for the third utterance, the method begins tracking the low F_0 of the double harmonic break episode, and does not follow the change to the higher F_0 value

when it changes a few time instants later.

3.5.2 Error Analysis

This subsection reviews and discusses the results presented in tables 3.1 to 3.6, and comments on which of the methods tested produces the best and the worst results, and explains why these methods fail where they do. First, table 3.1, which presents the gross pitch errors as defined in section 3.4.3, will be discussed. Briefly, the gross pitch error represent errors in the extracted pitch contour which differs from the reference pitch contour by more than 5 samples and usually flags errors such as pitch halving or doubling errors.

Gross Pitch Errors

From table 3.1, it can immediately be noticed that the method that gives the most pitch errors across all the test files (the sum column) is the correlogram-based pitch extraction method. This is not surprising based on the discussion of the extracted pitch contours from the test files presented in the previous section. Although this method has an intuitive appeal in that it performs cochlear band filtering in an attempt to mimic the nerve firing patters of the hair cells contained in the human ear, it is this complexity, however, that may actually lead to this method's downfall. The peaks in the correlogram subsequent to the cochlear filtering result in high values at multiples of the pitch period and lower values at the true pitch period. Consequently, although this method may seem appealing at first, the actual results reveal that this method is not very accurate in extracting F_0 from infant cries. The largest number of gross pitch errors occurs for file A07104, which implies that this method has particular difficulty extracting pitch from utterances containing high F_0 values.

The method that has the next highest rate of gross pitch errors is the linear

predictive coding (LPC) method. This is due to the fact that in cases where there are few harmonics present in the spectrum of an input signal frame, the number of poles used to model the spectrum do not, decreases, in turn, causing the poles to model the harmonic peaks. Consequently, when the signal is inverse filtered, all the effects of the harmonics are removed from the signal, leaving only periodicity present at multiples of the true pitch period leading to these types of errors. Note that reducing the bandwidth of the input signal, decimating, and reducing the number of poles used to model the spectrum, as is done in the simplified inverse filter tracking method (SIFT), reduces these errors substantially.

The improved crosscorrelation vector-based method provides the best results, producing no gross pitch errors in any of the files tested. This result is typical of those achieved with other files in the data set, as well, and whose results are not included here. Since this method uses post-processing in the form of thresholding, distance calculations, and distance analysis in order to determine the true pitch period, gross pitch errors seldom occur. Consequently, the improvement in this error measure when compared to some of the classical methods, and some of the newer, more complex methods, is immediately apparent, both in the tabular results, and when comparing the accuracy of the extracted pitch contours.

The super-resolution pitch extraction method, which also uses the crosscorrelation to generate pitch candidates, produces a number of gross pitch errors, particularly in a couple of cases. First, if while tracking a specific F_0, during a double harmonic break episode, for example, the fundamental frequency jumps to a value which is twice that of the old fundamental frequency, the method will continue to track the old F_0 value as being the true value of F_0, since periodicity exceeding the voiced-to-unvoiced threshold will still exist at this value. Also, in sections of the signal before the algorithm settles and begins to track a specific F_0 value, when the signal is either weak, or there is a narrow bandwidth F_1, other strong periodic peaks can be incorrectly labeled as being the fundamental frequency. Consequently, although the super-resolution method and the improved

crosscorrelation vector based method both use the crosscorrelation as a means for generating pitch candidates, the former uses the information located at that specific time index when determining the most likely F_0 value. The latter method uses a distance measure over the duration of the contour, so that the contour with the highest score will be tagged as corresponding to the pitch contour. As a result, the improved crosscorrelation vector-based pitch extractor is much less sensitive to noise, signal strength, and locally strong oscillations within pitch epochs, than the super-resolution method is.

The harmonic sieve also performs well in terms of the small number of gross pitch errors that it generates. This small gross pitch error rate is somewhat deceiving, however, because this method suffers from a large voiced-to-unvoiced error rate, as can be seen in table 3.4. Despite this reduction in the number of voiced frames available to this method from which it can determine the pitch, it still makes only a few pitch errors. Consequently, if the voiced-unvoiced determination method could be improved for the harmonic sieve pitch extractor, this method could possibly achieve reasonable results.

Fine Pitch Errors

To briefly re-state what constitutes a fine pitch error before discussing the results, a fine pitch error is a difference between the extracted pitch value and the reference pitch value of less than 5 samples. These types of errors occur when the method used avoids making gross pitch errors for these frames, but fails to extract the precise pitch value from the input frame due to imprecisions in the extraction method. The number and rate of fine pitch errors for the five test files are listed in table 3.2. In addition, the standard deviation, in samples, of the fine pitch error is listed in table 3.3. This value measures the accuracy of the pitch values extracted during voiced intervals.

Insofar as this type of error is concerned, there are no fine pitch errors in both

the improved crosscorrelation vector-based method and in the super-resolution method for the five test files. This is not surprising as both these methods use the normalized crosscorrelation as the core vehicle for determining pitch candidates. The normalized crosscorrelation provides an extremely accurate way of computing the vocal fundamental frequency since it computes the pitch based on the the location of maxima in the crosscorrelation vector which correspond to possible pitch candidates. This results in the absence of fine pitch errors and consequently no variation between the extracted and the reference pitch values for both of these methods if gross pitch errors do not occur.

Of the other pitch extraction methods tested, the one which generates the largest fine pitch error rate and the one which also has the highest variation between the extracted and reference pitch values, as indicated by the standard deviation of these errors, is the linear predictive coding method. This is due to a combination of effects. First, all frame-based methods have incorporated in them the averaging of the pitch periods contained within a given frame of samples. Some files, such as A07104 and C1213SQ3, contain segments where the pitch undergoes fast changes in value. This causes the harmonic peaks to be wider than they would be for a window containing a steady F_0 value. When the LPC poles attempt to model this spectrum, they may not be able to accurately model the wider harmonic peaks. When the input signal is inverse filtered, the periodicity present in the residual will be an average of the pitch values contained within the signal window further distorted by the spectral modelling process.

Note that most of the methods which perform some form of spectral transformation, such as SIFT, cepstrum, and spectral flattening autocorrelation methods, suffer from a higher pitch period error rate than the other methods that do not perform a spectral transformation, such as the harmonic sieve and the correlogram pitch extractor. The same observation can be made for the standard deviation of fine pitch errors.

Voiced-to-Unvoiced and Unvoiced-to-Voiced Errors

The tabulated results for the voiced-to-unvoiced errors and the unvoiced-to-voiced errors can be found in tables 3.4 and 3.5 respectively. These tables are closely related in that if a certain frame is not classified as voiced, it will be classified as unvoiced. Individually, however, these results can illustrate if and which of the methods tested are more biased towards one type of error over another.

Looking at the results for the improved crosscorrelation vector-based pitch extraction method, no voiced-to-unvoiced or unvoiced-to-voiced errors were found in the test files. This is because the post-processing phase of the method removes from the set of candidate pitch values, all contours, or sections, which last less than 8 time intervals, which effectively excludes locally periodic noise bursts, or simply spurious peaks, from being considered as pitch candidates. So, the post-processing phase does indeed do a good job at avoiding these types of errors as well.

For the other methods, the harmonic sieve gives the highest voice-to-unvoiced error rate. A particular frame will be classified as unvoiced if the harmonic content is poor in the input spectrum, or if there are a number of harmonic peaks that are removed from the input spectrum by the Goldstein theory of hearing processing stage prior to sieving. In these cases, if the remaining harmonic peaks are sparse, or if there are a small number of harmonics peaks present in the spectrum prior to sieving, the segment will be flagged as unvoiced. This is the major drawback of this method. Note, however, that there are fewer unvoiced-to-voiced errors.

The SIFT method also suffers from a relatively high number of voiced-to-unvoiced errors when compared to the number of unvoiced-to-voiced errors. If there are frames which contain high F_0 values and thus only one or two harmonics are present in the input spectrum, the poles will model the harmonic peaks exactly leaving no trace of periodicity when the signal is inverse filtered, causing the segment to be incorrectly labeled as unvoiced. Although, the SIFT method has low gross pitch errors because of the smaller signal bandwidth used, it suffers from a

considerable number of voiced-to-unvoiced errors. The converse can be said for LPC, which uses the full bandwidth signal.

The cepstrum uses the energy contained within a given signal frame to determine whether a signal is voiced or unvoiced. This method works reasonably well for voiced-to-unvoiced errors. For unvoiced-to-voiced errors, however, the error rate is larger as the method tends to pick up some locally periodic bursts of noise in a number of frames in the test files and incorrectly classifies these segments as being voiced.

The spectral flattening autocorrelation method uses the same voiced/unvoiced classification method that the cepstrum does, with the difference that the threshold for the energy contained within a given segment for this method is lower than that used by the cepstrum-based method, as was explained in section 3.4.2. This results in a smaller number of voiced-to-unvoiced errors, but a significant increase in the number of unvoiced-to-voiced errors.

Lastly, the super-resolution method shows its bias towards unvoiced-to-voiced errors which occur when any two adjacent signal segments have a crosscorrelation value greater than a threshold value, will cause that lag at that specific time index to be flagged as being voiced. This causes unvoiced portions of the signal which may be corrupted by noise, to be occasionally labelled as being voiced due to the high crosscorrelation values which occasionally occur in these cases.

Total Errors

The table cataloging the sum of the errors identified above is listed in table 3.6. This table sums the gross pitch errors, fine pitch errors, the voiced-to-unvoiced errors, and the unvoiced-to-voiced errors, dividing this total by the number of frames or windows in the utterance.

The best results were achieved for the improved crosscorrelation vector-based

pitch extraction method which generated no error of any kind on the test utterances. This is not to say that the performance of this method is perfect, but it gave excellent results for the test files. On some other files in the data set which were either corrupted by strong episodes of noise or ambient sounds, such as tones, these episodes, particularly if they were of long duration and relatively periodic, would be tracked as pitch contours during moments when there was no voicing in the signal. However, these files were not included in the test set since most of the methods used in the comparison would have made the same mistake. In short, this method outperforms all other methods tested, does a good job of removing the various types of errors, and yields optimal results, with the output being suitable for further processing, if so required. As well, it should be noted that on all the files of the data sets, as described in section 3.3, on which the various pitch extraction methods were tested, the improved crosscorrelation vector-based pitch extraction method outperformed all of the other methods in all of the error classes.

The method which yields the next best total error rate is the super-resolution method, mainly because of the large number of frames that are generated by using this other crosscorrelation based method. The main problem of this method occurs when the F_0 value changes at the end of a double harmonic break episode, and this change is not followed, as was mentioned above.

At the other end of the scale is the correlogram-based pitch extraction method which has the highest total error rate with the majority of errors occurring because of gross pitch errors. All the other methods have errors rates that are within a few percentage points of each other.

The file that lead to the most errors was file B056ST, where most of the methods had trouble dealing with the double harmonic break episode and with the inspiratory phonation. The file with the next highest error rate was file A07104 which generated problems for a number of methods due to the high fundamental frequency values contained in this recording.

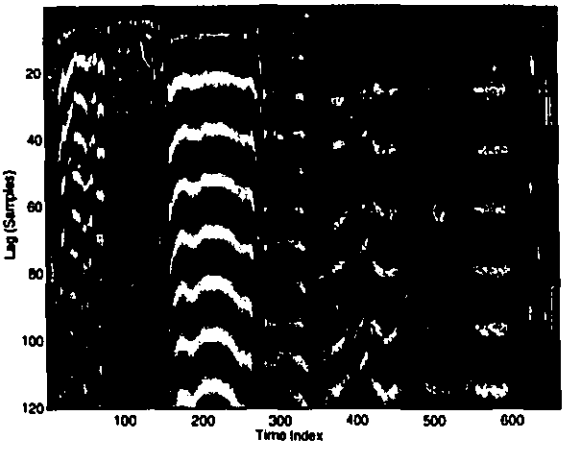


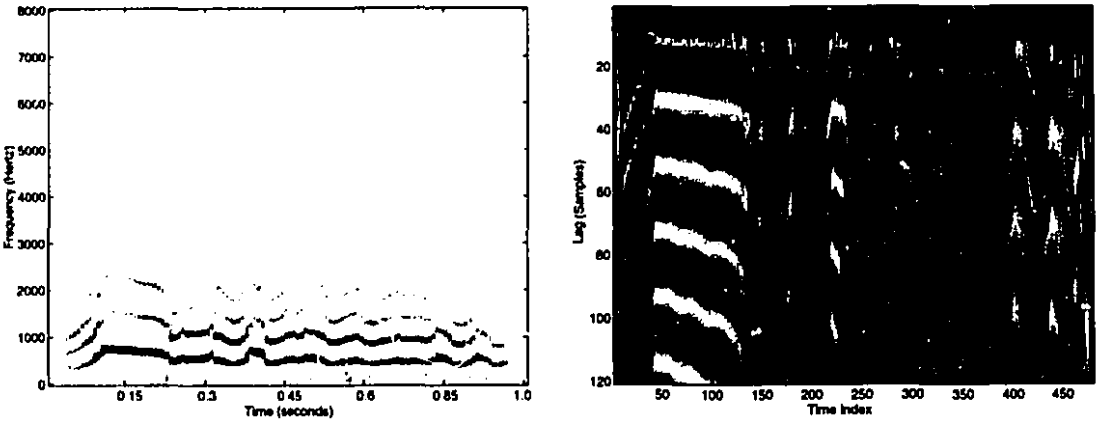
Figure 3.27: Crosscorrelogram of a Cry Uttered after a Heel Stick

3.6 Other Extensions of the Improved Crosscorrelation Vector-Based Fundamental Frequency Method

The following subsections present other advantages or spin-offs from the improved pitch period extraction method presented in section 3.1.

3.6.1 Improved Utterance Visualization Using the Crosscorrelogram

Aside from being a good method for accurately tracking the pitch period in infant cry utterances, the signal transformation phase of the crosscorrelation vector-based fundamental frequency extraction method, described in section 3.1.2, also provides information which is useful for improved visualization of infant cries [Petroni *et al.*, 1994b]. The sequence of crosscorrelation vectors placed together in a matrix as described in section 3.1.3, can be displayed in a three-dimensional plot of lag versus time versus intensity called a **crosscorrelogram**, which is shown in figure 3.27 [De Mori and Omologo, 1993]. This plot differs, however, from the one presented by De Mori and Omologo in that the time increments in the crosscorrelogram displayed in figure 3.27 are dependent on the lag value of the most likely pitch candidate for a given time index, not on a fixed value.



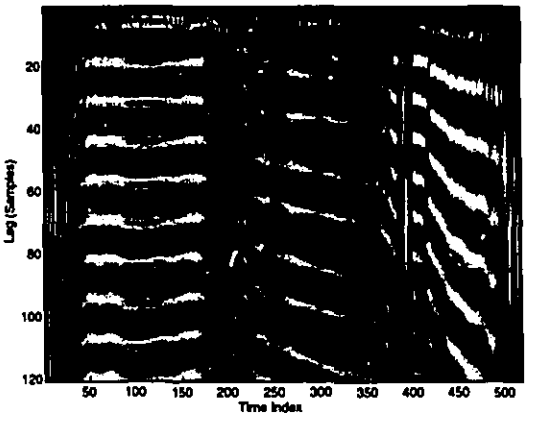
(a) Spectrogram of Cry Utterance

(b) Crosscorrelogram of Cry Utterance

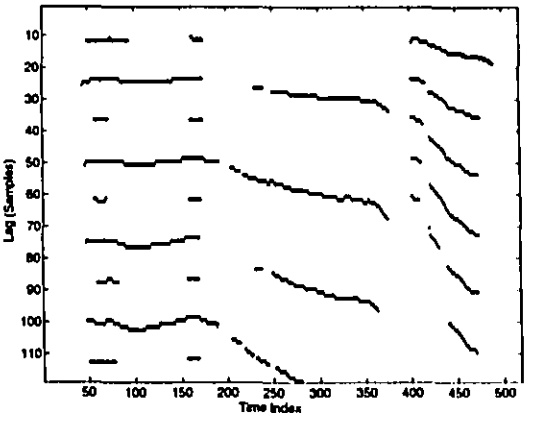
Figure 3.28: Comparison Between Spectrogram and Correlogram for the Second Cry Utterance of File C1213SQ3

The presence of periodicity in the crosscorrelogram is indicated by the occurrence of peaks (large positive crosscorrelation values) and valleys (large negative crosscorrelation values) denoted by the black and white shades respectively in the plot. Gray areas denote the occurrence of non-periodicity corresponding to either noise, silence, or dysphonic sections in the utterance. As is the case for the range of expected pitch period lag values in the utterance, the lag values indicated on the y-axis span values from 5 to 120, which correspond to frequencies of 3200 Hz to 133 Hz respectively, given a 16 kHz sampling rate.

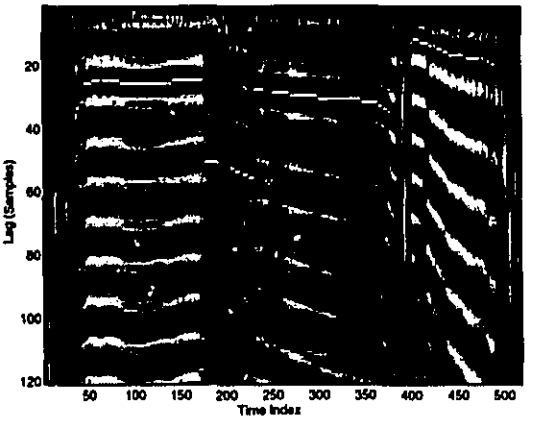
The crosscorrelogram gives a finer-grain view of the progression of the pitch period than the standard method of utterance vocalization, namely the spectrogram [Oppenheim, 1970], does. Comparing the spectrogram and crosscorrelogram for the same cry recording, as shown in figure 3.28, it is clear that the correlogram of figure 3.28(b) gives more detail than the spectrogram of figure 3.28(a) does. Intuitively this can be understood from the fact that the crosscorrelogram generates a vector for almost every pitch period, whereas the spectrogram uses a window of samples which inherently averages the F_0 values of the signal contained within a frame.



(a) Crosscorrelogram



(b) Thresholded Crosscorrelogram Matrix



(c) Extracted Pitch Period Contour Superimposed on Crosscorrelogram

Figure 3.29: Pitch Period Extraction Process from a Cry Recording Uttered After a Heel Stick (File B056ST)

In the crosscorrelogram, the identification of the pitch peak can usually be done visually, with the first strong peak from the top of the crosscorrelogram corresponding to the F_0 lag, although in the event of narrow bandwidth F_1, strong maxima appearing at smaller lag values may appear. Consequently, this inspection heuristic should be used with caution.

The results of the crosscorrelation matrix processing described in section 3.1.4 can be shown in the same manner and the resulting pitch period values can be superimposed on the the crosscorrelogram plot as well [Petroni *et al.*, 1994a]. The

tableau in figure 3.29 displays the crosscorrelogram in figure 3.29(a), the thresholded crosscorrelogram matrix in figure 3.29(b), and the actual extracted pitch period contour superimposed on the crosscorrelogram plot in figure 3.29(c), of an utterance which features a double harmonic break episode.

Note that the pitch contour is properly tracked through the double harmonic break episode. The crosscorrelogram contrasts to the series of plots, called a movie, which would be required by the correlogram-based algorithm described in section 3.2.5 since the correlogram generates a 3-dimensional plot for a single input frame and *not* for the entire utterance.

This chapter presented the description, implementation, and a sample of experimental results of the improved crosscorrelation vector-based fundamental frequency extraction method: a method which successfully and accurately tracks pitch contours in infant cries. The results of this method were compared against six other methods commonly used in the speech domain on five test cry utterances each with different spectral and fundamental frequency characteristics. This chapter has addressed the lack of an adequate method to extract the fundamental frequency from infant cry signals. Although, the method was designed to handle the large range of fundamental frequency values present in infant cry signal, this method can useful for other speech signals as well.

Chapter 4 Classification of Infant Cries Using Artificial Neural Networks

This chapter presents research undertaken for the purpose of classifying and distinguishing between different cry types using artificial neural networks (ANNs). The chapter begins with an introduction regarding the advantages of computing and classifying with neural networks, with the choice of this methodology over other classification methods being justified. This will be followed by a presentation of the paradigms used in cry classification experiments and their suitability for use with time varying signals will also be outlined, as will their relative strengths and weaknesses. The input features derived from the cry recordings and used as inputs to the networks will then be presented, followed by a discussion of their relative strengths and weaknesses. These features use information derived from the entire spectrum of the cry signal, since they provide a more comprehensive representation of both the fundamental frequency and of the vocal tract. Next, a brief overview of the software used to simulate these networks will be presented before proceeding to the presentation and discussion of the artificial neural network test results for the different input features.

It should be noted that the work presented in this chapter represents a novel application of artificial neural networks in a domain where they have not been used prior to this dissertation, or if such was the case, this fact has not been mentioned in the literature. Through the comparison of results achieved using different architectures and input features, certain conclusions can be drawn setting the stage for future research in this area.

4.1 Classification with Artificial Neural Networks: Introduction and Motivation

Artificial neural networks, or "neural nets" as they are commonly referred to, have and are currently being used to solve a number of different classification and computational issues in a variety of different domains [Lippmann, 1987]. This methodology is especially attractive due to its inherent parallelism, the simple computations involved during its operation, and resulting from recent advancements in very large scale integrated (VLSI) circuits which has allowed software simulations of neural networks to be implemented in fast hardware [Dayhoff, 1990, Morgan and Scofield, 1991]. These advancements have allowed the development of real-time ANN applications in the area of speech processing and recognition, but their application in the domain of infant crying, still remains undocumented in the literature [Petroni *et al.*, 1995].

Artificial neural networks are based on the present understanding of the way that biological neural systems behave, but the current state of the art is still far from equalling human performance in the area of recognizing speech, for example. The models of different neural network architectures are specified according to network topology, node characteristics, and the learning or training method employed.

Work on artificial neural network models dates back over 50 years with the paper of McCulloch and Pitts [McCulloch and Pitts, 1943] generally acknowledged as marking the start of work in this field, even if the authors make no mention of the practical uses of these models. During the past 15 years, neural nets have seen a renewed interest as the work of many researchers have brought about significant advances in this field, both for the development of new architectures, as well as in the formulation and implementation of improved training methods [Hecht-Nielsen, 1990].

Neural nets have become popular in pattern classification due to their inher-

ent parallelism and the simple computations required vis-a-vis traditional classifiers, which either require higher computational complexity, or which require longer times to perform their task which may preclude real-time operation. It was thought that because of these strengths, neural networks would be well suited to speech recognition, but their successes and the proliferation of this methodology in this particular domain has still not consistently surpassed that of hidden Markov models. This has been mostly due to the fact that the majority of the commonly used architectures only support static input pattern sizes. This causes a particular problem for applications such as word recognition, for example, or in any other application which has inputs that vary both in length and in where the "relevant features" required for correct classification occur within the input pattern [Morgan and Scofield, 1991]. Despite these limitations, however, a number of neural network architectures and neural network hybrid architectures have emerged over the past few years, as was presented in chapter 2.

Although hidden Markov models (HMMs) represent the method of choice for the majority of speech recognition applications, this methodology was thought to be too complex to segment and to train given the nature of the pattern that was to be classified in this particular domain. In speech, recognition of words requires the identification of specific phonetic events occurring in a particular order [O'Shaughnessy, 1987, Rabiner, 1989]. Hidden Markov models address this using a temporal ordering of the nodes, which is meant to correspond to the occurrence of certain acoustic events during the course of a word utterance. In these models, called *left-to-right* HMMs, an example of which is shown in figure 4.1, once a particular state has been traversed, it cannot be revisited, thus imposing a temporal order of phonetic events which is inherent in the words to be recognized by the system.

For infant cries, however, there seems to be no apparent temporal order in the features present in utterances of the same type of cry. As well, auditory identification or classification by adult listeners seems to focus on the *occurrence* or

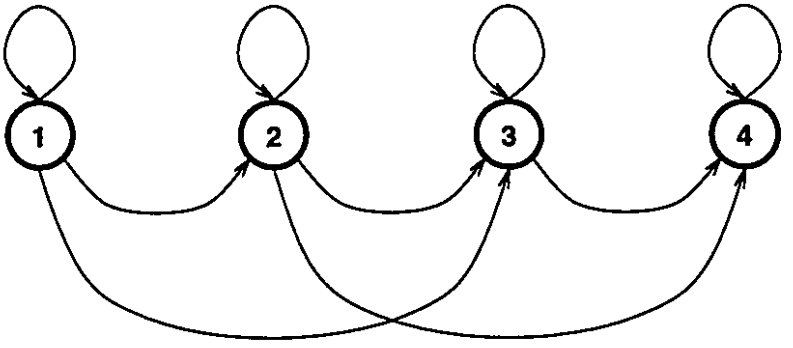


Figure 4.1: A Left-to-Right Hidden Markov Model

presence of specific acoustic events in a cry, coupled with cues based on intensity, and on the prosodic evolution of the F_0 contour [Wasz-Höckert *et al.*, 1968, Zeskind and Lester, 1978, Fuller, 1991]. Consequently, straight pattern identification techniques would seem to be better suited, potentially more successful, and also less complicated to train and test than a corresponding HMM. Consider, for example, the presence of dysphonic segments in an utterance. These events are a common occurrence in pain cries [Johnston and O'Shaughnessy, 1988], and are usually present at the the beginning of an utterance. However, these events are not limited to the start of the utterance; a dysphonic segment can occur anywhere in the course of an utterance.

Note the difference between the above and an application of word recognition in speech, where the position of where a phone is identified in a sequence of phones could determine whether one of two words in the vocabulary is recognized. For example, the identification of a phone such as "a" in a sequence could make the difference between an input word being classified as "able" or "bale". Consequently, for speech, both the order and presence of phonemes makes all the difference, whereas for acoustic events in infant cries, the occurrence or presence is important, regardless of order since the presence of certain acoustic events reflects articulator position and vocal tract tenseness, which, as was mentioned in section 2.1, is though to differ according to infant state. In addition to this, the intensity of the utterance and the prosodic evolution of F_0 is important as well, as opposed to word

recognition, where these features are considered a nuisance for recognition purposes, and are usually excluded from the feature set [Bourlard and Morgan, 1993]. Also, given the fact that newborn or premature infants have very poor control over their vocal tract articulators, lends additional support to the hypothesis that the problem of the correct classification of infant cries may not necessarily benefit from information of a sequential nature.

Neural networks have been used in both phoneme and speech recognition-related experiments, but the results published in the literature show more success for applications with the former than with the latter. Infant cry vocalization resemble vowel vocalizations, and can perhaps benefit from the phoneme classification-related work done in speech using artificial neural networks.

Although a cry utterance may not necessarily be considered as being a left-to-right first-order Markov process, where the probability of transition from the current state to another state depends solely on the current state, this should by no means imply that hidden Markov models could not be useful in this domain. One would have to replace the left-to-right model, shown in figure 4.1, used in the speech domain by an *ergodic* model, shown in figure 4.2, where transitions from a given state to all others are allowed.

The process of training such a model would be quite time consuming and there would be no guarantee that the results obtained using HMMs would be superior to those achieved using neural networks. As well, the training of the HMM requires a large number of training set data so that the training algorithm can learn to properly approximate the probability density functions of the observation symbols in the individual states and the state transition probabilities accurately. For infant cry utterances, it is often difficult to obtain large numbers of recordings, unlike speech, for example, where large databases of speech samples exist for testing and comparing the results of both feature extraction routines and speech recognition methods.

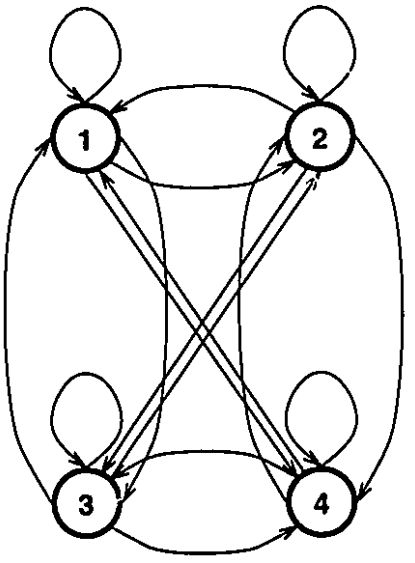


Figure 4.2: An Ergodic Hidden Markov Model

For these reasons, then, it was decided that first a series of tests using ANNs would be attempted for the purposes of classifying one of three different cry types, of which a detailed description is given in section 4.3. Pending the results of these tests, it would then be decided whether or not these tests would be abandoned in favour of a new sets of tests, this time performed using another classification methodology such as classification and regression trees [Breiman, 1984], or HMMs. The results obtained in initial tests were sufficiently good to warrant their continued testing. Further tests were performed on other neural network architectures for comparative purposes and in order to determine if certain architectures or input features yield better results and why. No other group to date has used ANNs in the domain of infant crying, or if they have, their results have not been published in the literature and it is important to have results which can be compared with other work, using the same or similar data sets.

Work done by Xie, Ward, and Laszlo used hidden Markov models to compute a cry's so-called *level-of-distress*, which is a subjective measure based on a parent's *perception* of the infant's physical and emotional state after listening to a cry [Xie *et al.*, 1993]. Although their method mentions the identification and use of "cry phonemes" in an HMM, no implementation details or error analysis measures

were discussed in the paper, even if the correct classification value of this level-of-distress measure was quoted as being over 80%. The validity of this measure, and the relation that this subjective measure has to specific infant states, such as pain or hunger, for example, is not clear. Nevertheless, this method was one of the first published in the literature which attempted to automate the classification of cries, even if the method is based more on the perceptions of the listener than on an understanding of the underlying cry production process which causes these cries to be uttered, or controlled experiments affecting the underlying physiological and emotional states of the infant producing the cry.

The results obtained as a result of the experimentation performed for this dissertation using ANNs, which are presented in section 4.6, achieve correct classification rates equalling or surpassing those achieved by Xie, Ward, and Laszlo. It should be noted, however, that the the classification experiments described in section 4.6 attempted to discriminate between three different cry states, and did not try to match the perceptual measures of infant distress as interpreted by adult listeners.

Moreover, neural networks were also selected because of their success in certain facets of the speech recognition problem, most notably in vowel recognition and phoneme recognition, as well as due to some successes in word recognition resulting from the use of architectures that incorporate time in them. Since ANNs possess the potential of equalling the classification rates of the best statistical recognizers for certain applications [Niles *et al.*, 1989], this implies that they are at least worth a closer look. While no neural network training method can yet guarantee that the set of weights generated after a given training session converges to the optimal set of weights, research is still on going in this problem to ensure that the weight values will approach this "optimum" as closely as possible.

The area of finding the optimal training methods, architectures, and input feature sets which will yield the best results for both speech and similar applications with time varying signals, such as infant cries, remains very much an open problem,

with limited success in the speech domain having been achieved, as mentioned in section 2.3.

The strengths of ANNs lie in their ability to map classification sets, and these optimal mappings are achieved if the training set is sufficiently large and representative of the data which is to be expected in test patterns or during actual use of the network [Haykin, 1994]. The premise is that if the training data is especially representative of the data which can be expected and if the network appropriately models of the classification spaces for the input and output sets, then the network will generalize and perform correct classification on the input patterns which it has previously never "seen". To perform this generalization task as well as possible, the training size should be comparable to the number of input weights [Hecht-Nielsen, 1990]. This is not possible for most applications, however, and other tasks must be performed to accurately determine the performance of a given network configuration. In the absence of sufficient input patterns, methods exit to assist in this determination [Weiss and Kulikowski, 1991]. Among these includes training for a fixed number of iterations, training until an error measure, such as the mean-square error, drops below a certain value, randomizing the sequence in which the patterns are presented, and **cross-validation** training. The latter method requires taking k patterns randomly from a data set of size n, and using $n - k$ patterns as the training set and the remaining k patterns as the test set, repeating the process until all the n patterns in the entire data set have been in the test set once.

The following section presents and discusses the neural network architectures and training algorithms used for the purposes of evaluating their ability for classifying infant cry vocalization uttered as a result of three different stimulus events.

4.2 Neural Network Paradigms Tested for the Classification of Infant Cry Vocalizations

This section presents the different neural network architectures and training paradigms tested for the classification of infant cries. A number of different architectures were investigated, but the results of only four are presented in the interest of presenting the most successful of the methods tested. Although the literature lists a number of different architectures which have been successfully used for either phoneme or word recognition, most of these architectures were not tested for the following reasons. First, the study undertaken for this dissertation was performed to investigate the feasibility of using ANNs for infant cry classification; if reasonable results could be obtained with these more "traditional" nets, further tests with more complicated methods could then be attempted in the future, as was mentioned in the previous section. Next, the availability of software implementations of certain neural network architectures and learning paradigms, such as probabilistic restricted Coulomb energy networks [Scofield *et al.*, 1988], or Viterbi networks [Lippmann and Singer, 1993], precluded the testing of these methods with the cry recordings. Finally, ANNs and learning paradigms that were available for testing via software implementations, available in the public domain, and suitable for time-dependent signals, such as speech, were tested.

The following subsections will present the four paradigms used in cry classification tests. They are the basic feedforward neural network (FF), the autoregressive neural network (RNN), the time-delay neural network (TDNN), and the cascade-correlation neural network (CC). A number of different algorithms for training these ANNs exist as well, all of which attempt to adapt the network weights according to a particular training regimen in order to achieve the desired end result of the training process. All of the learning methods used for training the various neural networks architectures for this particular application, which will be presented in the subsequent subsections, used supervised learning techniques. These

training methods attempt to find the optimal set of weights which leads to the convergence of the difference between the desired and actual output values for the inputs presented to the network, according to some error measure. Once the network has converged, the weights values will represent a minima in the error surface, and it is hoped that after training, the global minima of the error surface, for which the weights constitute the dimensions, is found.

Supervised learning techniques assume that the desired function of the network is to perform as an input/output system where the inputs to the network x_k have desired output values y_k associated with them. The stimulus pattern is presented at the input and the corresponding desired output values are presented to the output of the network. If the output of the network resulting from the presentation of the input pattern does not correspond to the desired output pattern within an acceptable error level, the network weights are then modified in such a way as to reduce the difference between the actual and the desired output values. The weights can also be modified after a group of input patterns are presented to the network, or after the entire set of patterns are presented. The following section will also discuss the various neural network training algorithms associated or used with the corresponding neural network architectures.

4.2.1 Feedforward Neural Network Architectures

Overview

This is perhaps the simplest of all neural network architectures, from which all other networks have evolved, and thus the most logical network architecture with which to begin. Many of the concepts presented here will be valid for the other networks as well, and as a result, this subsection will be somewhat longer than the subsequent ones.

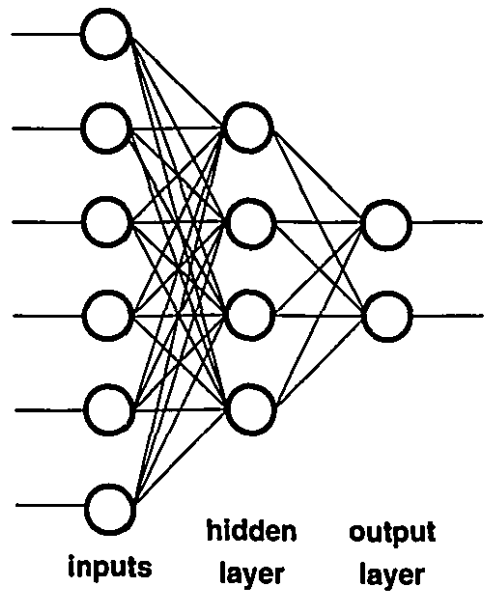


Figure 4.3: A Simple Feedforward Neural Network

An example of a simple feedforward neural network is shown in figure 4.3. The network is organized into layers according to where a certain node or cell in the network receives its inputs from. The connections in this network are unidirectional with the information flowing from left to right. In this architecture, connections are permitted only between neighbouring layers and these connections cannot loop backwards from nodes on the right to those on the left. The hidden layer receives its signals from the inputs and the output of the hidden nodes are in turn propagated to the output nodes.

The individual nodes in the hidden layer and output layer may have one of a number of transfer functions, $f(\cdot)$, which transforms the weighted sum of the signals that it receives from the previous layer, x_w^q, and presents this value at its output, x^{q+1}. Examples of commonly used transfer functions are the hyperbolic tan function, logical function, linear function, and the signum function [Simpson, 1990]. Usually, non-linear activation functions are used since this allows the network to compute high-order correlations of the inputs which are not possible using simple linear activation functions [Dayhoff, 1990]. It can be shown that a three-layer neural network, with non-linear activation functions, is sufficient to compute an arbitrary

mapping between input and output values [Haykin, 1994].

The choice of using a transfer function such as the hyperbolic tan function, which has outputs ranging between $[-1, 1]$, over the logistic transfer function, which has outputs ranging between $[0, 1]$, are that the range of the former is twice as large as that of the latter. As well, the tanh transfer function is asymmetric, that is $f(-v) = -f(v)$, and a multi-layer network will learn faster when trained using back propagation [Werbos, 1974], one of the most popular ANN training methods, which will be outlined below. A derivation of the method can be found in Haykin's book [Haykin, 1994], or in other text books dealing with neural networks.

Training Methods

The back-propagation algorithm for weight updates on a pattern-by-pattern basis behaves as follows. First the network is initialized with all the weight values set to small random values, in order to avoid the saturation of the majority of the network nodes. Next, the training examples are presented at the inputs, one pattern for every iteration, with the activation potentials, or the transfer function output values, of the nodes computed based on the weighted input sum at its inputs. This is done for all the layers in the network, proceeding from the inputs to the outputs. The activity of neuron j in layer l is given by the equation

$$y_j^{(l)}(n) = f\left(\sum_{i=0}^{p} w_{ji}^{(l)}(n) y_i^{(l-1)}(n)\right) \tag{4.1}$$

where $y_i^{(l-1)}(n)$ is the output neuron i in the previous layer $l-1$ at iteration n, $w_{ji}^{(l)}(n)$ is the weight of the connection between neuron i in layer $l-1$ and neuron j in layer l, $f(\cdot)$ is the activation or transfer function of the node, and $y_j^{(l)}(n)$ is the output of neuron j in layer l for iteration n.

At the output layer, layer L, the output for a node j can be defined as being $y_j^{(L)} = 0_j(n)$, and the error signal can be computed as $e_j(n) = d_j(n) - O_j(n)$, where

$d_j(n)$ is the desired output of node j at iteration n. This error is then propagated backwards to modify the weight values in such a way as to decrease the error between $d_j(n)$ and $O_j(n)$. This is accomplished by computing the local gradients, or the rate of change, of the weights, δ, by proceeding backwards on a layer by layer basis, starting at layer L - the output layer, as denoted by the superscript over the δ:

$$\delta_j^{(L)}(n) = c_j O_j(n)[1 - O_j(n)] \tag{4.2}$$

for the j^{th} neuron in the output layer L and

$$\delta_j^{(l)}(n) = y_j^{(l)}(n)[1 - y_j^{(l)}(n)] \sum_k \delta_k^{(l+1)}(n) w_{kj}^{(l+1)}(n) \tag{4.3}$$

for the j^{th} neuron in hidden layer l.

The weights in the network at layer l are then modified according to the generalized delta rule:

$$w_{ji}^{(l)}(n+1) = w_{ji}^{(l)}(n) + \alpha[w_{ji}^{(l)}(n) - w_{ji}^{(l)}(n-1)] + \eta \delta_j^{(l)}(n) y_i^{(l-1)}(n) \tag{4.4}$$

where α is the learning rate and η is the momentum. The learning rate determines how much of the difference between the two previous weight values is added to the current change, and the momentum term determines how much of the gradient contributes to the weight change.

This process terminates when the error for either all the patterns in the training set, or for the error summed over all the patterns in the training set, drops below a certain value. One common error measure used is the mean-square error:

$$\mathcal{E} = \frac{\sum_j |d_j^L(n) - O_j(n)|^2}{j}, \tag{4.5}$$

but other errors measures can be used as well [Morgan and Scofield, 1991].

Instead of updating the weights after the presentation of each pattern, the weight values can also be updated following the presentation of all the patterns in the training set. The latter is referred to as batch training, whereas the former is referred to pattern mode training. Pattern mode training was described at the beginning of this sub-subsection. In batch processing, one waits to propagate the error backwards after all the patterns in the training set have been presented. Both methods have their respective strengths and weaknesses and the better results obtained by the use of one method over another depend on the nature of the problem [Haykin, 1994, Weiss and Kulikowski, 1991]. In experiments performed for the classification of infant cries, which will be described in section 4.6, both pattern mode and batch mode training were employed.

In addition to the training method presented earlier, other training methods exist for updating the weight values; attempting to reduce the number of training iterations required before the network converges, as well as optimizing the value of the weights obtained following the completion of the training process, in such a way as to avoid the occurrence of getting stuck in a local minima. Such methods, which attempt to improve on the standard back-propagation, are QuickProp [Fahlman, 1988] and gradient descent line search training techniques which are based on optimization theory [Goryn and Kaveh, 1991].

Neural Network Connections and Configurations

The number and configuration of the connections between the different layers also distinguishes between different types of feedforward networks. Fully connected networks, such as the one shown in figure 4.3, are ones where all the nodes in a given layer have connections to all the nodes in the subsequent layer. One can also selectively choose to connect certain groups of nodes from one layer to a limited number of nodes in the following layer thus restricting or localizing the information that is passed from one layer to the next in this fashion. One formal

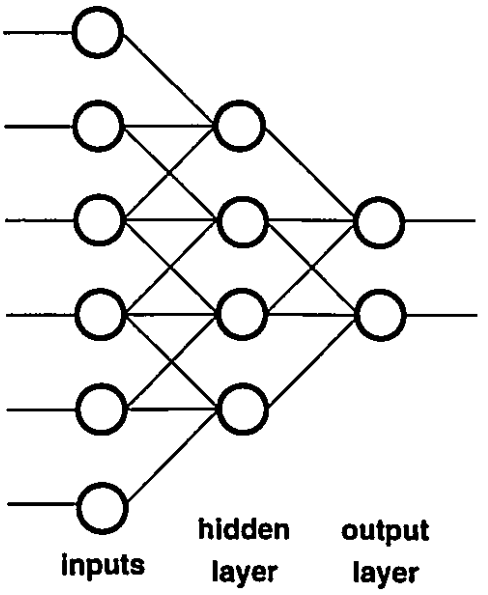


Figure 4.4: A Simple Feedforward Neural Network with Tessellated Connections

method of localizing the connections among a group of nodes is referred to as tessellation or "tiling" and involves connecting "tiles" of nodes from one layer to the next, as is shown in figure 4.4. In this figure, each node in a given layer, has connections from three nodes in the previous layer with two nodes from the previous layer overlapping between adjacent nodes in the subsequent layer. This allows the network to selectively integrate the activations from a group of nodes from the previous layer to the following layer. These types of connections are meant to model the receptive fields in the neural anatomy of the brain, and, by restricting the connections in this manner, the performance of tessellated-connection networks can either equal or better those using full connections [Dayhoff, 1990].

Lastly, a network with additional hidden layers can learn more complex mapping functions between the inputs and outputs than can be achieved with two-layer networks. For some applications, the use of multiple hidden layers has achieved better results as these networks perform the computation of higher-order correlation functions between the input and output.

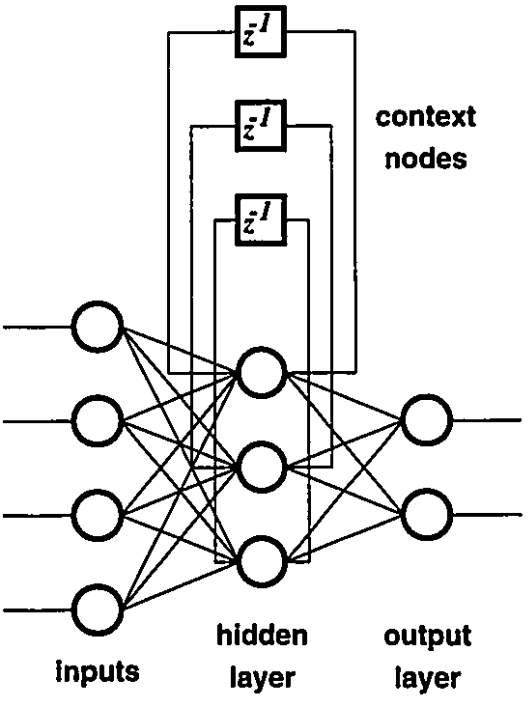


Figure 4.5: A Simple Recurrent Neural Network

4.2.2 Recurrent Neural Networks

A means of capturing time-dependent information in an input set of data, or for tracking sequences in an input pattern that varies with time, is through the use of recurrent neural networks (RNNs). Figure 4.5 shows an example of a simple recurrent neural network. This network takes the outputs from the node activations, or transfer functions, from the hidden layer, and feeds these values back to the inputs of the same node, delaying these values by one or more time instants. Figure 4.5 shows an example of only one delay unit per node, however, the outputs of the hidden layers could be delayed by additional time instants, simply by placing additional delay nodes, z^{-1}, in the network. The delay units store what is commonly referred to as *context information* since they store the context of the network at a particular time instant for use as future input values to the nodes. These context nodes can be added to the output units as well, enabling the state of these nodes to be captured, in addition to the state of the hidden layer.

These networks have the ability to learn the inherent "states" of a sequence of input vectors and are able to capture time-dependent information of sequences in a pattern that varies with time.

One of the methods with which these neural networks can be trained is through the use of the "temporal flow" network as described by Watrous and Shastri [Watrous and Shastri, 1987]. The behaviour of this method allows the context weights to vary and uses the standard back propagation algorithm to determine the value of all the weights in the network, treating the context weights simply as weights originating from the previous layer. After all the weights have been adjusted, the activation of the nodes are recalculated for the next pass of the weight update training algorithm.

Unlike static feedforward neural networks, where the output is based on the presentation of a static input pattern, the output values for recurrent neural networks vary with time. The selection of these time-varying target output values is quite arbitrary, although in their original experiments, Watrous and Shastri used a ramp function as their target response over the course of a pattern's presentation. For output values lying between 0 and 1, the ramp was initially started at 0.5 for all the outputs, and slowly increased towards 1 for the node whose desired response at the end of the pattern was 1, and decreased to 0 for the other output nodes. After the network was trained, the "winning" node in the output layer was the node with the largest output value of all the nodes for the subsequent testing phase,

Besides the ramp function, another time-varying function which has a desirable time-varying characteristic for the purposes of training the network is the Gaussian function. The nice feature of this function is that the values initially rise or fall quickly, before levelling off asymptotically towards either 1 or 0. The use of the Gaussian function has resulted in some faster convergence times than when the ramp function was used.

In the configuration of a recurrent neural network, the input size per time

interval and the overlap size between subsequent input patterns can also be varied. As well, the number of time delay units in the context and output nodes can be varied in order to determine the "granularity" of these parameters which yields the best results. Consequently, a static input pattern can be represented as a sequence of time dependent frames where both the frame size and the number of frames to be presented to a recurrent neural network can be varied. This allows the determination of whether or not time or sequence is important for a given classification problem, and which frame size, number of frames, and delay units in the hidden and output layers yield the best results. From this information, it can then be determined if and what granularity of time information is relevant for a given data set.

The time information extracted by this type of network is based on the activation values of the hidden nodes and of the output nodes, which differs from the time information encoded in the neural network architecture described in the following section.

4.2.3 Time-Delay Neural Networks (TDNNs)

This architecture was originally developed by Waibel, Hanazawa, Hinton, Shikano and Lang [Waibel *et al.*, 1987] in order to model the dynamic nature of speech by attempting to represent and capture the relationships between the different spectral and acoustics events in a given signal over time, while providing invariance to slight shifts in time between the various input frames. The latter feature of this network is designed to tolerate the imprecise segmentation and alignment of an input pattern so that the relevant acoustic events in the input frames for the same output class can occur either somewhat sooner or later in time without affecting recognition.

Figure 4.6 shows an example of a TDNN node. Unlike the feedforward neural network nodes, where the I inputs to the node are weighted and summed before

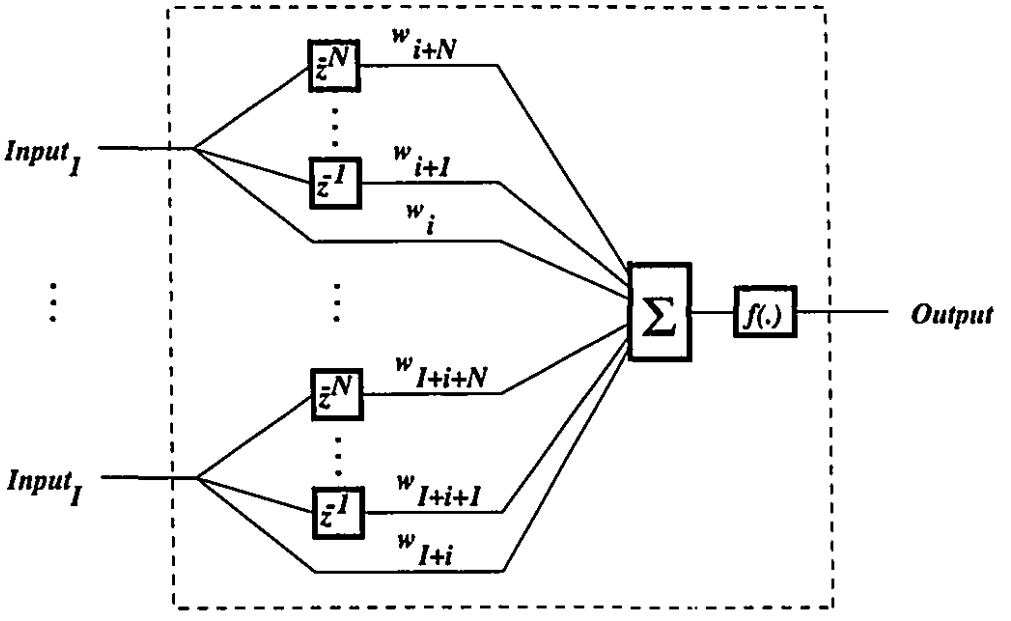


Figure 4.6: A Time-Delay Neural Network Node

being presented to the activation function, the TDNN node augments this definition by introducing a series of delays $z^{-1} \ldots z^{-N}$ for all the input signals leading to the node. In turn both the undelayed and the delayed inputs are weighted before being summed and presented to the node's activation function, so that the number of weights required for this node are $I(N+1)$, where N denotes the number of delays in the node.

A group of these nodes are then placed together in a TDNN, as shown in figure 4.7, where the number of delays in the hidden layer is represented by the number of input frames that are presented to a hidden layer node. This value is one less than the number of hidden layer frames. In figure 4.7, the hidden layer consists of three hidden nodes, counted horizontally, with the current input value and $N = 3$ delay units per node, counted vertically. With respect to the notation indicated in figure 4.6, figure 4.7 has $I = 12$ inputs per node, consisting of two six-element vectors, illustrated in the bold rectangle at the inputs, and $N = 3$ delay units in each of the three hidden layer nodes.

Consequently, a sequence of two six-element vectors, or twelve input values,

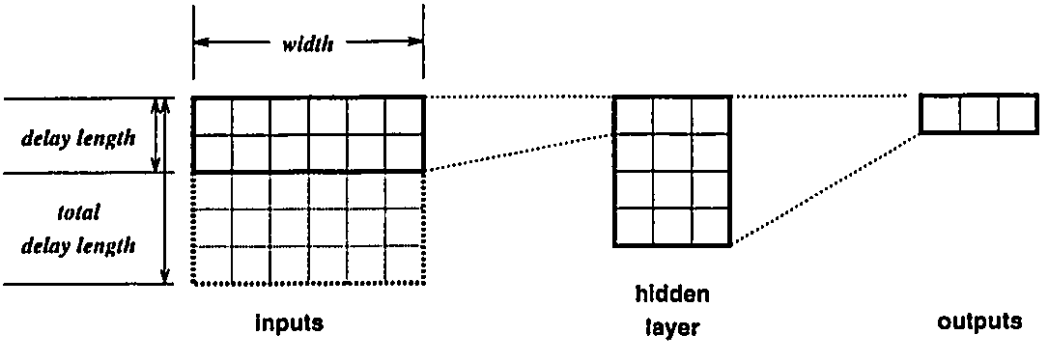


Figure 4.7: A Time-Delay Neural Network Definition

are presented at any given time instant. At the following time instant, the next group of two six-element vectors are presented, consisting of one of the input vectors from the previous frame, and one new input vector, with the activation of the previous group of the input data frame is delayed by one. This continues until all the groupings, or the total delay length of vectors, in the frame have been presented. In the example of figure 4.7, this occurs three time instants after the first group from the input data frame is presented. When the last grouping of two six element vectors has been presented to the network, the hidden layer activations of the first group from the input data frame have been delayed by three time units. Hence for this simple example with a delay length of two and a total delay length of five, a complete input frame consists of five vectors consisting of six elements per vector, presented in to the network in groups of two vectors.

Although the network of figure 4.7 uses only one hidden layer, other hidden layers can be added, with subsequently hidden layer nodes integrating the activations of the previous hidden layer nodes over time. In short, this architecture allows acoustic events occurring in the sequential input groups of a given input frames to be integrated over time. Although the network may seem complex, TDNNs are trained using back propagation.

In order to determine if this architecture is suitable for a given application, a number of different parameters in the network can be varied. First, the input delay length can be varied, which in turn changes the number of delays in the

hidden nodes, or the number of hidden node frames for the network. The larger the input delay size, the smaller the hidden layer delay size, and the *coarser* the time representation and time integration of features. Conversely, the smaller the delay length, the larger the hidden layer size, and, consequently, the *finer* the time representation, allowing the integration of spectral and acoustic features to be integrated over smaller time slices. In the latter cases, it may also be wise to experiment with adding a second hidden layer as well. This allows the network to capture both the input and time sequence representations by using a smaller number of delay units in the first hidden layer, using the second hidden layer to integrate a reduced dimension of activations from the first hidden layer.

These networks were originally shown to be successful for phoneme recognition [Waibel *et al.*, 1989] with performance topping that of hidden Markov models [Waibel *et al.*, 1988]. This improvement in phoneme recognition achieved by TDNNs over HMMs has not necessarily translated to improved results for word recognition, however. These TDNNs have the drawback that the learning procedure is rather lengthy, which is necessary in order to update the potentially large number of weights. As well, if true shift invariance is to be achieved, a large number of training tokens are necessary to both compensate for inaccurate segmentation techniques and for variable length utterances.

4.2.4 Cascade Correlation Neural Networks

Cascade correlation is an example of an algorithm that constructs its own hidden layer by adding hidden nodes based on the network error after a previous training iteration. This paradigm was developed by Scott Fahlman at Carnegie Mellon University [Fahlman and Lebiere, 1991]. An example of a cascade correlation network is shown in figure 4.8. This architecture has an intuitive appeal in that it will only create as many hidden nodes as it needs in order to get the network error to fall below a desired value. Consequently, one does not have to determine the optimal

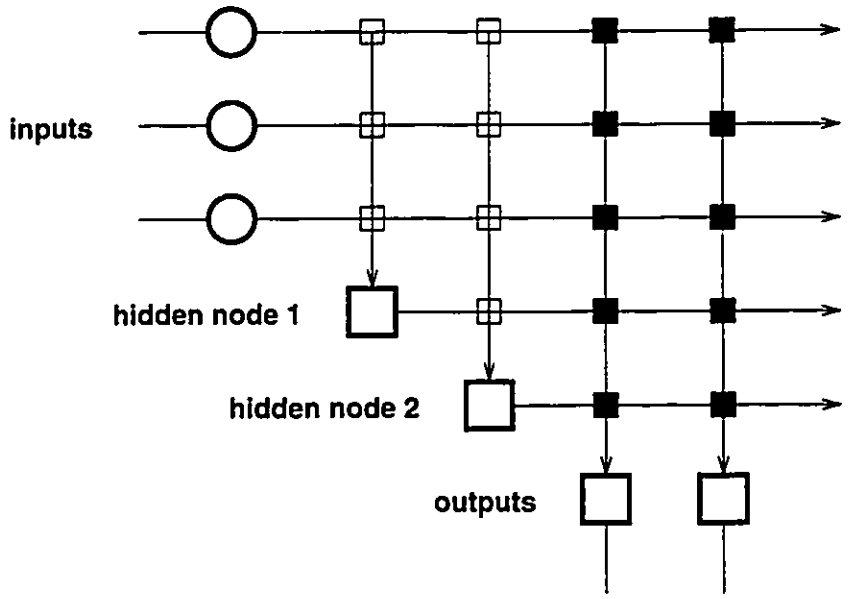


Figure 4.8: A Cascade Correlation Network

number of hidden nodes for an application to correctly or optimally learn the relations between the input and output values; the learning algorithm will determine this by proceeding in the following manner.

First, the network begins with only the input and output nodes, which are fully connected. The algorithm begins by adjusting the weights between the input and the output nodes, so as to minimize the network error, using either back propagation or another gradient descent learning algorithm. This portion of the training phase continues until either the network error no longer improves, a fixed number of iterations have occurred, or the network error goes below a predetermined value in which case the network has converged and training is stopped. Otherwise, the algorithm proceeds to train a set of candidate hidden nodes.

These candidate hidden nodes are linked to the existing network with connections coming from the inputs nodes only, with no connections to the output nodes during this training process. During the training of these candidate hidden nodes, the weights connecting the inputs to the candidate hidden units are adjusted so as to maximize the correlation between the activation of the candidate hidden units

and the residual error of the network, that is, the error of network at the output nodes when training stopped in the preceding stage. The values of the weights are adjusted using either back propagation, or a gradient descent learning method.

Training of these candidate nodes continues until the correlation value no longer increases, or, as is the case for the output weights, after a fixed number of iterations have occurred. The candidate node with the highest correlation value is added to the network with its weight connections from the inputs fixed at the values determined during the training of the candidate weights. The output of the hidden node is then connected to the inputs of the output nodes, so that the output nodes receive weighted inputs from the both the inputs and the hidden nodes, as figure 4.8 illustrates. The weights of these connections between the hidden and the output nodes and those of the connections between the inputs and the output nodes are then adjusted during the subsequent training session as the weights between the input and output layers were adjusted initially.

This cycle continues until the network error during the training of the the output nodes drops below a predetermined value. Subsequent hidden nodes which are added to the network receive inputs from both the input units and from the previous hidden nodes and hence the hidden nodes are cascaded in this fashion.

In short, the clear boxes in figure 4.8 denote the weights which are set a result of the candidate node training, and are fixed once a candidate node is added to the network, and the solid boxes denote the connections from either the input or hidden nodes to the output nodes and which are not fixed, and thus change after the hidden nodes are added to the network.

When compared to the other three neural network architectures presented in the previous subsections, the cascade correlation architecture and learning methods have not been used in numerous applications. However, the prospect of a network which grows a hidden layer in response to the way that the network error changes is indeed an appealing one. The purpose behind using this method is to determine

whether the classification rates of infant cries would indeed benefit from the use of this "tailor made" hidden layer.

4.3 Data Set and Experimental Set-Up

This section describes the data set used for the neural network tests for infant cry classification and the feature sets which were derived from the cry recordings for use as inputs into the various neural network architectures described in section 4.2.

The utterances to be classified in this experiment were a subset of the data set used for the fundamental frequency extraction tasks described in section 3.3. The set of recordings consisted of 238 utterances recorded at the Nôtre-Dame-de-Grâce CLSC (Community Health Clinic) from sixteen healthy infants ranging in age from two to six months, with no history of perinatal or postnatal complications. All the parents of the infants gave their informed consent to participate in this study. All cry vocalizations in this data set were due to one of three stimulus events: **pain / distress** from a routine immunization, **fear / startle** from a jack-in-the-box, and **anger / frustration** from a head restraint. Recordings were made on a Sony TCM-500DEV cassette recorder with an omni-directional Senheiser MKE 2 microphone placed 10 cm from the infant's mouth. Subsequent to low-pass filtering at 8 kHz, these cassette recordings were then digitized using a Data General D2701A card, using a 12-bit analogue-to-digital converter, on a personal computer, at a sampling rate of 16 kHz. These digitized signals were then transferred to a Sparc 10+ for subsequent analysis, feature extraction, and classification experiments.

Recordings containing cry utterances with a minimum duration of 0.75 seconds were used for feature extraction. Of the 238 recordings in this data set, 195 had vocalizations with durations that satisfied this criterion. The other 37 recordings were discarded from the study.

4.4 Parametric Representations

Since important events in the cry signal are thought to occur in the first second of the utterance following the onset of the cry after the stimulus event [Johnston and O'Shaughnessy, 1988], the first second of utterances lasting at least 0.75 seconds after the cry onset were used for the subsequent feature extraction data set to be used in the classification experiments. For utterances that did not last for at least one full second but which lasted at least 0.75 seconds, the last frame of parameters was extended to fill the empty frames. If important features useful for classification lie in this portion of the cry, classification could be accurately performed using the extracted features, provided, of course, that these features are relevant from an auditory point of view. The motivation behind using these types of features is that since a human listener can distinguish between certain types of cries [Zeskind *et al.*, 1985], or can learn to differentiate between different types of cries [Ostwald and Murry, 1985], then if a classification method is presented with features derived from an understanding of the human auditory system, hopefully, the automatic classification system will also "learn" to identify the relevant features in the data and perform classification based on these features.

Two feature sets have been successfully used for speech recognition and derive from an understanding for the human auditory system in general, and in the frequency response of the cochlea in particular [Davis and Mermelstein, 1980, O'Shaughnessy, 1987]. They are the mel-based cepstral coefficients [Davis and Mermelstein, 1980], and the mel-scale filter-band energies [O'Shaughnessy, 1987]. The cochlea's frequency response characteristic is such that the hair cells at low frequencies have a higher resolution than those at higher frequencies. One of the first representations of the hair cell center frequency values and bandwidths was published by Zwicker [Zwicker, 1961]. This brief article shows a linear spacing of frequency bands with very narrow bandwidths for frequencies below 1 kHz, and logarithmic spacing and corresponding bandwidths

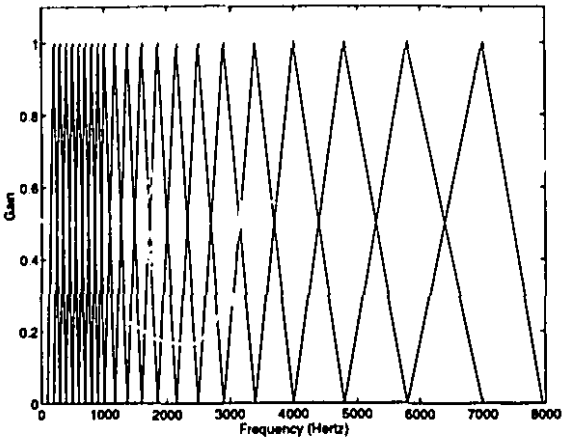


Figure 4.9: Filter Bank for Mel-Cepstrum Coefficient Generation

for values above 1 kHz. This representation is also known as the "Bark" scale or "mel" scale. A number of studies have shown that using these representations for the purposes of speech recognition yields superior results to using a fixed band or linear scale representation of similar features [Davis and Mermelstein, 1980].

Using the research done for the parametric representation of speech for speech recognition as a starting point for the representation of infant cry vocalizations, the following two feature sets were extracted from the signal; 10 mel-cepstrum coefficients and 19 filter-band energies per frame of cry utterance data.

To extract these feature sets, the first second of the aforementioned recordings containing utterances lasting at least 0.75 seconds were segmented into a series of 16 ms or 256 sample frames, with subsequent frames overlapping by 50%. Consequently, for a 1 second portion of the cry utterance, 125 frames of feature vectors would be generated.

Generation of the mel-based cepstrum coefficients begins by taking the discrete Fourier transform (DFT) of the Hamming-windowed signal frame. Then the output spectrum is passed through as series of triangular band-pass filters which model the bark or mel-scale, and the log energy output values of these filters are calculated. Figure 4.9 shows a representation of the 21 triangular band-pass filters used to filter the DFT of each frame of the utterance. The overall frequency response of

these band-pass filters sums to unity for the portion of the signal in the area of interest. Once the 21 critical band energies, E_k, are calculated, the 10 mel-cepstrum coefficients are calculated according to the following formula:

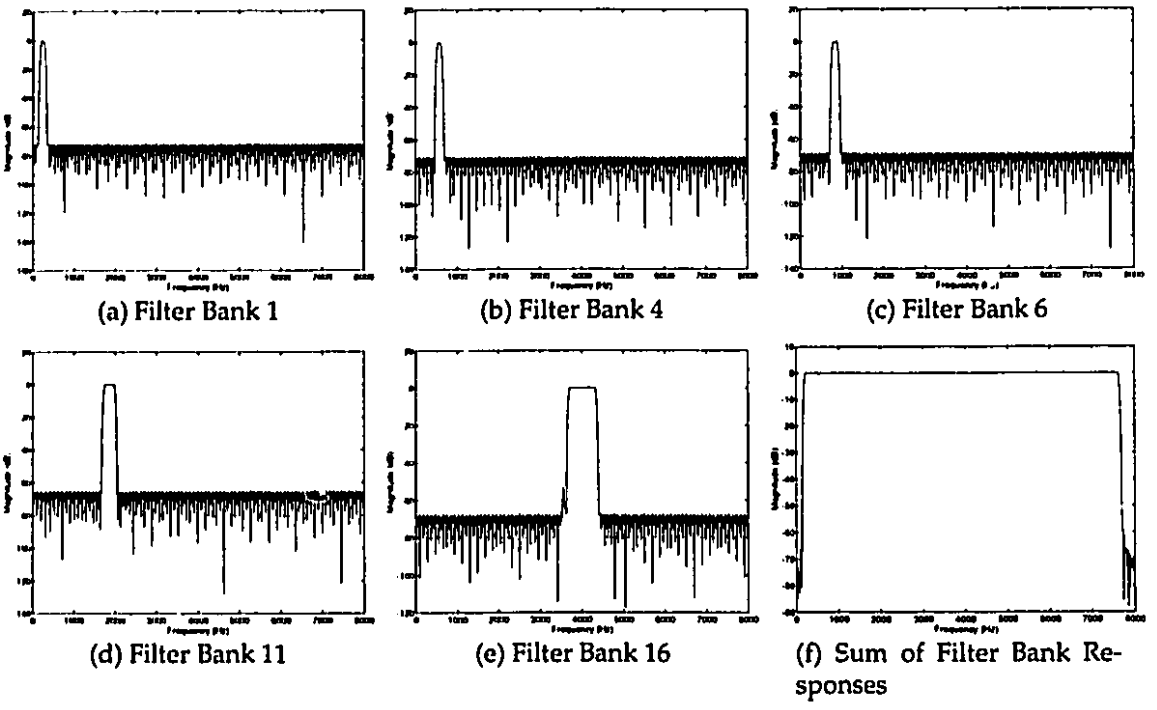
$$c_n = \sum_{k=1}^{21} \log(E_k) \cos\left[n(k - \frac{1}{2})\frac{\pi}{21}\right] \tag{4.6}$$

for $n = 1, 2, \ldots, M$, where in this particular case, $M = 10$. As well, the zeroth component of the mel-cepstrum coefficients, c_0, corresponding to the average energy of the frame is also included, so that in all, 11 coefficients, $c_0, \ldots, c_{10}$, constitute the mel-cepstrum vector for a given input frame or window of data.

As was mentioned earlier, the 19 mel-scale filter-band energy values are also generated as features. Generation of this feature set is done be first generating a series of band-pass filters whose center frequency and bandwidths approximately followed the critical band values. These filters were then generated using the Remez Exchange Algorithm for generating linear-phase finite impulse response (FIR) filters.

The difference between the number of *critical-band* filters used in the mel-cepstrum coefficient computation and in the computation of the *mel-scale* filter-band energy values is due to the fact that for the filter-bands in the range of 0 Hz to 1000 Hz, the filters generated using the Remez Exchange Algorithm required slightly larger bandwidths than those specified by Zwicker, which were used in the mel-cepstrum computation, in order to obtain unity gain in the pass-band. Moreover, using slightly larger bandwidths decreased the number of filters used for the mel-scale filter-band energy values in the range of 0 Hz to 1000 Hz from 9 to 7, but these larger bandwidth filters resulted in unity gain in the pass-bands of these filters, and consequently the sum of the magnitude of the frequency responses in this range summed to unity as desired. This criterion could not have been satisfied if narrower bandwidth filters were generated using the Remez-exchange algorithm.

The motivation behind using FIR filters, as was also mentioned in section 3.3,



(a) Filter Bank 1 (b) Filter Bank 4 (c) Filter Bank 6

(d) Filter Bank 11 (e) Filter Bank 16 (f) Sum of Filter Bank Responses

Figure 4.10: Some Filter Bank Responses for Mel-Scale Filters

is that because of their linear phase characteristic, all the frequency components in the signal are delayed by the same amount of time during the filtering process, and thus, no signal distortion occurs. A comprehensive article by Dautrich, Rabiner, and Martin discuss the benefits of using FIR filter over infinite impulse response (IIR) filters for the purposes of speech recognition [Dautrich *et al.*, 1983].

Despite the desirable linear-phase characteristics of FIR filters, these filters have the drawback that they require an order of magnitude more taps than IIR filters do, in order to achieve the same stop-band attenuation values. The 19 FIR band-pass filters required 601 taps to achieve a stop-band attenuation of over 70 dB. The frequency response of some of these filters are shown in figure 4.10, with the characteristics of all of the bands are listed in table 4.1. The sum of the frequency responses sums to unity in the frequency range of interest, namely from 180 Hz to 7500 Hz.

To generate the energy values for the individual bands, the signal was presented

Filter Bank No.	Lower Stopband Frequency (Hz)	Lower Passband Frequency (Hz)	Upper Passband Frequency (Hz)	Upper Stopband Frequency (Hz)
1	130	230	240	250
2	240	340	350	450
3	350	450	460	560
4	460	560	580	680
5	580	680	715	815
6	715	815	865	965
7	865	965	1030	1130
8	1030	1130	1215	1315
9	1215	1315	1425	1525
10	1425	1525	1670	1770
11	1670	1770	1940	2040
12	1940	2040	2260	2360
13	2260	2360	2640	2740
14	2640	2740	3075	3175
15	3075	3175	3625	3725
16	3675	3725	4300	4400
17	4300	4400	5200	5300
18	5270	5300	6300	6400
19	6300	6400	7600	7700

Table 4.1: Characteristics of Mel-Scale Filter Bands

to the band-pass filters and then the energy was computed for the 16 ms or 256 sample frames from the individual band-pass filtered signals. The set of 19 energy values per frame were then augmented by an adding another value containing the total energy for the frame, so that in all, 20 energy values constituted the mel-scale filter-band energy vector for a given input signal frame, or window.

Some considerations are in order before presenting the data to the neural network architectures for training and subsequent classification tests. First, the dynamic range for the input values can be rather large, given that the range of possible energy values can vary appreciably from one utterance to the next, between different portions of the same episode, and between different infants as well. In order to normalize these effects and to decrease the dynamic range of the inputs, so that the training or learning of features does not focus on the overly large input values, the input values of the feature sets were either scaled or normalized [Weiss and Kulikowski, 1991, Morgan and Scofield, 1991].

For the 1-second frames of mel-cepstrum coefficients, two alternatives were were investigated. In the first, the individual 1-second collection of computed mel-cepstrum coefficients were treated in two ways. First, these values were scaled by max(c_0), the maximum of the average energy value within the 1-second, or 125-vector, input data frame so that the maximum value contained in a given input frame would not exceed 1.0. In the second, these frames were also subject to normalization so that all the values within a given frame would lie between ± 1.0. This normalization was performed by first determining the range of values in a given 1-second input data frame, done by finding the largest value (max) and smallest value (min) in the frame, computing the "mean" or offset, and then subtracting the mean from all the values in the frame. This operation has the effect of shifting all input values in the frame to lie within the same positive and negative number. Dividing all the values in the frame by this number has the effect of normalizing values to lie between ± 1.0.

The individual 1-second collection of mel-scale filter-band energy values were treated in three ways. First, these values were scaled by the maximum value of a given 1-second, or 125-vector, input data frame, so that all the values would be at most 1.0. Then, dynamic range reduction was also achieved by taking the logarithm of the energy values, producing the second data set derived from mel-scale filter-band energies. Lastly, all the input frames had the mean of the maximum and minimum values contained within the 1-second frame subtracted and then normalized so that all values within the frame would lie between ± 1.0.

All of the parameter extraction, filter design, and signal processing of the cry utterances was performed on a Sparc 10+ using MATLAB. The following section briefly describes the neural network simulation software used to create, train, and test the various neural network architectures discussed in section 4.2.

4.5 Neural Network Simulation Software

In order to train and test the articles and paradigms that were mentioned in section 4.2, three different public domain neural network software simulators were used. The following subsections will briefly describe their features.

4.5.1 Aspirin/Migraines

The Aspirin/Migraines package is a system of tools developed at the MITRE Corporation in order to facilitate the generation, training, and testing of both small, trivial neural networks, and large, more complex neural networks [Leighton, 1992]. In this package, a neural network is specified according to a specific syntax which describes the network to be created in terms of input size, hidden layer size, and output layer size. The type of connections between layers, such as full or tessellated, the transfer or activation functions of the nodes, and the learning methods can also be specified. As well, if the transfer functions provided by the package do not provide sufficient resolution, or other types of activation functions with different characteristics than the one provided by the package are desired, they may be specified by the user.

The file that contains the description of the network in Aspirin format is then parsed and compiled to create a series of "C" language functions which are necessary to simulate the network. These generated functions are then compiled and linked to the application code using a set of user interface libraries referred to as Migraines.

The user interface provided by the latest version, release 6.0a, of this set of tools is text based. However, the Migraines interface can provide output in various formats which are supported by a number of popular plotting packages, so that network data can be visualized. As well, a number of formats are supported for

the input data file specification from simple ASCII to MATLAB-formatted data.

Unfortunately, the number of architectures and learning methods in this version of the software is quite limited. Using this package, only feedforward networks, with full and tessellated connections, and recurrent networks were simulated, both of which were trained using the generalized delta rule back propagation learning method.

4.5.2 Xerion

Xerion is a collection of simulators which is built using "C" language libraries developed at the University of Toronto by Geoffrey Hinton's research group [van Camp, 1993]. Each neural network architecture supported by this collection has its own individual simulator network, with the libraries providing the user with a consistent interface for interacting with the simulators, and for displaying the network properties using an X-based graphical user interface (GUI). As well, the collection of libraries provided by this package allows the researcher to code complex and experimental network definitions simply and quickly, and also allows the addition and generation of other architectures to be performed using the Xerion interface.

A neural network is specified in a file using a set of objects and creating and connecting these objects in the desired manner. A user can select from a number of different transfer function types and learning methods, which can either be selected prior to training using either X-based GUI menus and panels, command-line inputs, or through the use of command files. The input data files used for training and testing the network must be in plain ASCII; no other file formats are supported in the latest version of the software, version 3.1.

Although the standard Xerion distribution consists of eight different network simulators, some of which were used to test the suitability of certain architectures

for cry utterance classification, the results of only one of the more successful architectures will be presented in in section 4.6, namely cascade correlation neural networks which were trained using a conjugate-gradient descent learning method.

4.5.3 The Stuttgart Neural Network Simulator (SNNS)

The Stuttgart Neural Network Simulator is a comprehensive neural network simulator program which was developed by a team of researchers at the University of Stuttgart in Germany [Zell *et al.*, 1994]. This simulator package uses an X-based graphical user interface to interact with the user regarding the creation of networks, the loading and saving of network definitions, patterns, and configuration files, the training and testing of networks, and the selection and specification of node transfer functions, learning methods, and learning method parameters.

Although the graphical user interface provides the most elaborate and elegant means of interacting with the SNNS kernel, an automated batch process is provided for the purpose of training the networks as well. Version 3.2 of SNNS provides access to eleven neural network architectures and to a number of learning methods which are based on back propagation, and improvements to standard back propagation, such as QuickProp, which was developed in the interest of improving convergence times [Fahlman, 1988]. Version 3.3, which was released in November 1994, has added few more architectures to their collection, and has also provided some network pruning algorithms.

Despite the complete and comprehensive nature of this package, SNNS only supports input data files for training and testing in plain ASCII format.

The SNNS-provided architectures used in this research, whose results will be presented in this following section, were the cascade correlation using both back propagation and QuickProp, and time-delay neural networks (TDNNs).

4.6 Results

This section presents the results of the tests done using the feature sets and the neural network architectures presented in section 4.2. First, the experimentation procedures will be described, detailing what error measures will be used for the results which will be presented in the subsequent subsections. Then, the results obtained from input sets generated as described in section 4.3 will be presented in section 4.6.2 and section 4.6.3. The results will be discussed in section 4.7.

4.6.1 Experimentation Procedures and Error Measures

The various combinations of neural network architectures and input data sets were trained using a resampling strategy of 10-fold cross validation [Weiss and Kulikowski, 1991]. This involves splitting the data set into ten mutually exclusive sample sets of roughly the same size and using nine of these sets to train the network with the remaining set to be used as a test set. This process is repeated until all the sets have been used as the test set once, in order to perform an error rate estimation which is as close as possible to an unbiased estimator of the true classification rate. Since there are 195 input frames, or files, in the data set, at any one time approximately 90% of them, namely 175, 176, or 177 of these input frames, were used to train the network, and the remaining 10%, either 18, 19, or 20 input frames, were used to test the network once the training process was completed. The results from the testing process from these 10 data sets were then accumulated in order to determine the correct classification rates and error rates of a particular combination of an input data set with a neural network configuration.

The correct classification rate corresponds to the number of correct classification of test files, divided by the total number of test files. The error rate, on the other hand, is simply the correct classification rate subtracted from 1.

Using only one training and test set for all the experiments has the danger of being an especially biased test and, consequently, results of these types of experiments may not reflect the true performance if another data set consisting of similar data were to be used. These tests can lead to unrealistic and over optimistic results.

For the 10-fold cross-validation tests, since the data set consists of 54 anger cries, 16 fear cries, and 125 pain cries, the test sets contained anywhere from 4 to 6 anger cries, 1 or 2 fear cries, and 12 to 13 pain cries chosen randomly from the pool of 195 cries, respecting the proportion of these sets in the total data set.

The results will be presented in two separate subsections, one for the tests run using the mel-cepstrum coefficients as inputs, and the other for the tests run using the mel-scale filter-band coefficients. In these subsections, the results will be presented according to the architecture and the input features derived from either the mel-cepstrum coefficients or from the mel-scale filter-band energies, in a tabular format referred to as a confusion matrix, from which the optimal hidden layer size, learning rate, and momentum can be identified. As well, for the individual architectures, error rates will be tabulated as some parameters in the various networks were varied, such as hidden layer size for feedforward networks, input frame size and overlap for recurrent and time-delay neural networks, and as different training methods were used for cascade correlation networks. Lastly, the individual subsections will conclude by summarizing all the results in a table for comparative purposes.

Displaying the test results of the neural network outputs in a confusion matrix allows the identification of the different types of errors which occurred during the testing phase of the different neural networks architectures using different input data sets. A confusion matrix lists the correct classifications against the predicted classification for each output class, which for the this application corresponds to A, F, and P, for anger, fear, and pain cries, respectively. The number of correct classifications falls along the diagonal of this matrix.

Along with the confusion matrix, the correct classification rate for the anger, fear, and pain classes, as well as the total correct classification rate , resulting from 10-fold cross-validation training and testing performed on these networks will be given.

Since it is also useful to determine how these networks perform on a two-class classification problem, results from a "pain" and "no-pain" class perspective will be derived from the three-class classification results and presented along with the three-class classification results. Grouping the anger and fear classes together in the "no-pain" class implies that although correct classification of anger, fear, and pain are desired, it is of particular importance, especially in a clinical setting, that pain and non-pain utterances are not confused once the network has been trained. For the two-class confusion matrices, the "pain" and "no-pain" classes are denoted by P+ and P- respectively.

Consequently, when compiling the three-class results into a two-class grouping, the anger and fear outputs will be labelled as "no-pain" utterances. In this two-class grouping, anger utterances classified as fear, and vice-versa, will not be considered as being incorrectly classified since both anger and fear utterances fall into the "no-pain" class.

The two possible errors which occur in two-class classification problems are frequently given the names adopted from classification a medical context: *false positives* or *false negatives*. In this pain and no-pain context, false positives represent no-pain utterances which are incorrectly classified as pain utterances, and false negatives represent pain utterances incorrectly classified as no-pain utterances. Note that here the *true positives* is the number of pain files correctly classified as pain, and the *true negatives* are the number of no-pain files correctly classified as no-pain. From this, the following formal measures of classification performance can be defined [Weiss and Kulikowski, 1991]:

1. **Sensitivity**: which is the number of true positives divided by the total number

of actual pain files,

2. **Specificity**: which is the number of true negatives divided by the total number of actual no-pain files,
3. **Positive Predictive Value**: which is the number of true positives divided by the total number of files predicted as pain,
4. **Negative Predictive Value**: which is the number of true negatives divided by the total number of files predicted as no-pain,
5. **Accuracy**: the sum of the true positives and the true negatives divided by the total number of pain and no-pain files.

These measures are useful for identifying a neural network architecture and input data set which yields a high sensitivity, but which may have a poor specificity, in the event that numerous no-pain files are classified as pain.

4.6.2 Mel-Cepstrum Coefficient Input Data Set

This subsection presents the results for the input data sets derived from the mel-cepstrum coefficients, which were generated as described in section 4.3. The results are presented according to the architecture and the input data sets used to train and test the respective networks. The two input input data sets derived from the mel-cepstrum coefficients correspond to either the mel-cepstrum coefficients scaled by the maximum value in the input frame of data so that the input values do not exceed a maximum of 1.0, and the mel-cepstrum coefficient values with the mean removed and normalized to lie between ± 1.0.

First, the neural network results will be presented, followed by the results of the variation of some neural network parameters.

Number of hidden units: 45; Learning rate: $\alpha = 0.015$; Momentum: $\eta = 0.9$.

Predicted \ Actual	A	F	P
A	32	6	19
F	0	4	0
P	11	6	99
ϕ	11	0	7

$\text{Total}_{correct} = \frac{135}{195} = 69.23\%$

$A_{correct} = \frac{32}{54} = 59.25\%$

$F_{correct} = \frac{4}{16} = 25.00\%$

$P_{correct} = \frac{99}{125} = 79.20\%$

Predicted \ Actual	P-	P+
P-	42	19
P+	17	99
ϕ	11	7

Sensitivity $= \frac{99}{125} = 0.7920$

Specificity $= \frac{42}{70} = 0.6000$

Predictive value (+) $= \frac{99}{116} = 0.8534$

Predictive value (-) $= \frac{42}{61} = 0.6885$

Accuracy $= \frac{141}{195} = 0.7231$

Table 4.2: Results for Fully Connected Feedforward Neural Network using Mel-Cepstrum Inputs Scaled to a Maximum Value of 1.0

Number of hidden units: 45; Learning rate: $\alpha = 0.015$; Momentum: $\eta = 0.9$.

Predicted \ Actual	A	F	P
A	41	5	5
F	0	7	0
P	5	4	115
ϕ	8	0	5

$\text{Total}_{correct} = \frac{163}{195} = 83.59\%$

$A_{correct} = \frac{41}{54} = 75.93\%$

$F_{correct} = \frac{7}{16} = 43.75\%$

$P_{correct} = \frac{115}{125} = 92.00\%$

Predicted \ Actual	P-	P+
P-	53	5
P+	9	115
ϕ	8	5

Sensitivity $= \frac{115}{125} = 0.9200$

Specificity $= \frac{53}{70} = 0.7571$

Predictive value (+) $= \frac{115}{124} = 0.9274$

Predictive value (-) $= \frac{53}{58} = 0.9138$

Accuracy $= \frac{168}{195} = 0.8615$

Table 4.3: Results for Fully Connected Feedforward Neural Network using Mel-Cepstrum Inputs with Mean Removed and Normalized to Lie Between ± 1.0

Neural Network Results

As mentioned at the start of section 4.6, the results are displayed in confusion matrices and the entries of these matrices represents the sum of the test phases of the ten 10-fold cross-validation tests with separate matrices for both three-class and two-class confusion matrices. To the right of the respective matrices are error measures derived from these matrices. This is done for the configuration indicated at the top of the tables, and for the neural network architectures and input data sets indicated in the caption of the table.

For the feedforward neural network result tables, an additional row has been

Number of hidden units: 24 with [11 × 10] tessellation and 11 X-overlap and 3 Y-overlap.
Learning rate: $\alpha = 0.015$; **Momentum:** $\eta = 0.95$.

Predicted \ Actual	A	F	P
A	32	6	20
F	2	2	0
P	20	8	98
ϕ	0	0	7

$\text{Total}_{correct} = \frac{132}{195} = 67.69\%$
$A_{correct} = \frac{32}{54} = 59.25\%$
$F_{correct} = \frac{2}{16} = 12.50\%$
$P_{correct} = \frac{98}{125} = 78.40\%$

Predicted \ Actual	P-	P+
P-	42	20
P+	28	98
ϕ	0	7

Sensitivity $= \frac{98}{125} = 0.7840$
Specificity $= \frac{42}{70} = 0.6000$
Predictive value (+) $= \frac{98}{126} = 0.7778$
Predictive value (-) $= \frac{42}{62} = 0.6774$
Accuracy $= \frac{140}{177} = 0.7910$

Table 4.4: Results for a Feedforward Neural Network with Tessellated Connections using Mel-Cepstrum Inputs Scaled to a Maximum value of 1.0

Number of hidden units: 24 with [11 × 10] tessellation and 11 X-overlap and 3 Y-overlap.
Learning rate: $\alpha = 0.015$; **Momentum:** $\eta = 0.95$.

Predicted \ Actual	A	F	P
A	34	2	10
F	0	2	0
P	13	12	105
ϕ	7	0	10

$\text{Total}_{correct} = \frac{141}{195} = 73.31\%$
$A_{correct} = \frac{34}{54} = 62.96\%$
$F_{correct} = \frac{2}{16} = 12.50\%$
$P_{correct} = \frac{105}{125} = 84.00\%$

Predicted \ Actual	P-	P+
P-	38	10
P+	25	105
ϕ	7	10

Sensitivity $= \frac{105}{125} = 0.8400$
Specificity $= \frac{38}{70} = 0.5429$
Predictive value (+) $= \frac{105}{130} = 0.8077$
Predictive value (-) $= \frac{38}{48} = 0.7917$
Accuracy $= \frac{143}{195} = 0.7333$

Table 4.5: Results for a Feedforward Neural Network with Tessellated Connections using Mel-Cepstrum Inputs with Mean Removed and Normalized to Lie Between ±1.0

included in the respective confusion matrices. This row, labeled as "predicted ϕ", indicates the number of test files whose output was undefined, that is, which had an output value which did not single out one of the three output classes, when an input frame of extracted parameters was presented at the inputs during the testing phase. This problem arises when more than one output values saturates at the positive output value, +1.

The best results of the fully connected feedforward neural networks, which use the scaled and normalized mel-cepstrum input data sets are given in table 4.2 and

Inputs size: [75 × 11] with an overlap of 50 vectors per input frame.
Number of hidden units: 36 with 3 delay units per node.
Learning rate: $\alpha = 0.1$; **Momentum:** $\eta = 0.5$.

		Actual		
		A	F	P
Predicted	A	20	0	20
	F	0	2	0
	P	25	6	104
	ϕ	9	8	2

$\text{Total}_{correct} = \frac{125}{195} = 64.61\%$
$A_{correct} = \frac{20}{54} = 37.03\%$
$F_{correct} = \frac{2}{16} = 12.50\%$
$P_{correct} = \frac{104}{125} = 83.20\%$

		Actual	
		P-	P+
Predicted	P-	22	20
	P+	31	104
	ϕ	17	2

Sensitivity $= \frac{104}{125} = 0.8320$
Specificity $= \frac{22}{70} = 0.3143$
Predictive value (+) $= \frac{104}{135} = 0.7704$
Predictive value (-) $= \frac{22}{42} = 0.5238$
Accuracy $= \frac{126}{195} = 0.6461$

Table 4.6: Results for a Fully Connected Recurrent Neural Network using Mel-Cepstrum Inputs Scaled to a Maximum Value of 1.0

table 4.3 respectively. Best results for the feedforward networks with tessellated connections are given in table 4.4 and table 4.5 for the scaled and normalized mel-cepstrum input data sets respectively. For the feedforward architecture, the training method which yielded the best correct classification rate employed a pattern-by-pattern updating of the weights, which is similar to the weight update method presented in section 4.2.1. The presentation sequence of the input patterns during the training phase was randomized in order to both speed the convergence time of the network as well as improving the generalization capabilities of the trained network.

Table 4.6 and table 4.7 list the best results for the recurrent neural networks, which, for both the scaled and normalized input parameter sets derived from the mel-cepstrum coefficients, have an input frame size of 75 mel-cepstrum vectors, with subsequent input frames overlapping by 50 vectors. Consequently, one entire 1-second, or 125-vector, input frame for a given utterance is traversed in three 75-vector frames. For both the scaled and the normalized input parameter sets, the optimal configuration for the recurrent neural network consisted of 36 hidden units with each node in the hidden and output units having two delays per node. This results in the two previous outputs of these nodes being fed back as input to

Inputs size: [75 × 11] with an overlap of 50 vectors per input frame.
Number of hidden units: 36 with 3 delay units per node.
Learning rate: $\alpha = 0.1$; **Momentum:** $\eta = 0.5$.

	Actual A	Actual F	Actual P
Predicted A	29	0	15
Predicted F	0	2	0
Predicted P	13	0	106
Predicted ϕ	12	14	4

$\text{Total}_{correct} = \frac{137}{195} = 70.26\%$
$A_{correct} = \frac{29}{54} = 53.70\%$
$F_{correct} = \frac{2}{16} = 12.50\%$
$P_{correct} = \frac{106}{125} = 84.80\%$

	Actual P-	Actual P+
Predicted P-	31	15
Predicted P+	13	106
Predicted ϕ	26	4

Sensitivity $= \frac{106}{125} = 0.8480$
Specificity $= \frac{31}{70} = 0.4429$
Predictive value (+) $= \frac{106}{119} = 0.8908$
Predictive value (-) $= \frac{31}{46} = 0.6739$
Accuracy $= \frac{137}{195} = 0.7026$

Table 4.7: Results for a Fully Connected Recurrent Neural Network using Mel-Cepstrum Inputs with Mean Removed and Normalized to Lie Between ±1.0

Inputs size: [105 × 11]; **Overlap:** [104 × 11].
Number of hidden units: [21 × 5]; **Learning rate:** $\alpha = 0.015$.

	Actual A	Actual F	Actual P
Predicted A	32	16	53
Predicted F	12	0	0
Predicted P	10	0	72

$\text{Total}_{correct} = \frac{105}{195} = 61.02\%$
$A_{correct} = \frac{32}{54} = 61.11\%$
$F_{correct} = \frac{0}{16} = 0\%$
$P_{correct} = \frac{72}{125} = 57.60\%$

	Actual P-	Actual P+
Predicted P-	60	53
Predicted P+	10	72

Sensitivity $= \frac{72}{125} = 0.5760$
Specificity $= \frac{60}{70} = 0.8571$
Predictive value (+) $= \frac{72}{82} = 0.8780$
Predictive value (-) $= \frac{60}{113} = 0.5310$
Accuracy $= \frac{132}{195} = 0.6792$

Table 4.8: Results for a Time-Delay Neural Network using Mel-Cepstrum Inputs Scaled to a Maximum Value of 1.0

that same unit one time instant later. During the testing of this particular network architecture, the weights of the neural networks were updated after the presentation of one complete 125-vector input frame, that is after three 75-vector frames with the subsequent 125-vector training frames presented randomly to the network. For the recurrent, time-delay, and cascade correlation neural networks, once the network had converged and the test patterns were presented at the inputs, the winning output or "class" was determined as being the output having the largest value.

The results for the time-delay neural networks are given in table 4.8 and table 4.9

Inputs size: [105 × 11]; **Overlap:** [104 × 11].
Number of hidden units: [21 × 5]; **Learning rate:** $\alpha = 0.015$

		Actual		
		A	F	P
Predicted	A	35	16	37
	F	10	0	0
	P	9	0	88

$\text{Total}_{correct} = \frac{123}{195} = 63.07\%$

$A_{correct} = \frac{35}{54} = 64.81\%$

$F_{correct} = \frac{0}{16} = 0\%$

$P_{correct} = \frac{88}{125} = 70.40\%$

	Actual	
	P-	P+
Predicted P-	61	37
Predicted P+	9	88

Sensitivity $= \frac{88}{125} = 0.7040$

Specificity $= \frac{61}{70} = 0.8714$

Predictive value (+) $= \frac{88}{97} = 0.9072$

Predictive value (-) $= \frac{61}{98} = 0.6224$

Accuracy $= \frac{149}{195} = 0.7641$

Table 4.9: Results for a Time-Delay Neural Network using Mel-Cepstrum Inputs with Mean Removed and Normalized to Lie Between ±1.0

for the scaled and normalized mel-cepstrum inputs, respectively. For both input parameters derived from the mel-cepstrum coefficients, the TDNN configuration which yielded the best results was for an input frame size of 105 mel-cepstrum vectors, with an overlap of 104 vectors per frame, and with a hidden layer consisting of 5 hidden units integrating the activations of 21 input frames, corresponding to 20 delayed activation values plus the undelayed activations of the last input frame. Consequently, the entire 125-vector input frame is presented after 125 time instants; 105 to fill the first input frame and 20 more to fill the 20 activation delay values in the hidden layer. The network was trained using a variant of back propagation specifically formulated for TDNNs, with the weights being updated after the presentation of all the input patterns in the training set [Waibel *et al.*, 1987]. The 125-vector input patterns were presented to the network in random order during the training process.

Table 4.10 and table 4.11 show the best results obtained using the cascade correlation network, for the scaled and normalized mel-cepstrum coefficient input data sets, respectively. For both of these input data sets, the optimal results were achieved using a network that was trained using a conjugate gradient descent training algorithm with the patterns presented in a random fashion and with the weights updated after all the patterns in the training set were presented.

Number of hidden units created: 28.
Training method: Congugate Gradient.

Predicted \ Actual	A	F	P
A	22	7	48
F	9	0	27
P	23	9	50

$\text{Total}_{correct} = \frac{72}{195} = 36.92\%$

$A_{correct} = \frac{22}{54} = 40.74\%$

$F_{correct} = \frac{0}{16} = 0\%$

$P_{correct} = \frac{50}{125} = 40.00\%$

Predicted \ Actual	P-	P+
P-	38	75
P+	32	50

Sensitivity $= \frac{50}{125} = 0.4000$

Specificity $= \frac{38}{70} = 0.5429$

Predictive value (+) $= \frac{50}{82} = 0.6097$

Predictive value (-) $= \frac{38}{113} = 0.3363$

Accuracy $= \frac{88}{195} = 0.4513$

Table 4.10: Results for a Cascade Correlation Neural Network using Mel-Cepstrum Inputs Scaled to a Maximum Value of 1.0

Number of hidden units created: 24.
Training method: Congugate Gradient.

Predicted \ Actual	A	F	P
A	24	8	41
F	24	0	21
P	6	8	63

$\text{Total}_{correct} = \frac{78}{195} = 40.00\%$

$A_{correct} = \frac{24}{54} = 44.44\%$

$F_{correct} = \frac{0}{16} = 0\%$

$P_{correct} = \frac{63}{125} = 50.40\%$

Predicted \ Actual	P-	P+
P-	56	62
P+	14	63

Sensitivity $= \frac{63}{125} = 0.5040$

Specificity $= \frac{56}{70} = 0.8000$

Predictive value (+) $= \frac{63}{77} = 0.8182$

Predictive value (-) $= \frac{56}{118} = 0.4746$

Accuracy $= \frac{119}{195} = 0.6103$

Table 4.11: Results for a Cascade Correlation Recurrent Neural Network using Mel-Cepstrum Inputs with Mean Removed and Normalized to Lie Between ± 1.0

Neural Network Parameter Variations

In order to compare how the number hidden layer nodes affects the error rate for the two input data sets derived from the mel-cepstrum coefficients, tables of hidden layer size and error rates for the fully connected feedforward neural network is shown in figure 4.12(a) and figure 4.12(b) for the scaled inputs and for the normalized inputs respectively. The neural network configuration with two hidden layers had 125 and 17 nodes in the first and second hidden layers, respectively.

Table 4.13(a) and table 4.13(b) show the same for the feedforward networks

Learning Rate: $\alpha = 0.015$; Momentum: $\eta = 0.95$.

Hidden Nodes	Error Rate
17	$\frac{94}{195} = 0.4821$
32	$\frac{75}{195} = 0.3846$
45	$\frac{60}{195} = 0.3077$
74	$\frac{96}{195} = 0.4923$
2-layer 125 & 17	$\frac{95}{195} = 0.4872$

(a) Scaled Input Values

Hidden Nodes	Error Rate
17	$\frac{75}{195} = 0.3846$
32	$\frac{69}{195} = 0.3538$
45	$\frac{32}{195} = 0.1641$
74	$\frac{44}{195} = 0.2256$
2-layer 125 & 17	$\frac{65}{195} = 0.3333$

(b) Normalized Input Values

Table 4.12: Hidden Layer Size and Error Rates for Fully Connected Feed-forward Neural Networks using Mel-Cepstrum Coefficient Inputs

with tessellated connections. The input nodes are organized in a two dimensional array of 11 × 125, corresponding to 125 vectors of 11 mel-cepstrum coefficients. The tessellation configurations, and the number of overlapping nodes between in adjacent groupings of input nodes, for the respective hidden layer size, are indicated in the tables. The tessellation column shows the number input nodes which were grouped together, and the overlap column indicates the number of overlapping nodes between adjacent groups of nodes. For example, the first row of table 4.13(a) indicates that a grouping, or tiling, of 20 mel-cepstrum coefficients vectors ([11×20]) with adjacent "tiles" having 15 overlapping mel-cepstrum vectors, generates a hidden layer of 22 hidden units.

Table 4.14 and table 4.15 illustrate how the input frame size, corresponding to the time granularity or resolution of the given input parameters, affects the error rate, and how the hidden layer size affects the error rate for the optimal frame size of 75-vectors for both the scaled and normalized mel-cepstrum inputs, respectively. In table 4.14(a) and table 4.15(a), the input frame column indicates the number of mel-cepstrum coefficient vectors which comprised the input frame size. The overlap column displays the number of vectors from the current input frame which would included in the next input frame. The number of delay nodes in both output nodes and for the number of hidden nodes indicated under the hidden layer column are also indicated in these tables.

Learning Rate: $\alpha = 0.015$; **Momentum:** $\eta = 0.95$.

Hidden Nodes	Tessellation	Overlap	Error Rate
[1 × 22]	[11 × 20]	11X-15Y	$\frac{75}{195} = 0.3846$
[1 × 23]	[11 × 25]	11X-10Y	$\frac{66}{195} = 0.3384$
[1 × 24]	[11 × 25]	11X-5Y	$\frac{63}{195} = 0.3231$
[1 × 61]	[11 × 5]	11X-3Y	$\frac{92}{195} = 0.4722$

(a) Scaled Input Values

Hidden Nodes	Tessellation	Overlap	Error Rate
[1 × 22]	[11 × 20]	11X-15Y	$\frac{77}{195} = 0.3949$
[1 × 23]	[11 × 25]	11X-10Y	$\frac{60}{195} = 0.3077$
[1 × 24]	[11 × 25]	11X-5Y	$\frac{54}{195} = 0.2769$
[1 × 61]	[11 × 5]	11X-3Y	$\frac{90}{195} = 0.4615$

(b) Normalized Input Values

Table 4.13: Hidden Layer Size and Error Rates for Feedforward Neural Networks with Tessellated Connections using Mel-Cepstrum Coefficient Inputs

Learning Rate: $\alpha = 0.1$; **Momentum:** $\eta = 0.5$.

Input Frame	Overlap	Delay Nodes	Hidden Nodes	Error Rate
10 × 11	5	4	18	$\frac{89}{195} = 0.4564$
25 × 11	0	4	28	$\frac{95}{195} = 0.4872$
75 × 11	50	2	36	$\frac{69}{195} = 0.3538$

(a) Input Frame Size and Error Rates

Hidden Nodes	Error Rate
27	$\frac{90}{195} = 0.4615$
36	$\frac{69}{195} = 0.3538$
45	$\frac{79}{195} = 0.4051$

(b) Hidden Layer Size and Error Rate for a [75 × 11] Input Frame Size

Table 4.14: Parameter Variations and Error Rates for Recurrent Neural Network using Mel-Cepstrum Coefficients Scaled to a Maximum Value of 1.0

Learning Rate: $\alpha = 0.1$; **Momentum:** $\eta = 0.5$.

Input Frame	Overlap	Delay Nodes	Hidden Nodes	Error Rate
10 × 11	5	4	18	$\frac{78}{195} = 0.4000$
25 × 11	0	4	28	$\frac{72}{195} = 0.3692$
75 × 11	50	2	36	$\frac{58}{195} = 0.2974$

(a) Input Frame Size and Error Rates

Hidden Nodes	Error Rate
27	$\frac{75}{195} = 0.3846$
36	$\frac{58}{195} = 0.2974$
45	$\frac{70}{195} = 0.3590$

(b) Hidden Layer Size and Error Rate for a [75 × 11] Input Frame Size

Table 4.15: Parameter Variations and Error Rates for Recurrent Neural Network using Mel-Cepstrum Coefficients With Mean Removed and Normalized to Lie Between ±1.0

Learning Rate: $\alpha = 0.015$.

Input Frame	Overlap	Hidden Nodes	Error Rate
10 × 11	9 × 11	(1) 56 × 8 (2) 60 × 4	$\frac{120}{195} = 0.6154$
105 × 11	104 × 11	21 × 5	$\frac{90}{195} = 0.4615$

(a) Scaled Input Values

Input Frame	Overlap	Hidden Nodes	Error Rate
10 × 11	9 × 11	(1) 56 × 8 (2) 60 × 4	$\frac{105}{195} = 0.5385$
105 × 11	104 × 11	21 × 5	$\frac{76}{195} = 0.3897$

(b) Normalized Input Values

Table 4.16: Network Variations and Error Rates for the Time-Delay Neural Network using Parameters Derived from the Mel-Cepstrum Coefficients

Table 4.16 shows the results of another time-delay neural network configuration using a smaller input frame size, which corresponds to a finer time resolution or representation of the input data set, than the 75 × 11 which yielded the lowest error rate for this architecture. Since a smaller delay width for this neural network corresponds to a larger number of input frames to integrate over, requiring a larger number of delay units for a given input frame size, two hidden layers were used for this network configuration. The first hidden layer, with a width of 8 units, integrates activity from 56 input frames, and the second hidden layer, with a width of 4 units, integrates the activation of the first hidden layer over 60 time units. Since the training times of these networks are significantly longer when compared to those of the other architectures investigated for these experiments, only the two configurations listed in the table were tested.

Table 4.17 displays the results of cascade correlation networks trained using different methods for the two input data sets derived from the mel-cepstrum coefficients. The table lists the learning rate and momentum parameters for the training methods, if applicable, and also lists the number of hidden units created by the respective methods, followed by the error rates obtained during testing.

Training Method	Parameters		Hidden Nodes Created	Error Rate
	α	η		
BackProp	0.1	0.7	37	$\frac{127}{195} = 0.6513$
QuickProp	0.0001	1.9	4	$\frac{130}{195} = 0.6667$
ConjGrad	–	–	28	$\frac{123}{195} = 0.6308$

(a) Scaled Input Values

Training Method	Parameters		Hidden Nodes Created	Error Rate
	α	η		
BackProp	0.1	0.7	41	$\frac{122}{195} = 0.6256$
QuickProp	0.0001	1.9	15	$\frac{129}{195} = 0.6615$
ConjGrad	–	–	24	$\frac{117}{195} = 0.6000$

(b) Normalized Input Values

Table 4.17: Training Methods and Error Rates for the Cascade Correlation Neural Network using Mel-Cepstrum Coefficient Derived Inputs

4.6.3 Mel-Scale Filter-Band Energy Input Data Set

This subsection presents the results for the mel-scale filter-band energy-based input data set, which was generated as described in section 4.3. As was the case for the presentation of the results for the input data sets derived from the mel-cepstrum coefficients, the results in this subsection will be presented according to the architecture and the input parameters used in the training and testing of the respective architectures. The input data sets derived from the 19 mel-scale filter-band energies are the mel-scale filter-band energy values scaled by the maximum value of the input data frame so that the maximum value of the input data frame does not exceed 1.0, the logarithm of the mel-scaled filter-band energies, and the logarithm of the mel-scale filter-band with the mean of a given input frame removed and normalized so that the values in the input frame lie between ± 1.0.

As was the case for the presentation of results from the input data sets derived from the mel-cepstrum coefficients presented in preceding subsection, first, the neural network results will be presented, followed by the error rates resulting from the variation of some of the neural network parameters, such as hidden layer size and, input frame size and overlap, if applicable.

Number of hidden units: 25.
Learning rate: $\alpha = 0.015$; Momentum: $\eta = 0.95$.

		Actual		
		A	F	P
Predicted	A	41	5	0
	F	0	2	0
	P	13	5	109
	ϕ	0	4	16

$\text{Total}_{correct} = \frac{152}{195} = 77.95\%$
$A_{correct} = \frac{41}{54} = 75.92\%$
$F_{correct} = \frac{2}{16} = 12.50\%$
$P_{correct} = \frac{109}{125} = 87.20\%$

		Actual	
		P-	P+
Predicted	P-	48	0
	P+	18	109
	ϕ	4	16

Sensitivity $= \frac{109}{125} = 0.8720$
Specificity $= \frac{48}{70} = 0.6857$
Predictive value (+) $= \frac{109}{127} = 0.8583$
Predictive value (-) $= \frac{48}{48} = 1.0000$
Accuracy $= \frac{157}{195} = 0.8051$

Table 4.18: Results for a Fully Connected Feedforward Neural Network using Mel Filter-Band Inputs Scaled to a Maximum Value of 1.0

Number of hidden units: 25.
Learning rate: $\alpha = 0.015$; Momentum: $\eta = 0.95$.

		Actual		
		A	F	P
Predicted	A	48	7	10
	F	0	2	0
	P	6	7	105
	ϕ	0	0	7

$\text{Total}_{correct} = \frac{155}{195} = 79.49\%$
$A_{correct} = \frac{48}{54} = 88.89\%$
$F_{correct} = \frac{2}{16} = 12.50\%$
$P_{correct} = \frac{105}{125} = 84.00\%$

		Actual	
		P-	P+
Predicted	P-	57	10
	P+	13	105
	ϕ	0	7

Sensitivity $= \frac{105}{125} = 0.8400$
Specificity $= \frac{57}{70} = 0.8143$
Predictive value (+) $= \frac{105}{118} = 0.8898$
Predictive value (-) $= \frac{57}{67} = 0.8507$
Accuracy $= \frac{162}{195} = 0.8308$

Table 4.19: Results for a Fully Connected Feedforward Neural Network using the Log of the Mel Filter-Band Inputs

Neural Network Results

To reiterate what was mentioned at the start of section 4.6, the values displayed in the confusion matrices are the sum of the test phase of the ten 10-fold cross validation tests, be it for the three-class confusion matrices or for the two-class confusion matrices displayed. The error measures indicated in the tables are derived from the respective three or two-class confusion matrices.

Table 4.18, table 4.19, and table 4.20, show the results of the feedforward neural network, with full connections between adjacent layers, for the scaled, log, and

Number of hidden units: 25.
Learning rate: $\alpha = 0.015$; **Momentum:** $\eta = 0.95$.

Predicted \ Actual	A	F	P
A	43	6	6
F	0	2	0
P	11	8	109
ϕ	0	0	10

$\text{Total}_{correct} = \frac{154}{195} = 78.97\%$

$\text{A}_{correct} = \frac{43}{54} = 79.63\%$

$\text{F}_{correct} = \frac{2}{16} = 12.50\%$

$\text{P}_{correct} = \frac{109}{125} = 87.20\%$

Predicted \ Actual	P-	P+
P-	51	6
P+	19	109
ϕ	0	10

Sensitivity $= \frac{109}{125} = 0.8720$

Specificity $= \frac{51}{70} = 0.7286$

Predictive value (+) $= \frac{109}{128} = 0.8516$

Predictive value (-) $= \frac{51}{57} = 0.8947$

Accuracy $= \frac{160}{195} = 0.8205$

Table 4.20: Results for a Fully Connected Feedforward Neural Network using the Log of the Mel Filter-Band Inputs with Mean Removed and Normalized to Lie Between ± 1.0

Number of hidden units: 22 with [20 × 20] tessellation and 20 X-overlap and 15 Y-overlap.
Learning rate: $\alpha = 0.015$; **Momentum:** $\eta = 0.95$.

Predicted \ Actual	A	F	P
A	41	8	11
F	0	0	0
P	13	8	97
ϕ	0	0	17

$\text{Total}_{correct} = \frac{138}{195} = 70.77\%$

$\text{A}_{correct} = \frac{41}{54} = 75.92\%$

$\text{F}_{correct} = \frac{0}{16} = 0\%$

$\text{P}_{correct} = \frac{97}{125} = 77.60\%$

Predicted \ Actual	P-	P+
P-	49	11
P+	21	97
ϕ	0	17

Sensitivity $= \frac{97}{125} = 0.7760$

Specificity $= \frac{49}{70} = 0.7000$

Predictive value (+) $= \frac{97}{118} = 0.8220$

Predictive value (-) $= \frac{49}{60} = 0.8167$

Accuracy $= \frac{157}{195} = 0.8051$

Table 4.21: Results for Feedforward Neural Network with Tessellated Connections using Mel Filter-Band Inputs Scaled to a Maximum Value of 1.0

normalized log of the mel-scale filter-band energies. The results for the corresponding input data sets trained on a feedforward neural network with tessellated connections are shown in table 4.21, table 4.22, and table 4.23, respectively. For all these feedforward networks, the weights of the network were updated following the presentation of an input frame, in a manner similar to that described in section 4.2.1. Subsequent input frames were presented randomly to the network. As well, all feedforward nets used the same learning rate and momentum parameters, as indicated in the tables.

Number of hidden units: 22 with [20 × 20] tessellation and 20 X-overlap and 15 Y-overlap.
Learning rate: $\alpha = 0.015$; **Momentum:** $\eta = 0.95$.

Predicted \ Actual	A	F	P
A	42	8	27
F	0	0	0
P	3	8	98
ϕ	9	0	0

$\text{Total}_{correct} = \frac{140}{195} = 71.79\%$
$A_{correct} = \frac{42}{54} = 77.78\%$
$F_{correct} = \frac{0}{16} = 0\%$
$P_{correct} = \frac{98}{125} = 78.40\%$

Predicted \ Actual	P-	P+
P-	50	27
P+	11	98
ϕ	9	0

Sensitivity $= \frac{98}{125} = 0.7840$
Specificity $= \frac{50}{70} = 0.7143$
Predictive value (+) $= \frac{98}{109} = 0.8991$
Predictive value (-) $= \frac{50}{77} = 0.6494$
Accuracy $= \frac{148}{195} = 0.7590$

Table 4.22: Results for a Feedforward Neural Network with Tessellated Connections using the Log of the Mel Filter-Band Inputs

Number of hidden units: 22 with [20 × 20] tessellation and 20 X-overlap and 15 Y-overlap.
Learning rate: $\alpha = 0.015$; **Momentum:** $\eta = 0.95$.

Predicted \ Actual	A	F	P
A	34	8	9
F	0	0	0
P	10	8	107
ϕ	10	0	9

$\text{Total}_{correct} = \frac{141}{195} = 72.31\%$
$A_{correct} = \frac{34}{54} = 62.96\%$
$F_{correct} = \frac{0}{16} = 0\%$
$P_{correct} = \frac{107}{125} = 85.60\%$

Predicted \ Actual	P-	P+
P-	42	9
P+	18	107
ϕ	10	9

Sensitivity $= \frac{107}{125} = 0.8560$
Specificity $= \frac{42}{70} = 0.6000$
Predictive value (+) $= \frac{107}{125} = 0.8560$
Predictive value (-) $= \frac{42}{51} = 0.8253$
Accuracy $= \frac{149}{195} = 0.7641$

Table 4.23: Results for a Fully Connected Feedforward Neural Network using the Log of the Mel Filter-Band Inputs with Mean Removed and Normalized to Lie Between ±1.0

As was the case for the feedforward nets in the preceding subsection, an additional row has been included in the confusion matrices for these networks. This row, labeled ϕ, indicates the number of test files whose output was undefined, that is, which had an output value which did not correspond to one of the three output classes during testing.

Next, the results of the recurrent neural networks for the three input parameter sets derived from the mel-scale filter-band energies are given in table 4.24, table 4.25, and table 4.26. Here, all three input parameter sets yielded the highest

Inputs size: [75 × 20] with an overlap of 50 vectors per input frame.
Number of hidden units: 18 with 3 delay units per node.
Learning rate: $\alpha = 0.1$; **Momentum:** $\eta = 0.5$.

Predicted \ Actual	A	F	P
A	26	8	11
F	0	0	0
P	28	8	104
ϕ	0	0	10

$\text{Total}_{correct} = \frac{130}{195} = 66.67\%$
$A_{correct} = \frac{26}{54} = 48.15\%$
$F_{correct} = \frac{0}{16} = 0\%$
$P_{correct} = \frac{104}{125} = 83.20\%$

Predicted \ Actual	P-	P+
P-	34	11
P+	36	104
ϕ	0	10

Sensitivity = $\frac{104}{125} = 0.8320$
Specificity = $\frac{34}{70} = 0.4857$
Predictive value (+) = $\frac{104}{140} = 0.7429$
Predictive value (-) = $\frac{34}{45} = 0.7556$
Accuracy = $\frac{138}{195} = 0.7077$

Table 4.24: Results for a Recurrent Neural Network using Mel Filter-Band Inputs Scaled to a Maximum Value of 1.0

Inputs size: [75 × 20] with an overlap of 50 vectors per input frame.
Number of hidden units: 18 with 3 delay units per node.
Learning rate: $\alpha = 0.1$; **Momentum:** $\eta = 0.5$.

Predicted \ Actual	A	F	P
A	20	2	20
F	0	0	0
P	34	8	104
ϕ	0	6	0

$\text{Total}_{correct} = \frac{124}{195} = 63.59\%$
$A_{correct} = \frac{20}{54} = 37.04\%$
$F_{correct} = \frac{0}{16} = 0\%$
$P_{correct} = \frac{104}{125} = 83.20\%$

Predicted \ Actual	P-	P+
P-	22	20
P+	42	104
ϕ	0	7

Sensitivity = $\frac{104}{125} = 0.8320$
Specificity = $\frac{22}{70} = 0.3143$
Predictive value (+) = $\frac{104}{146} = 0.7123$
Predictive value (-) = $\frac{22}{42} = 0.5238$
Accuracy = $\frac{124}{195} = 0.6359$

Table 4.25: Results for a Recurrent Neural Network using the Log of the Mel Filter-Band Inputs

correct classification rates for this architecture using the same network configuration. This recurrent neural network had input frames consisting of 75 vectors, with subsequent input frames overlapping by 50 vectors, so that the entire 125-vector input pattern was visited after three 75-vector frames. These networks also had 18 hidden layer nodes and three delay units for each hidden layer and output layer node. The weights for this network were updated following the presentation of the entire 125-vector input frame, corresponding to the presentation of one complete input pattern. Subsequent 125-vector input patterns were presented randomly to

Inputs size: [75 × 20] with an overlap of 50 vectors per input frame.
Number of hidden units: 18 with 3 delay units per node.
Learning rate: $\alpha = 0.1$; **Momentum:** $\eta = 0.5$.

Predicted \ Actual	A	F	P
A	31	6	19
F	0	0	0
P	23	10	106
ϕ	0	0	0

$\text{Total}_{correct} = \frac{137}{195} = 70.26\%$
$A_{correct} = \frac{31}{54} = 57.41\%$
$F_{correct} = \frac{0}{16} = 0\%$
$P_{correct} = \frac{106}{125} = 84.80\%$

Predicted \ Actual	P-	P+
P-	37	19
P+	33	106
ϕ	0	00

Sensitivity = $\frac{106}{125} = 0.8480$
Specificity = $\frac{37}{70} = 0.5286$
Predictive value (+) = $\frac{106}{139} = 0.7626$
Predictive value (-) = $\frac{37}{56} = 0.6607$
Accuracy = $\frac{143}{195} = 0.7333$

Table 4.26: Results for a Recurrent Neural Network using the Log of the Mel Filter-Band Inputs with Mean Removed and Normalized to Lie Between ±1.0

Inputs size: [105 × 20]; **Overlap:** [104 × 20].
Number of hidden units: [21 × 10]; **Learning rate:** $\alpha = 0.2$.

Predicted \ Actual	A	F	P
A	26	16	39
F	0	0	0
P	29	0	86

$\text{Total}_{correct} = \frac{111}{195} = 56.92\%$
$A_{correct} = \frac{25}{54} = 46.30\%$
$F_{correct} = \frac{0}{16} = 0\%$
$P_{correct} = \frac{111}{125} = 68.80\%$

Predicted \ Actual	P-	P+
P-	41	39
P+	29	86

Sensitivity = $\frac{86}{125} = 0.6880$
Specificity = $\frac{41}{70} = 0.5857$
Predictive value (+) = $\frac{86}{115} = 0.7478$
Predictive value (-) = $\frac{41}{80} = 0.5125$
Accuracy = $\frac{127}{195} = 0.6513$

Table 4.27: Results for a Time-Delay Neural Network using Mel Filter-Band Inputs Scaled to a Maximum Value of 1.0

the network. As was the case for the recurrent, time-delay, and cascade correlation neural networks used with the mel-cepstrum coefficient derived parameter sets, and explained in subsection 4.6.2, the winning class during the testing phase, was determined as being the output with the largest value.

The results for the time-delay neural networks are given in table 4.27, table 4.28, and table 4.29 for the input parameters derived from the mel-scale filter-band inputs. For the three input data sets, the TDNN configuration which yielded the highest correct classification rate had an input frame size of 105 vectors, with sub-

Inputs size: [105 × 20]; **Overlap:** [104 × 20].
Number of hidden units: [21 × 10]; **Learning rate:** $\alpha = 0.2$.

		Actual		
		A	F	P
Predicted	A	14	0	23
	F	0	0	0
	P	40	16	102

$\text{Total}_{correct} = \frac{116}{195} = 59.49\%$
$A_{correct} = \frac{14}{54} = 25.93\%$
$F_{correct} = \frac{0}{16} = 0\%$
$P_{correct} = \frac{102}{125} = 81.60\%$

		Actual	
		P-	P+
Predicted	P-	14	23
	P+	56	102

Sensitivity = $\frac{102}{125} = 0.8160$
Specificity = $\frac{14}{70} = 0.2000$
Predictive value (+) = $\frac{102}{158} = 0.6456$
Predictive value (-) = $\frac{14}{37} = 0.3784$
Accuracy = $\frac{116}{195} = 0.5949$

Table 4.28: Results for a Time-Delay Neural Network using the Log of the Mel Filter-Band Inputs

Inputs size: [105 × 20]; **Overlap:** [104 × 20].
Number of hidden units: [21 × 10]; **Learning rate:** $\alpha = 0.2$.

		Actual		
		A	F	P
Predicted	A	10	0	15
	F	0	0	0
	P	44	16	110

$\text{Total}_{correct} = \frac{120}{195} = 61.54\%$
$A_{correct} = \frac{10}{54} = 18.52\%$
$F_{correct} = \frac{0}{16} = 0\%$
$P_{correct} = \frac{110}{125} = 88.00\%$

		Actual	
		P-	P+
Predicted	P-	10	15
	P+	60	110

Sensitivity = $\frac{110}{125} = 0.8800$
Specificity = $\frac{10}{70} = 0.1429$
Predictive value (+) = $\frac{110}{170} = 0.6471$
Predictive value (-) = $\frac{10}{25} = 0.4000$
Accuracy = $\frac{120}{195} = 0.6154$

Table 4.29: Results for a Time-Delay Neural Network using the Log of the Mel Filter-Band Inputs with Mean Removed and Normalized to Lie Between ±1.0

sequent input frames overlapping by 104 vectors. For this TDNN, the hidden layer consisted of 10 hidden units integrating the activations of 21 input frames, corresponding to 20 delayed activation values plus the activations of the last input frame, allowing the total input delay length of 125 vectors to be considered after 125 time instants. The network was trained using a variant of back propagation specifically formulated for TDNNs, with the weights being updated after the presentation of all the input patterns in the training set [Waibel *et al.*, 1987]. The input patterns were presented to the network in random order during the training process.

Table 4.39(a), table 4.39(b), and table 4.39(c) show the best results obtained using

Number of hidden units created: 11.
Training method: Back Propagation; Learning rate: $\alpha = 0.1$**; Momentum:** $\eta = 0.7$.

Predicted \ Actual	A	F	P
A	9	6	30
F	0	0	0
P	45	10	95

$\text{Total}_{correct} = \frac{104}{195} = 53.33\%$

$A_{correct} = \frac{9}{54} = 16.67\%$

$F_{correct} = \frac{0}{16} = 0\%$

$P_{correct} = \frac{95}{125} = 76.00\%$

Predicted \ Actual	P-	P+
P-	15	30
P+	55	95

Sensitivity $= \frac{95}{125} = 0.7600$

Specificity $= \frac{15}{70} = 0.2143$

Predictive value (+) $= \frac{95}{150} = 0.6333$

Predictive value (-) $= \frac{15}{45} = 0.3333$

Accuracy $= \frac{110}{195} = 0.5641$

Table 4.30: Results for a Cascade Correlation Neural Network using Mel Filter-Band Inputs Scaled to a Maximum Value of 1.0

Number of hidden units created: 9.
Training method: Back Propagation; Learning rate: $\alpha = 0.1$**; Momentum:** $\eta = 0.7$.

Predicted \ Actual	A	F	P
A	6	7	24
F	0	0	0
P	48	9	101

$\text{Total}_{correct} = \frac{107}{195} = 54.87\%$

$A_{correct} = \frac{6}{54} = 11.11\%$

$F_{correct} = \frac{0}{16} = 0\%$

$P_{correct} = \frac{101}{125} = 80.80\%$

Predicted \ Actual	P-	P+
P-	13	24
P+	57	101

Sensitivity $= \frac{101}{125} = 0.8080$

Specificity $= \frac{13}{70} = 0.1857$

Predictive value (+) $= \frac{101}{158} = 0.6392$

Predictive value (-) $= \frac{13}{37} = 0.3513$

Accuracy $= \frac{116}{195} = 0.5949$

Table 4.31: Results for a Cascade Correlation Neural Network using the Log of the Mel Filter-Band Inputs

Number of hidden units created: 4.
Training method: Back Propagation; Learning rate: $\alpha = 0.1$**; Momentum:** $\eta = 0.7$.

Predicted \ Actual	A	F	P
A	10	8	19
F	0	0	0
P	44	8	106

$\text{Total}_{correct} = \frac{116}{195} = 59.49\%$

$A_{correct} = \frac{10}{54} = 18.52\%$

$F_{correct} = \frac{0}{16} = 0\%$

$P_{correct} = \frac{106}{125} = 84.80\%$

Predicted \ Actual	P-	P+
P-	18	19
P+	52	106

Sensitivity $= \frac{106}{125} = 0.8480$

Specificity $= \frac{18}{70} = 0.2571$

Predictive value (+) $= \frac{106}{158} = 0.6709$

Predictive value (-) $= \frac{18}{37} = 0.4865$

Accuracy $= \frac{124}{195} = 0.6359$

Table 4.32: Results for a Cascade Correlation Neural Network using the Log of the Mel Filter-Band Inputs with Mean Removed and Normalized to Lie Between ± 1.0

Learning Rate: $\alpha = 0.015$; **Momentum:** $\eta = 0.95$.

Hidden Nodes	Error Rate
25	$\frac{43}{195} = 0.2205$
55	$\frac{58}{195} = 0.2974$
85	$\frac{56}{195} = 0.2872$
2-layer 120 & 25	$\frac{63}{195} = 0.3231$

(a) Scaled Input Values

Hidden Nodes	Error Rate
25	$\frac{40}{195} = 0.2051$
55	$\frac{56}{195} = 0.2872$
85	$\frac{62}{195} = 0.3179$
2-layer 120 & 25	$\frac{52}{195} = 0.2667$

(b) Log Input Values

Hidden Nodes	Error Rate
25	$\frac{41}{195} = 0.2103$
55	$\frac{48}{195} = 0.2462$
85	$\frac{58}{195} = 0.2974$
2-layer 120 & 25	$\frac{50}{195} = 0.2564$

(c) Normalized Log Input Values

Table 4.33: Hidden Layer Size and Error Rates for Fully Connected Feed-forward Neural Networks using Mel-Scale Filter-Band Inputs

the cascade correlation network, for this set of input parameters. For all of these input parameter sets derived from the mel-scale filter-band energy values, the optimal results were achieved using a network that was trained using the back propagation training algorithm with the patterns presented in a random fashion, and with the weights updated after all the patterns in the training set were presented to the network.

Neural Network Parameter Variations

In order to compare how the hidden layer size affects the error rate for the three input parameter sets derived from the mel-scale filter-band energy values, tables showing the hidden layer size versus error rates for the fully connected feedforward neural network are shown in figure 4.33(a), figure 4.33(b), and figure 4.33(c) for the scaled, log, and normalized log input data sets respectively. Table 4.34(a), table 4.34(b), and table 4.34(c) show hidden layer size and error rates for the feed-forward networks with tessellated connections for the scaled, log, and normalized log of the mel-scale filter-band energies respectively.

The input nodes are organized in a two dimensional array of 20×125, corresponding to 125 vectors of 20 mel-scale filter-band energy values. The tessellation configurations, and the number of overlapping nodes between adjacent groupings of input nodes, for the respective hidden layer size are indicated in the tables. The

Learning Rate: $\alpha = 0.015$; **Momentum:** $\eta = 0.95$.

Hidden Nodes	Tessellation	Overlap	Error Rate
[1 × 22]	[20 × 20]	20X-15Y	$\frac{57}{195} = 0.2923$
[2 × 22]	[10 × 20]	0X-15Y	$\frac{63}{195} = 0.3231$
[1 × 61]	[20 × 5]	20X-3Y	$\frac{68}{195} = 0.3487$
[5 × 5]	[4 × 25]	0X-0Y	$\frac{65}{195} = 0.3333$

(a) Scaled Input Values

Hidden Nodes	Tessellation	Overlap	Error Rate
[1 × 22]	[11 × 20]	11X-15Y	$\frac{55}{195} = 0.2821$
[1 × 23]	[11 × 25]	11X-10Y	$\frac{58}{195} = 0.2974$
[1 × 24]	[11 × 25]	11X-5Y	$\frac{66}{195} = 0.3385$
[1 × 61]	[11 × 5]	11X-3Y	$\frac{70}{195} = 0.3590$

(b) Log Input Values

Hidden Nodes	Tessellation	Overlap	Error Rate
[1 × 22]	[20 × 20]	20X-15Y	$\frac{54}{195} = 0.2769$
[2 × 22]	[10 × 20]	0X-15Y	$\frac{57}{195} = 0.2923$
[1 × 61]	[20 × 5]	20X-3Y	$\frac{62}{195} = 0.3179$
[5 × 5]	[4 × 25]	0X-0Y	$\frac{64}{195} = 0.3282$

(c) Normalized Log Input Values

Table 4.34: Hidden Layer Size and Error Rates for Feedforward Neural Networks with Tessellated Connections using Mel-Scale Filter-Band Inputs

tessellation column shows the number input nodes which were grouped together, and the overlap column indicates the number of overlapping nodes between adjacent groups of nodes. For example, the last row of table 4.34(a) indicates that a grouping, or tiling, of 4 of the 20 mel-scale filter-band energy values in a vector over 25 vectors ([4 × 25]) with no overlap occurring between adjacent "tiles" generates a hidden layer of 25 hidden units organized in a two-dimensional array of 5 × 5 nodes, which effectively "cover" the 20 × 125 input nodes.

Table 4.35, table 4.36, and table 4.37 illustrate how the input frame size, corresponding to the time granularity or resolution of the given input data sets, affects the error rate, and how the hidden layer size affects the error rate for the optimal frame size for both the scaled, log, and normalized log of the mel-scale filter-band energies, respectively.

Learning Rate: $\alpha = 0.1$; **Momentum:** $\eta = 0.5$.

Input Frame	Overlap	Delay Nodes	Hidden Nodes	Error Rate
10 × 20	5	4	15	$\frac{88}{195} = 0.4513$
25 × 20	0	4	18	$\frac{80}{195} = 0.4103$
75 × 20	50	2	18	$\frac{65}{195} = 0.3333$

(a) Input Frame Size and Error Rates

Hidden Nodes	Error Rate
12	$\frac{85}{195} = 0.4359$
18	$\frac{65}{195} = 0.3333$
36	$\frac{81}{195} = 0.4153$

(b) Hidden Layer Size and Error Rate for an Input Size of [75 × 20]

Table 4.35: Parameter Variations and Error Rates for Recurrent Neural Network using Mel-Scale Filter Band Energy Values Scaled to a Maximum Value of 1.0

Learning Rate: $\alpha = 0.1$; **Momentum:** $\eta = 0.5$.

Input Frame	Overlap	Delay Nodes	Hidden Nodes	Error Rate
10 × 20	5	4	15	$\frac{91}{195} = 0.4667$
25 × 20	0	4	18	$\frac{87}{195} = 0.4461$
75 × 20	50	2	18	$\frac{71}{195} = 0.3641$

(a) Input Frame Size and Error Rates

Hidden Nodes	Error Rate
12	$\frac{90}{195} = 0.4615$
18	$\frac{71}{195} = 0.3641$
36	$\frac{87}{195} = 0.4461$

(b) Hidden Layer Size and Error Rate for an Input Size of [75 × 20]

Table 4.36: Parameter Variations and Error Rates for Recurrent Neural Network using the Logarithm of the Mel-Scale Filter-Band Energies

Learning Rate: $\alpha = 0.1$; **Momentum:** $\eta = 0.5$.

Input Frame	Overlap	Delay Nodes	Hidden Nodes	Error Rate
10 × 20	5	4	15	$\frac{79}{195} = 0.4051$
25 × 20	0	4	18	$\frac{66}{195} = 0.3385$
75 × 20	50	2	18	$\frac{58}{195} = 0.2974$

(a) Input Frame Size and Error Rates

Hidden Nodes	Error Rate
12	$\frac{65}{195} = 0.3333$
18	$\frac{58}{195} = 0.2974$
36	$\frac{62}{195} = 0.3179$

(b) Hidden Layer Size and Error Rate for an Input Size of [75 × 20]

Table 4.37: Parameter Variations and Error Rates for Recurrent Neural Network using the Logarithm of the Mel-Scale Filter-Band Energies With Mean Removed and Normalized to Lie Between ± 1.0

In table 4.38, the result of another time-delay neural network configuration using a smaller input frame, which corresponds to a finer time resolution of the input parameters, than that which yielded the highest correct classification rate for this architecture. Since a smaller input delay width for this neural network corresponds to a larger number of input frames to integrate over, thus requiring a larger number of delays per hidden layer node, two hidden layers were used for

Learning Rate: $\alpha = 0.015$.

Input Frame	Overlap	Hidden Nodes	Error Rate
10 × 20	9 × 20	(1) 56 × 10 (2) 60 × 5	$\frac{89}{195} = 0.4564$
105 × 20	104 × 20	21 × 10	$\frac{84}{195} = 0.4308$

(a) Scaled Input Values

Input Frame	Overlap	Hidden Nodes	Error Rate
10 × 20	9 × 20	(1) 56 × 10 (2) 60 × 5	$\frac{87}{195} = 0.4462$
105 × 20	104 × 20	21 × 10	$\frac{79}{195} = 0.4051$

(b) Log Input Values

Input Frame	Overlap	Hidden Nodes	Error Rate
10 × 20	9 × 20	(1) 56 × 10 (2) 60 × 5	$\frac{79}{195} = 0.4051$
105 × 20	104 × 20	21 × 10	$\frac{75}{195} = 0.3846$

(c) Normalized Log Input Values

Table 4.38: Network Variations and Error Rates for the Time-Delay Neural Network using Parameters Derived From the Mel-Scale Filter-Band Energy Values

this network configuration. The first hidden layer of this TDNN, with a width of 8 units, integrates activity from 56 input frames, and the second hidden layer, with a width of 4 units, integrates the activation of the first hidden layer over 60 time instants. Since the training times of these networks are significantly longer when compared to those of the other architectures, only the two TDNN configurations listed in the table were tested.

Table 4.39 displays the results of cascade correlation networks trained using different weight update methods for the three input parameter sets derived from the mel-scale filter-band energy values. The table lists the learning rate and momentum parameters for the training methods, if applicable, and also lists the number of hidden units created by the respective methods, followed by their respective error rates.

Training Method	Parameters		Hidden Nodes Created	Error Rate
	α	η		
BackProp	0.1	0.7	11	$\frac{91}{195} = 0.4667$
QuickProp	0.0001	1.9	4	$\frac{95}{195} = 0.4872$
ConjGrad	–	–	18	$\frac{115}{195} = 0.5897$

(a) Scaled Input Values

Training Method	Parameters		Hidden Nodes Created	Error Rate
	α	η		
BackProp	0.1	0.7	9	$\frac{88}{195} = 0.4513$
QuickProp	0.0001	1.9	8	$\frac{93}{195} = 0.4769$
ConjGrad	–	–	22	$\frac{99}{195} = 0.5077$

(b) Log Input Values

Training Method	Parameters		Hidden Nodes Created	Error Rate
	α	η		
BackProp	0.1	0.7	4	$\frac{79}{195} = 0.4051$
QuickProp	0.0001	1.9	2	$\frac{87}{195} = 0.4462$
ConjGrad	–	–	18	$\frac{90}{195} = 0.4615$

(c) Normalized Log Input Values

Table 4.39: Training Method and Error Rates for the Cascade Correlation Neural Network using Mel-Cepstrum Coefficient Derived Parameters

4.7 Discussion

This section discusses the results of the neural network training and testing presented in section 4.6. First, the results obtained from the various input parameter sets, derived from both the mel-cepstrum coefficients and from the mel-scale filter-band energy values, which were trained and tested on the neural network architectures presented in section 4.2 will be discussed. Then, the best results obtained for the various architectures will be discussed, followed by a discussion of the results of the neural network configurations and input frame size variations for the different input data sets. The final subsection compares the results obtained in the experiments presented in the previous section to that of other work done by other researchers.

4.7.1 Neural Network Architectures

The results presented in section 4.6.2 and section 4.6.3 have been summarized in table 4.40 and table 4.41 for the input parameter sets derived from the mel-cepstrum coefficients and mel-scale filter-band energy values respectively. In these tables, the columns correspond to the neural network architectures and represent, going from left to right, the best results for the fully connected feedforward ANN (FF), the feedforward ANN with tessellated connections (FT), the recurrent neural network (RNN), the time-delay neural network (TDNN), and the cascade correlation neural network (CC). The rows represent the three-class classification rates and two-class error measures indicated in the rows of the table, which were defined in subsection 4.6.1.

The three-class classification results in these two tables are given in decimal form and not as a percentage as was done when these values were presented in the individual tables of section 4.6.2 and section 4.6.3, in the interest of remaining consistent with the format of the results for the two-class classification rates.

Mel-Cepstrum Coefficient-Derived Input Parameters

Table 4.40 summarizes the results for the two input data sets derived from the mel-cepstrum coefficients, with the results for the mel-cepstrum input frames scaled to a maximum value of 1.0 given in table 4.40(a), and the results for the mel-cepstrum coefficient input frames which have had their mean removed and normalized to lie between values of ± 1.0, given in table 4.40(b).

It can immediately be observed that the results for the normalized mel-cepstrum inputs are, for the most part, better than for the corresponding scaled mel-cepstrum input values. Looking at the three-class classification results for the anger, fear, and pain classes, the majority of neural network architectures have better classification rates for pain outputs than for either anger and fear, with the latter class consistently

	FF	FT	RNN	TDNN	CC
$Total_{correct}$	0.6923	0.6769	0.6461	0.6102	0.3692
$A_{correct}$	0.5925	0.5925	0.3703	0.6111	0.4074
$F_{correct}$	0.2500	0.1250	0.1250	0.0000	0.0000
$P_{correct}$	0.7920	0.7840	0.8320	0.5760	0.4000
Sensitivity	0.7920	0.7840	0.8320	0.5760	0.4000
Specificity	0.6000	0.6000	0.3143	0.8571	0.5429
Pred. Val.(+)	0.8534	0.7778	0.7704	0.8780	0.6097
Pred. Val.(-)	0.6885	0.6774	0.5238	0.5310	0.3363
Accuracy	0.7231	0.7910	0.6461	0.6792	0.4513

(a) Scaled Input Values

	FF	FT	RNN	TDNN	CC
$Total_{correct}$	0.8359	0.7331	0.7026	0.6307	0.4000
$A_{correct}$	0.7953	0.6296	0.5370	0.6481	0.4444
$F_{correct}$	0.4375	0.1250	0.1250	0.0000	0.0000
$P_{correct}$	0.9200	0.8400	0.8480	0.7040	0.5040
Sensitivity	0.9200	0.8400	0.8480	0.7040	0.5040
Specificity	0.7571	0.5429	0.4429	0.8714	0.8000
Pred. Val.(+)	0.9274	0.8077	0.8908	0.9072	0.8182
Pred. Val.(-)	0.9138	0.7917	0.6739	0.6224	0.4746
Accuracy	0.8615	0.7333	0.7026	0.7641	0.6103

(b) Normalized Input Values

Table 4.40: Result Summary for Neural Networks using Mel-Cepstrum Coefficient Inputs

having classification rates below 0.5. This may be due to the small number of fear recordings present in the data set, when compared to either pain or anger classes.

For the scaled mel-cepstrum inputs whose results are summarized in table 4.40(a), the best classification rate is achieved for a fully connected feedforward neural network with a hidden layer size of 45 nodes. Next, the feedforward neural network with tessellated connections does only slightly worse than the fully connected feedforward, with the decreased fear and pain classification rates forcing the total correct classification rate down. The recurrent neural network generates the third-highest correct classification rate, and also features the highest correct classification rate for pain cries for this input data set. The correct classification of anger cries drops substantially for this neural network architecture however. For the time-delay neural network, the correct classification rate drops from that of the recurrent neural network, due to the substantial drop in the number of pain cries,

which is slightly offset by an increase in the number of correctly classified anger cries. The correct classification rate drops substantially for the cascade correlation neural network, with the majority of pain cries not being correctly classified.

Looking at the two-class results for the scaled input data set listed in the bottom half of table 4.40(a), the highest accuracy is achieved for the feedforward neural network with tessellated connections. The neural network architecture with the highest sensitivity is the recurrent neural network, which implies that for this particular input data set, this architecture has the best rate of correctly classifying pain cries. This same architecture, however, has a low specificity, implying that the method has a tendency to classify anger and fear cries as pain cries, which is not a desirable characteristic. In this light, the best tradeoff between a good sensitivity and a good specificity is achieved for the fully connected feedforward network, even if this architecture has a considerable number of undefined outputs.

For the three-class results of the normalized mel-cepstrum inputs, whose results are shown in table 4.40(b), the highest correct classification rate is achieved by the fully connected feedforward neural network with a hidden layer size consisting of 45 nodes. Note that the correct classification rates for all three output classes are appreciably larger than for the scaled mel-cepstrum inputs. The feedforward ANN with tessellated connections has the next highest correct classification rate for this input data set, with drops in all the classification rates of all three output classes contributing to the decline in the total correct classification rate. Although the classification rate for the recurrent neural network is lower than that of the feedforward net with tessellated connections, the recurrent neural network has a higher correct classification rate for pain utterances, a result similar to that achieved for the scaled input data set. The time-delay and cascade correlation results for the normalized input data set do not differ significantly from those for the scaled input data set. However, the normalized inputs have a higher number of correctly classified pain cries than the scaled input data set does.

For the two-class classification results obtained from the normalized mel-cepstrum input data set listed in bottom half of table 4.40(b), the architecture that has the highest pain classification sensitivity is the fully connected feedforward neural network. Both the feedforward network with tessellated connections and the recurrent neural network have good sensitivity values, but their specificities are rather low, implying that these two methods have a high incidence of classifying non-pain cries as pain cries. The time-delay neural network has a high specificity, but although this network can classify non-pain cries correctly, it cannot do the same for pain cries. A similar observation can be made for the cascade correlation network, except that this network has a lower accuracy, corresponding to a lower correct classification rate for pain cries than the TDNN does. The fully connected feedforward network has the highest accuracy, and the best specificity and sensitivity combination with high values for both the positive and negative predictive values. The latter values imply that for recordings which are classified or predicted as being pain or no-pain, over 90% of these classifications correspond to actual pain or no-pain utterances.

Mel-Scale Filter-Band Energy-Derived Input Parameters

The results from the three parameter sets derived from the mel-scale filter-band energy values are summarized in table 4.41. In this table, the results for the mel-scale filter-band energies data set which have either been scaled to a maximum of 1.0 are given in table 4.41(a), the results for the data set corresponding to the logarithm of the mel-scale filter-band energies are given in table 4.41(b), and the input data set where the mean of the logarithm of the mel-scale filter-band energies has been removed and for which the values been normalized to lie between ± 1.0 is given in table 4.41(c).

Examining all the results in these tables, it can be observed that for all the architectures, except for the fully connected feedforward neural network, the input

	FF	FT	RNN	TDNN	CC
$Total_{correct}$	0.7795	0.7077	0.6667	0.5692	0.5333
$A_{correct}$	0.7592	0.7592	0.4815	0.4630	0.1667
$F_{correct}$	0.1250	0.0000	0.0000	0.0000	0.0000
$P_{correct}$	0.8720	0.7760	0.8320	0.6880	0.7600
Sensitivity	0.8720	0.7760	0.8320	0.6880	0.7600
Specificity	0.6857	0.7000	0.4857	0.5857	0.2143
Pred. Val.(+)	0.8583	0.8220	0.7429	0.7478	0.6333
Pred. Val.(-)	1.0000	0.8167	0.7556	0.5125	0.3333
Accuracy	0.8051	0.8051	0.7077	0.6513	0.5641

(a) Scaled Input Values

	FF	FT	RNN	TDNN	CC
$Total_{correct}$	0.7949	0.7179	0.6359	0.5949	0.5487
$A_{correct}$	0.8889	0.7778	0.3740	0.2593	0.1111
$F_{correct}$	0.1250	0.0000	0.0000	0.0000	0.0000
$P_{correct}$	0.8400	0.7840	0.8320	0.8160	0.8080
Sensitivity	0.8400	0.7840	0.8320	0.8160	0.8080
Specificity	0.8143	0.7143	0.3143	0.2000	0.1857
Pred. Val.(+)	0.8898	0.8991	0.7123	0.6456	0.6392
Pred. Val.(-)	0.8507	0.6494	0.5238	0.3784	0.3513
Accuracy	0.8308	0.7590	0.6359	0.5949	0.5949

(b) Log Input Values

	FF	FT	RNN	TDNN	CC
$Total_{correct}$	0.7897	0.7231	0.7026	0.6154	0.5949
$A_{correct}$	0.7963	0.6296	0.5471	0.1852	0.1111
$F_{correct}$	0.1250	0.0000	0.0000	0.0000	0.0000
$P_{correct}$	0.8720	0.8560	0.8480	0.8800	0.8080
Sensitivity	0.8720	0.8560	0.8480	0.8800	0.8080
Specificity	0.7286	0.6000	0.5286	0.1429	0.2000
Pred. Val.(+)	0.8516	0.8560	0.7626	0.6471	0.6456
Pred. Val.(-)	0.8947	0.8253	0.6607	0.4000	0.3784
Accuracy	0.8205	0.7641	0.7333	0.6154	0.5949

(c) Normalized Log Input Values

Table 4.41: Result Summary for Neural Networks using Mel-Cepstrum Coefficient Inputs

data set which has the highest correct classification rates for both the two and three-class groupings are obtained for the normalized log of the mel-scale filter-band energy values. The next highest correct classification rates are obtained using the scaled mel-scale filter-band values, with the log of the mel-scale filter-band energies yielding the lowest correct classification rates, for the majority of neural network architectures listed in the tables.

For all three input data sets derived from the mel-scale filter-band energy values, the fully connected feedforward neural network gives the best results, with the one using the log of the mel-scale filter-band energies having the highest total correct classification and accuracy rates of the three input data sets. This results from a higher number of correctly classified anger cries for the log input data sets, which offsets the smaller number of correctly classified pain cries.

Looking at the results of the input data sets derived from the mel-scale filter-band energy values individually, it can be noticed from the three-class results of the scaled values, as tabulated in table 4.41(a), that the fully connected feedforward neural network with a hidden layer size of 25 nodes has the highest correct classification rate of all the architectures. The feedforward ANN with tessellated connections has the same correct classification rate for anger cries, but the drop in the overall correct classification rate is due to the drop in the number of correctly classified pain utterances and from the absence of any correctly classified fear utterances.

For the recurrent neural network, the number of correctly classified anger cries drops significantly, but the number of correctly classified pain cries lies between that of the fully connected feedforward neural network and that of the feedforward neural network with tessellated connections. As was the case for the TDNN using the mel-cepstrum derived input data sets, the time-delay neural network produces a low classification rate for anger cries. The rate for pain cries drops from that of the feedforward network with tessellated connections for the TDNN. The cascade correlation network has a very poor classification rate for anger cries, but has a larger number of correctly classified pain cries than the time-delay neural network.

For the two-class results listed in the bottom half of table 4.41(a), the method that has the highest sensitivity is the fully connected feedforward ANN. This architecture and data set also features a reasonable specificity, implying that this method can correctly classify pain from pain utterances, and also does reasonably well at correctly identifying no-pain utterances. Note that this architecture has

a perfect negative predictive value, corresponding to the result that all predicted no-pain utterances corresponded to no-pain utterances. Both feedforward neural network architectures, either with full or tessellated connections, have the same accuracy, but overall, however, the numbers of the fully connected network are better, even if the one with tessellated connections has a higher specificity.

For the three-class results of the log mel-scale filter-band energy input data set, as shown in table 4.41(b), the fully connected feedforward neural network has the best correct classification rates of all the neural network architectures. Note that for this input data set, the feedforward network with tessellated connections has the lowest classification rate for pain utterances, even if it comes second in terms of the overall correct classification rate. Also, for the recurrent, time-delay, and cascade correlation neural networks, the classification rate falls dramatically from that of the feedforward networks, even if the correct classification rate for pain falls only slightly, when compared to the feedforward nets. The only method that can correctly classify any fear files is the fully connected feedforward neural network; all other architectures fail to classify fear utterances.

In the bottom half of table 4.41(b), the two-class results, for the log of the mel-scale filter-band energy data set, indicate that the fully connected feedforward ANN gives the highest accuracy, sensitivity, and specificity when compared to all the other architectures which use the same input data set. Note that the recurrent, time-delay, and cascade correlation neural networks all have sensitivity values above 0.8, but the specificity of these nets are extremely low. This implies that although these nets can correctly classify pain utterances, they do poorly at correctly classifying non-pain utterances. Also from the low positive and negative prediction values indicated in the table for these three ANNs, it can be said that these networks have a larger number of misclassifications than the feedforward neural networks with full and tessellated connections.

In table 4.41(c), it is readily seen that once again the fully connected feedforward

neural network with 25 hidden layer nodes yields the highest total correct classification rate for the normalized log mel-scale filter-band energy input values. Also, this architecture has the highest classification rate for anger and fear cries for this input data set, as well as having the second highest correct classification rate for pain cries of all the architectures that use this input data set. The classification rates for the other architectures decrease as one scans the table from left to right, with the sole exception being the correct classification rate for pain utterances for the time-delay neural network. Although the classification rate for anger cries drops by more than half from that of the recurrent neural network to that of the time-delay and cascade correlation neural networks, the correct classification rate for the pain cries does not drop below 0.8 for all the architectures with this input data set. Also, as was the case for the other two input data sets derived from the mel-scale filter-band energies, only the fully connected feedforward neural network classifies some of the fear utterances correctly.

In the bottom half of table 4.41(c), the best accuracy, and the best specificity and sensitivity combination, is obtained for the feedforward neural network with full connections between the nodes of adjacent layers. Note that the predictive value rates are also relatively high for this ANN, implying that this architecture, as is the case when the other two input data sets are used, performs few misclassification errors. Although the other architectures all have sensitivity values above 0.8, the specificity values are very poor for the time-delay and the cascade correlation neural network, implying that these methods have the tendency of classifying non-pain utterances as pain; an observation made for the log of the mel-scale filter-band energy values input data set.

Comparison of the Input Data Sets

Examining the results obtained from the five data sets used to train and test the various neural network architectures, the best correct classification rates were achieved

for the fully connected feedforward neural network using the mel-cepstrum coefficients which had their mean removed and which were normalized to lie between values of ± 1, as the input data set. This combination of network architecture and input data set has the highest correct classification rates for two of the three utterance classes, namely fear and pain cries, across all the architectures and all the input data sets, with 92% of pain utterances in the test set being correctly classified, and a total correct classification rate of 83.59%. From the two-class classification rates, this neural network architecture and input data set yields a high sensitivity with a good specificity, implying that there are few misclassifications performed by this method, and, as well, that over 91% of utterances that are classified as either pain or no-pain actually correspond to pain and no-pain utterances. Consequently, this architecture and input data set also have the highest of all the two-class accuracy values.

Comparing the total correct classification rates of the input data sets, the normalized mel-cepstrum coefficient values yield higher values for all the architectures except for the cascade correlation networks, which in any event, yield poor results in all cases.

For all the input data sets, the best results were achieved when a fully connected feedforward neural network was used. Often, a feedforward neural network with tessellated connections would provide the next best classification rates, which at best would be slightly less, but never equalling those of the fully connected network. Hence, it would seem that an organization of the neural network connections which attempts to model the receptive fields of the brain may work well for some applications, but fail to match the performance of full connections for the purposes of cry classification. Full connections between nodes in adjacent layers seems to better model the input-output characteristics than tessellated connections do, as is identified by the larger correct classification rates using this node connection methodology. Tessellated connections seem to miss some of the correlations between the input features which seem to be captured by the full connections.

Next, the issue of the usefulness of time information for the purposes of correct classification will be addressed. Looking at the results obtained for the recurrent neural networks and for the time-delay neural networks, it is noted that the total correct classification rate for the recurrent neural networks are consistently better than those of the time-delay neural networks. The time-delay neural network was formulated to accurately capture the specific acoustic sequences of a given phoneme for speech, whereas the recurrent neural network was formulated as a means of handling time-dependent input in general, with no provisions made to provide shift invariance or to capture fine features in the input sequence.

The stricter encoding of the sequence of acoustic features present in cry utterances which is inherent in the structure of TDNNs, would not seem to benefit the classification of infant anger, fear, and pain cries. Although the correct classification rates for these two architectures are less than those of the feedforward networks having either full or tessellated connections, the recurrent neural network would seem to make better use of time information than does the time-delay neural network. This observation is intuitive if one thinks for the type of information contained in cry utterances and speech signals.

Speech is defined in terms of phonemes, and a specific sequence of acoustic events denotes a specific phoneme. Some phonemes can contain similar acoustic events, but it is the sequence of these events, or the occurrence of these events, followed or preceded by other events, which distinguishes phonemes from each other. Hence, time-delay neural networks are an effective architecture for capturing this information in an input frame of parameters derived from the speech signal.

For neonates, however, vocal tract shape is affected by a number of physiological or psychological effects, which may be reflexive and not under the direct volitional control of the infant [Zeskind, 1985]. Consequently, the *occurrence* of specific acoustic events in cries of the same class would seem to be more important than the *sequence* in which these events occur.

That being said, it is understandable that recurrent networks fare better than time-delay neural networks, since the former encodes time information on a more general level than the sequential information encoded by a TDNN. This is further supported by the observation that larger, or coarse, time frame sizes give better results than smaller, or finer, time frame sizes, as will be further elaborated in the following subsection.

As well, since the occurrence of certain acoustic events would appear to be more relevant than the sequence with which these events occur, feedforward neural networks, with fully connected nodes between adjacent layers, yield better results than their time-dependent counterparts. Fully connected feedforward neural networks are capable of computing more complex relations between the inputs and outputs than what is possible when sparser connections are used, thus yielding better results for this particular application.

Looking at the results obtained from the cascade correlation neural network, it would seem that this particular paradigm is not suitable for the classification of infant anger, fear, and pain cry utterances. Although the idea behind using a learning method that grows its own hidden layer, thus taking the guesswork out of determining the optimal number of hidden layer nodes which is required for a given network and application, is indeed appealing, the resulting correct classification rates obtained are rather disappointing.

For both the mel-cepstrum coefficients and for the mel-scale filter-band energy input data sets, the best classification rates are obtained when the normalized input data sets are used. Normalization ensures that all values in a given input data frame will lie between ± 1, so that all the input data frames will have the same dynamic range. Scaling the values, or dividing by the maximum value of a given input frame only ensures that the largest value in the frame will be 1, making no claims on the range of values of the input data frames.

Comparing the results between the best of the mel-cepstrum coefficient derived

input data sets and the best of the input data sets derived from the mel-scale filter-band energy values, both of which are for the normalized input data sets, one can notice that the results for the normalized mel-cepstrum coefficients yield better results for all but the the cascade correlation neural network. The encoding of the relevant spectral characteristics of an input signal window for the purposes of classification, would seem to be better captured by the mel-cepstrum coefficients than by the normalized log mel-scale filter-band energy values. This observation has been also made by researchers in the speech domain [Davis and Mermelstein, 1980], so it would appear that for the purposes of classifying and discriminating between infant anger, fear, and pain cries, mel-cepstrum coefficients yield better results than filter-band coefficients as well.

Overall Observations and Comments

Looking at the classification results for all the architectures and all the input data sets overall and in general, a number of patterns emerge. First, the correct classification rate of anger cries seldom exceeds that for pain cries. In fact, the correct classification rate for anger cries exceeds that of pain cries in only three cases: for the time-delay and cascade neural networks which use the scaled mel-cepstrum coefficient input data set, and for the fully connected feedforward neural network which uses the log of the mel-scale filter-band energy as the input data set.

This may be partly due to the fact that there were more than twice as many pain cries in the data set than there were anger cries. Consequently, the neural networks may have been better able to generalize on pain cries than with anger cries, especially for the input sets derived from the mel-scale filter-band energies. For these particular data sets, the correct classification rate of pain cries never went below 0.68.

The correct classification rate for fear cries is consistently very poor, never exceeding the 0.45 mark. Again, this is most likely due to the small number of

available fear utterances in the data set. Consequently, there were too few utterances for the neural network to pull a sufficiently general number of features from this class in order to perform correct classification on test utterances. The correct classification rate for fear utterances is especially poor for the data sets derived from the mel-scale filter-band energy values. For these data sets, the only architecture that consistently classified at least one fear utterance correctly was the fully connected feedforward neural network. All other architectures failed to correctly a single one fear utterance using these data sets.

One observation, which was made over the course of numerous classification training and test sessions over a number of architectures, was that certain utterances would be consistently misclassified, irrespective of the parametric representation used for the signal of the neural network architecture used. To mention just one example, one particular pain utterance, when present in the test set, has always been classified as an anger utterance. This raises some thoughts as to the degree of pain which may be present in a given utterance. Perhaps for this particular event, this particular infant did not perceive the heel stick as a painful event and found this procedure to be more bothersome than painful. After all, if one looks at adults, the same procedure or event may be more painful for one person than for another, so perhaps the same can be said for infants as well.

4.7.2 Neural Network Parameter Variations

In this subsection, the results of some neural network parameter variations, such as the hidden layer size and number of input vectors, if applicable, will be discussed, with some considerations resulting from the tables presented in the latter portions of section 4.6.2 and section 4.6.3. All the tables relating the results for these parameter variations in these sections report error rates, which correspond to the total correct classification rate for the three-class classification problem subtracted from 1. As was the case for the previous two subsections, this subsection will be divided

according to the input parameters sets derived from the input signal from which the input data sets are derived.

Mel-Cepstrum Coefficient-Derived Input Data Sets

Table 4.12 presents the error rate, as the hidden node size is varied, for the fully connected feedforward neural network. The results for the scaled mel-cepstrum input data set are shown in table 4.12(a) and those for the normalized mel-cepstrum input data set are shown in table 4.12(b).

For both input data sets, the error rate starts off relatively high, and then reaches a minimum for a hidden layer size of 45 nodes, before increasing once again for a larger hidden layer size of 74 and for a fully connected feedforward neural network with two hidden layers with the first and second hidden layers consisting of 125 and 17 nodes respectively. This behaviour is typical for the variation of the number of hidden layer nodes; generally, the error rate will fall as the hidden layer size starts from a small number, and then increases to a larger number of nodes. At a particular hidden layer size, a minimum error rate will be reached and the error rate will then begin to increase once again as the hidden layer size continues is increased.

Also, for both the data sets derived from the mel-cepstrum coefficients, the error rate does not seem to improve once an additional hidden layer is added. Hence the interim mappings generated by the addition of another hidden layer in the network does not improve classification results. In turn, the results obtained from the use of a 2-hidden layer configuration would not warrant the substantial time and computations required to train this network. As is the case for most other applications, a single hidden layer is sufficient to capture the mappings between the inputs and the outputs for the classification of anger, fear, and pain cries. As can also be observed in table 4.12, for a given hidden layer size, the error rate of the normalized mel-cepstrum input data set is lower than that of the scaled input

data set.

Next, the error rates for the feedforward neural networks with tessellated connections, shown in table 4.13, will be considered. As can be seen in table 4.13(a) and table 4.13(b), the error rate initially starts at a high value for a small hidden layer size before reaching a minimum, and then increasing once again as the hidden layer size increases. For both the scaled and normalized mel-cepstrum coefficient input data sets, a larger grouping of input vectors seems to decrease the error rate, with the grouping of 25 mel-cepstrum input vectors achieving the best results. For this grouping of input nodes, a smaller overlap between subsequent input node groupings results in a decrease of the error rate, with an overlap of 5 input vectors yielding a lower error rate than when the overlap between adjacent "tiles" consisted of 10 vectors. In any event, as was discussed in section 4.7, the results obtained for this type of neural network connections do not improve the error rate when compared to the full connection of nodes between adjacent layers.

The results for the parameter variations performed on the recurrent neural network for the scaled and normalized mel-cepstrum input data sets are presented in table 4.14 and table 4.15, respectively. The subtables illustrate the error rate variation as the number of input data vectors in the network changes according to the number of overlapping vectors between subsequent input data frames, delay nodes, and hidden layer sizes listed. These tables also list how the hidden layer size affects the error rate for the optimal frame size of 75 input vectors.

For both the input data sets derived from the mel-cepstrum coefficients, the error rate falls as the number of input data vectors used in the recurrent neural network increases, with the best results achieved for a network using an input frame size of 75 vectors with subsequent input frames overlapping by 50 vectors. This result would seem to imply that the classification of infant anger, fear, and pain cries does not benefit from the use of a "fine" time resolution of input features. The error rates obtained for input frame sizes of 10 and 25 vectors are both larger than the

one achieved for an input frame size consisting of 75 vectors. Although the error rate for this architecture does not fall below that of the fully connected feedforward neural network, it would seem that this application does make some use of time information, but this information is better captured by a larger or coarser time window, than with a smaller one.

The results of table 4.14(b) and table 4.15(b) both follow the same pattern of error rate versus hidden layer size that was observed for the fully connected feedforward neural network: the error rate falls as the hidden layer size increases, reaching a minimum, before increasing once again as the hidden layer size is increased.

The table listing the error rates as the input frame size is varied for time-delay neural networks, using the two input data sets derived from the mel-cepstrum coefficients, is presented in table 4.16. Due to the large amount of time required to train time-delay neural networks, only two configurations were trained. One time-delay neural network was presented with 10 vectors of data at a time, or an input delay length of 10, with subsequent input frames containing 9 of the previous vectors and one new vector of input data. In order to process a one second segment of a cry signal, which consists of 125 vectors, it was decided to use two hidden layers in order to decrease the large number of delay nodes, namely 115, which would be required if a single hidden layer would have been used.

Consequently, for this fine time resolution input representation of 10 vectors, or alternatively, an input delay width of 10 vectors, two hidden layers were used, The first hidden layer consisted of 8 units containing 55 delay units and one undelayed unit, which integrated information over 56 input frames. The second hidden layer consisted of 4 hidden units with 59 delay units and one unit with no delay, which integrated the activations of the first hidden layer over 60 time units.

The other time-delay neural network constructed was designed to use a much larger input vector size thus integrating a coarser grouping of input data. This network had an input frame size, or delay length, of 105 vectors with subsequent

frames consisting of 104 vectors from the previous frame and one new vector of input data, be it for the scaled or normalized mel-cepstrum coefficient input data sets. The hidden layer consisted of 5 units containing 20 delay units and one unit with no delay, thus integrating the activations of 21 input frames.

For both the input data sets derived from the mel-cepstrum coefficients, the error rates of the coarser, or larger, time window are substantially better than the error rates obtained by the network with the smaller time window. This is consistent with the results achieved for the recurrent neural network, and is due to the observation made earlier regarding the type of information which is found in infant cry vocalizations.

For speech, TDNNs are especially good at capturing the sequence of acoustic features which constitute a phoneme, since this sequence is common to a particular phoneme, even if it is spoken by a number of different speakers. For infant cries, however, the occurrence of certain acoustic features is more important than the sequence in which they occur, which follows if one considers that infant control of vocal articulators is poor. The fact that sequence is not important for correct classification of infant cries is also reinforced by the observation that when larger groups of vectors are used as inputs, the error rate drops for both TDNNs and for recurrent neural networks. The fine time integration provided by TDNNs is important to capture the subtle differences between phonemes in speech such as /p/, /t/, and /k/, or /b/, /d/, and /g/. Since infants lack the precise articulator control required to produce acoustic events with extremely short durations and specific articulator positioning such as those mentioned above, the correct classification rate for anger, fear, and pain cries does not benefit as a result of the use of this precise and fine time information.

Moreover, comparing the results for the larger, or coarser, time window sizes for the recurrent and the time-delay neural networks, shows that the smaller error rate of the recurrent neural network with a smaller overlap rate performs better than

the time-delay neural network with a large number of input vectors and a large overlap rate. This observation also reinforces the statement that the occurrence of acoustic events in cries is more important than sequence for the classification of anger, fear, and pain cries. The large overlap rate of time-delay neural networks allows it to capture the precise sequence of acoustic events, whereas for recurrent neural networks, the delay nodes capture the activations of the nodes, with no explicit sequence being modeled. Consequently, from these results, time information improves classification rates if the input consists of a large number of input vectors. However, the incorporation of time information does not improve the correct classification rate over that of the feedforward neural networks.

This observation should not imply that sequence is not useful for the classification of cry utterances. Some studies have shown that there is indeed a correlation between fundamental frequency patterns over the course of an utterance, and pathology [Michelsson *et al.*, 1980, Michelsson *et al.*, 1984, Ostwald and Murry, 1985]. Since the data set available for this work consisted only of healthy infants, the use and usefulness of time information for the classification of pathology could not be tested.

Lastly, table 4.17 displays the effects of different training methods on the error rate and on the number of hidden layer nodes created by the cascade correlation paradigm before the network error fell below the desired level and training was stopped. For both the input data sets derived from the mel-cepstrum coefficients, the conjugate gradient learning method yielded a lower error rate than when either standard back propagation, or its variant QuickProp, were used to train the networks. The latter training method, QuickProp, generated the smallest number of hidden layer nodes, but the quality of the trained network was the worst of the three training methods, having the largest error rate of the three methods. Standard back propagation, on the other hand, generated the largest number of hidden units with the second largest error rate. For the two data sets derived from the mel-cepstrum coefficients, it would appear that conjugate gradient learning yields the

best correct classification rates, but these results are still far poorer than any of the rates for the other architectures tested.

Also, the error rates for all the architectures and configuration variations using the data sets derived from the mel-cepstrum coefficients all follow the pattern that using the normalized input data set generates lower error rates than when the scaled input data set is used, implying that the process of removing the mean and normalizing the input values to lie between ± 1 allows a given network to better capture relevant features from the inputs.

Mel-Scale Filter-Band Energy-Derived Input Data Sets

Table 4.33(a), table 4.33(b), and table 4.33(c) of table 4.33 shows the error rates versus the number of hidden layer nodes for the fully connected feedforward neural network. Here, the minimum error rate for all three cases was reached for a hidden layer size of 25 nodes. As was the case for the mel-cepstrum coefficient-derived input data sets, once the minimum error rate was reached for a given hidden layer size, further increasing the hidden layer size would cause the error rate to increase. Also, none of the error rates for the three mel-scale filter-band energy-based inputs benefit from the use of a neural network with two hidden layers. This same pattern is observed for the number of hidden layer nodes for the feedforward neural network with tessellated connections, the results of which are listed in table 4.34.

The results for the parameter variations on the recurrent neural network are indicated in table 4.35, table 4.36, and table 4.37 for the scaled, log, and normalized log of the mel-scale filter-band energy values, respectively. For these data sets, the same observation regarding the input frame size and error rate can be made as was stated for the mel-cepstrum coefficient derived data set discussed in the previous sub-subsection. The optimal error rates for the three data sets derived from the mel-scale filter-band energy values are such that the best results are achieved for

an input frame size of consisting of 75 vectors, with subsequent input frames consisting of 50 vectors from the previous input frame and 25 new vectors.

As well, the results for the hidden layer size and the error rate for an input frame size consisting of 75 input data vectors follows the same pattern as it did for the mel-cepstrum coefficient-derived input sets; a small initial hidden layer size has a high error rate, decreasing and reaching a minimum as the size was increased, and then increasing as the hidden layer size was further increased.

Table 4.38 presents the results of tests done on two time-delay neural networks as the input frame size for the input data sets derived from the mel-scale filter-band energies was varied. The TDNN with the larger input frame size, consisting of 105 input data vectors, produced the lowest error rates for all three input data sets. These results, coupled with those obtained with the recurrent neural network, further reinforce the statement made in the previous sub-subsection that coarse, or large, groupings of input vectors yields better results. Also recurrent nets produce lower error rates, since this network integrates the activation of nodes, and not the input activations, as the TDNN does. Furthermore, the fine time integration of acoustic features provided by the large overlap size of TDNNs does not benefit the classification of anger, fear, and pain from infant cry utterances for the input data sets derived from both the mel-scale filter-band energies and from the mel-cepstrum coefficients.

Lastly, the error rates obtained from the different learning methods used to train the cascade correlation neural network are shown in table 4.39. For the input data sets derived from the mel-scale filter-band energies, the network trained using standard back propagation yielded the lowest error rate. The hidden layer size of a cascade correlation network trained with standard back propagation fell between that of the QuickProp method, which yielded the next best error rate with the lowest hidden layer size generated, and the conjugate gradient training method, which yielded the highest error rate, and also the highest number of hidden layer

nodes created.

The error rates for all the architectures and configuration variations using the data sets derived from the mel-scale filter-band energy values, show that the best results were achieved when the normalized log inputs were used. The log inputs yielded better results than the scaled inputs did, presumably because scaling the energy values may result in some extremely small values in the inputs, whereas taking the logarithm of these energies better compresses the dynamic range of these values.

Comparison of the Input Data Sets

For the most part the results obtained using the data sets derived from the two sets of features derived from the cry signals are comparable. The results for normalized inputs yield the best results in both cases. When considering the usefulness of time information for the purposes of classification, both input sets have lower error rates when a coarse, or larger, grouping of vectors is input into the network.

To reiterate, this reinforces the statement that the occurrence, not the sequence, of acoustic features in the input frame is important for improved classification, which is somewhat intuitive given the difference in the articulator control required to produce cries and to produce speech. Also, the finer time integration of acoustic features performed by the time-delay neural network, leads to a higher error rate than the coarser integration of node activations performed by the recurrent neural network, which encodes sequences in a more general manner than the time-delay neural network does.

Lastly, the input data sets achieve different results insofar as the best results obtained from the cascade correlation learning methods are concerned. The mel-cepstrum derived input data sets both achieve their best results when the conjugate gradient learning method is used. The input data sets derived from mel-scale filter-

band energy values achieve their best results when standard back propagation is used. The observations regarding the input data sets and the cascade correlation network and the different learning methods are moot, however, as this architecture yields extremely poor results in any event.

4.7.3 Comparison to Other Classification Attempts

As was mentioned in section 2.3, and again in section 4.1, there has been a very limited attempt by researchers to automate the process of infant cry classification. If this attempt has indeed been more widespread, the results of the research have not been published in the literature. Consequently, the research undertaken for the classification of anger, fear, and pain from infant cry utterances performed for this dissertation represents the first attempt at the automatic classification of an infant state and is also the first attempt at using artificial neural networks in the infant cry domain.

The results presented in section 4.6 show that the best correct classification rate for anger, fear, and pain cries of infant ranging in age from two to six months was achieved for a feedforward neural network with a hidden layer consisting of 45 nodes, which used 125 vectors consisting of 11 mel-cepstrum coefficient values for which the mean was removed and then normalized to lie between ± 1. The correct classification rate obtained was 0.8359 or 83.59%.

The only other recently published work which cites an attempt a performing the automatic classification of infant cries is a conference publication authored by Xie, Ward, and Laszlo [Xie *et al.*, 1993]. This research group uses hidden Markov models to compute a cry's so-called level-of-distress, which corresponds to an adult's perception as to the infant's distress level, and quotes a correct classification rate of over 80%, without detailing their results. It should be noted that this group did not attempt to classify either infant state of pathology based on a cry

utterance; only perceptive measure of infant distress was computed. Consequently, it is difficult to compare the results of this research with that performed for this dissertation. Nonetheless, the correct classification rate of 83.59% achieved for anger, fear, and pain classification here, surpasses their classification rate for this subjective measure.

There is some concern that arises from the choice of Xie, Ward, and Laszlo to use a perceptive measure of infant distress in an automatic cry classification system, instead of trying to classify infant state or pathology directly. A number of researchers over the past 15 years have questioned the validity of using a parent's or an adult's perception of an infant's cry to determine if an infant is indeed in distress for a number of reasons. First, the relationship between the actual and perceived features of the infant's cries and the behavioural response is affected by a number of factors [Murray, 1985]. The response, or the perception of aversiveness or distress may be dependent upon the length of exposure to certain types of cries. As well, the perceived meaning of the cries changes as adult listeners are more frequently exposed to these utterances in general. Furthermore, with some listeners, the cry may elicit a nurturing response, whereas with others, the same cry may elicit a hostile response, with research undertaken in this area noting that crying is often cited as a major trigger for child abuse [Donovan and Leavitt, 1985b, Frodi, 1985, Murray, 1985]. It has also been observed that people from different cultures react differently to infant cries [Murray, 1985].

In brief, then, attempting to model a perceptive measure may not be a proper solution to the classification problem as the cry can have a paradoxical impact on the listener.

Another means of comparing the results obtained through the neural network-based classification experiments performed and reported in section 3.4, given the lack of automatic classification results for anger, fear, and pain cries, is to compare these results with those of studies where the classification of these types of cries

is attempted by adults who themselves have infants. A review of cry perception research, performed in 1985 by Boukydis [Boukydis, 1985], reveals the presence of one study where the recognition of anger and pain cries was performed by adults who themselves had infants [Weisenfeld *et al.*, 1981].

In this study, Weisenfeld, Zander Malatesta, and DeLoach report that mothers correctly identified the anger and pain cries of infants approximately 77% of the time. The correct classification of their own infant's cries was significantly higher than that for other infants, 82.5% versus 72% respectively. Fathers, on the other hand, did very poorly in correctly classifying anger and pain from cries, with the reported correct classification results being approximately 50%. Unlike their spouses, the fathers showed no difference between the correct classification of anger and pain cries of their infant versus that of other infants. It should be noted that for this study, the infant was considered as producing an anger cry upon either being physically restrained or when its pacifier was removed, and considered as producing a pain cry, when its heel was snapped with a rubber band. The latter differs from the data used in this dissertation, where the infant was considered to have produced a pain cry after a heel stick.

Comparing the best results of the Weisenfeld, Zander Malatesta, and DeLoach study with the best three-class neural network classification results, the neural network still performs slightly better than the rate which is quoted for mother identifying the cries of their own infant (83.59% versus 82.5%). If the general correct classification rate for the mothers in the study is used, then the neural network's performance is much better than that of the mothers'.

In short, then, the results of the best neural network-based anger, fear, and pain classifier exceeds the results of both the hidden Markov model-based distress classification system of Xie, Ward, and Laszlo, as well as surpassing the best classification rates of parents on similar types of vocalizations.

Chapter 5 Future Work

This chapter presents some future work which could be undertaken as a result of the research performed for this dissertation, and, which was presented in chapter 3 and in chapter 4. The chapter will be divided into two sections, one tackling the possible future extensions for the improved crosscorrelation vector-based fundamental frequency extraction method, and another addressing the extensions for the neural network-based classification of infant cries.

5.1 Future Extensions for the Improved Crosscorrelation Vector-Based Fundamental Frequency Method

This subsection deals with the possible future extensions for the improved crosscorrelation vector-based fundamental frequency extraction method presented in section 3.1. To reiterate, this method is capable of tracking rapid changes in the fundamental frequency of infant cry utterances, handling the large range of F_0 values present in infant cry signals, generates values of F_0 for almost every pitch period in voiced utterances, and is also useful for improved visualization of cry utterances.

5.1.1 Improvements in Speed

The current implementation of the improved crosscorrelation vector-based fundamental frequency extraction method is computationally intensive, as was mentioned in subsection 3.1.6, since, during the signal transformation phase, a cross-

correlation value is generated for every possible lag in the range of expected fundamental frequency values. Over the length of a recording , this amounts to a large number of computations, which consumes the majority of the computation time of the pitch extraction algorithm.

One method of reducing the number of computations required for a given time index would be to calculate crosscorrelation values for every other lag in the expected pitch period range, instead of calculating crosscorrelation values for every lag. Thus would reduce the required number of computations for this stage of the algorithm by half and would also reduce the amount of memory required to store the collection of crosscorrelation vectors. However, this savings in both the number of computations and memory comes at the expense of the resolution of the extracted pitch period values in the subsequent post-processing stage, and correspondingly, a reduction in the resolution of the crosscorrelogram. If execution speed is of importance for a particular application, for example, then this extension could easily be implemented and tested for its effectiveness.

5.1.2 Pitch-Synchronous Processing

Another extension of this method would be to use it for pitch-synchronous processing subsequent to the extraction of the pitch period in a given recording. As well, a further pass over the time indexes, or the n_0s, as was introduced on page 49 in section 3.1.2, and the input cry utterance could result in the following extensions, illustrated in figure 5.1.

First, one could supplement the missing pitch values due to large time increments resulting from large maxima at multiples of the actual pitch period value, or correct those due to small increments resulting from narrow bandwidth F_1 values. As well, one could use techniques to interpolate between sample values in order to obtain so-called "infinite" resolution in the extracted pitch period values

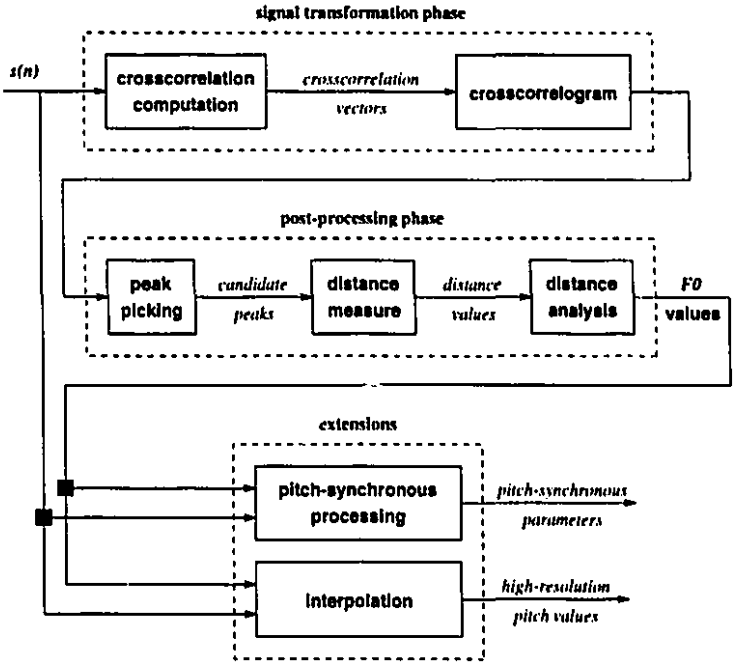


Figure 5.1: Extension to Improved Pitch Period Processing Method

[Medan *et al.*, 1991]. The improved crosscorrelation vector-based pitch extraction algorithm synchronizes itself to the maximum value of a pitch period when it first begins to find periodicity within a portion of the signal. Consequently, one could use the time indexes used by the algorithm, which mark the beginning of a pitch period during voiced sections of the recording, to extract pitch-synchronous features such as formant values, or proceed to perform another pass over the data and obtain all the infinite resolution pitch values for further processing.

Other, more detailed measures of parameters based on F_0 such as, jitter and shimmer, could be determined on a period-by-period basis, allowing a more precise picture as to how these parameters evolve over the length of the cry episode than was previously possible. This insight could shed more light into the precise way that these and other parameters behave for different types of cries, recorded in different contexts. This could be especially useful for cries of infants with pathological or genetic problems [Lind *et al.*, 1970, Zeskind and Lester, 1978] and, as well, for adults with varying degrees of pathology of the vocal tract [Kasuya *et al.*, 1983].

This pitch synchronous processing could also be useful for eventual coding of the signal, should it be necessary to transmit an utterance from a remote center over a low bandwidth connection, for example, to a central processing system for further analysis, or for archival purposes.

5.1.3 Other Fundamental Frequency Extraction Methods

Although the improved crosscorrelation-based fundamental frequency extraction method represents a significant improvement in the determination of pitch period values from infant cry utterances, other emerging methods are currently being researched for improved fundamental frequency extraction from speech signals which could possibly be tested on cry utterances as well. One particular method which appears to be promising is the application of wavelets [Boashash, 1992a, Boashash, 1992b].

Some research groups have attempted to extract the fundamental frequency from speech signals using this method with reasonable results [Kadambe and Boudreaux-Bartels, 1991]. Recently, however, good results have been achieved through the application of new wavelet functions for speech coding applications which may prove to be useful for fundamental frequency extraction [Kinsner and Langi, 1993].

5.2 Future Work for Neural Network-Based Infant Cry Classification

This section presents some future work for the automated classification of infant cry signals using neural networks which emerges as a result of the work presented in chapter 4. Since attempts at automatic classification of cry signals in general are just beginning, there is much that could be said on this topic. However, this section

will briefly touch on some points which could be investigated in future attempts at addressing this problem.

5.2.1 Other Neural Network Architectures

As as result of the correct classification rate of 83.59% achieved using a feedforward neural network, with a single hidden layer consisting of 45 units, using an input data set of mel-cepstrum coefficients with the mean of the input vectors removed, and normalized to lie between ± 1, it can be deduced that artificial neural networks are suitable for the discrimination of anger, fear, and pain cries.

Consequently, other activation functions, such as radial basis functions [Morgan and Scofield, 1991], and neural network architectures, such as Coulomb energy networks [Scofield *et al.*, 1988], or Viterbi networks [Lippmann and Singer, 1993] could be trained and tested with the input parameter sets used for the neural network tests of chapter 4.

5.2.2 Other Parametric Representations

The data sets used in the neural network tests of chapter 4 were derived from either 11 mel-cepstrum coefficients of 20 mel-scale filter-band energy values. Another series of tests which could be performed would be to reduce the dimensionality of both the parameter vectors. The number of mel-scale cepstrum coefficients extracted from a given signal frame could be reduced from 11 to, say, 7. As well, further tests could be conducted on augmenting this set of 7 mel-cepstrum coefficients with 7 differential mel-cepstrum coefficients. The differential mel-cepstrum coefficients simply correspond to the first time-derivative of these features. The augmentation of mel-cepstrum coefficients with differential mel-cepstrum coefficients has provided good results for speech recognition applications [Flaherty and Roe, 1993],

and may be worth investigating for the classification of infant cries as well.

Alternatively, other input features such as the linear predictive coding (LPC) coefficients, or higher order representations from the input spectrum [Nikias and Mendel, 1993] could also be generated and tested as input for different artificial neural networks, and whose correct classification results could be compared to those presented in section 4.6.

The portion of the voiced utterances which were parametrized for input into the different neural networks, consisted of the first second of an utterance which lasted at least 0.75 seconds after the stimulus event. Here, parametrization of different portions of the cry utterance could be tested as well; taking a 1 second portion of the signal centered about the signal frame with the largest energy value, for example. Alternatively, a longer portion of the utterance could be parametrized using no overlap between subsequent signal windows. If this were done for the parametric representations used in chapter 4, a two second segment of the cry signal could be taken, without increasing the number of input vectors presented to the neural network architectures tested.

5.2.3 Expanding the Study

Other neural network-based classification experiments could be expanded to attempt the classification of other infant states, such as hunger, or to classify various pathological or genetic disorders, if a collection of data is available. As well, future experiments could be expanded to include premature infants.

One drawback of research undertaken in this domain, is the lack of a standardized data set, on which a number of researchers could compare the results of new methods of either classification, or parameter extraction, as is the case for speech, or image processing, with the availability of speech databases or standard image files. This allows the improvements in processing or classification techniques to not

only be done on a de facto standard set of data, but also allows researchers all over the world to have access to the same data and to determine whether improvements are due to genuine improvements or a result of a limited data set.

Lastly, since the data sets are painstakingly collected in a clinical setting, it is often difficult to obtain more than a few recordings of a specific cry type, which causes problems for the training process of classification methods. These methods often require a large number of representative data for the training and testing processes in order to determine the unbiased estimation of the classification method being tested. Consequently, it would be desirable that when automatic classification methods are being investigated in the future, that a large number of recordings be available for this purpose, and that equal numbers of recordings be available for the different types or classes of cry utterances slated for classification.

Chapter 6 Conclusion

One of the goals of this work was to address the problems inherent in the processing of infant cry signals, most notably for the extraction of vocal fundamental frequency, since this is a very important parameter in the determination of infant state and future developmental outcome as was mentioned chapter 2.

This thesis developed a method which was capable of accurately extracting this parameter using a multi-stage time-domain method called the improved cross-correlation vector-based fundamental frequency extraction method presented in chapter 3. The method uses a distance scoring method of the pitch candidates in order to extract the correct fundamental frequency values over the length of an utterance, and is able to correctly deal with discontinuities in the pitch contour due to rapid or sudden variations in pitch, double harmonic break episodes, and disphonation. As well, the method generates pitch values for almost every pitch period in the voiced sections of cry utterances, which can be further refined by implementing the future extensions for this method mentioned in section 5.1.

The improved crosscorrelation vector-based fundamental frequency extraction method overcame the limitations of the standard frame-based pitch extraction methods, the most common of which were presented in section 3.2. Typically, traditional pitch extraction methods are not well suited to the large range of fundamental frequency values of infant cry utterances. Moreover, characteristics of certain types of cry utterance signals, such as narrow bandwidth high energy formant values, create confusion for other F_0 extraction methods. As well, since there are no comparisons between different pitch extraction methods which exist for infant cry utterances, this work was also performed and reported on in order to demonstrate the improvements in the results of the new pitch extraction method,

and the more popular existing methods. In addition to providing improvements in the extracted pitch values, this method also generates a series of crosscorrelation vectors which are extremely useful for improved visualization of cry utterances. The improvements need not be limited to the processing of cry signals, however. This method may also be useful for the extraction of pitch values from adult speakers who may have aperiodicities in their vocalizations and for whom a very fine analysis of pitch period variations may be required to determine the extent of vocal tract pathology.

Another of the goals of this work was to perform the accurate automated classification of infant anger, fear, and pain cries, which was achieved using feedforward artificial neural networks and the first second of a voiced utterance parametrized using 11 mel-cepstrum coefficients for which an overall correct classification rate of 83.59% was achieved, illustrating the suitability of this paradigm for this particular application.

The comparison between two different parametric representations, the mel-cepstrum coefficients and mel-scale filter-band energy values, extracted from a 1 second portion of a voiced cry utterance trained on four different neural networks was also reported and compared. To the best of our knowledge, this work represents the first attempt at, and comparison of, automated infant cry classification using artificial neural networks. Based on the results of the various input parameter sets and neural network architectures investigated, the relevance of certain types of information were discussed and presented in section 4.7. Some ideas for possible future work relating to both the parametric representations and neural network-based classification experiments were also presented.

References

[Ahalt and Jung, 1991] S. C. Ahalt and T.-P. Jung, "A comparison of MLP and FCL-LVQ neural networks for vowel classification," in *Neural Networks: Concepts, Applications, and Implementations* (P. Antognetti and V. Multinovic, eds.), vol. 3, ch. 6, pp. 124–143, Englewood Cliffs, New Jersey: Prentice Hall Inc., 1991.

[Anand and Hickey, 1987] K. J. S. Anand and P. R. Hickey, "Pain and its effects in the human neonate and fetus," *New England Journal of Medicine*, vol. 317, pp. 1321–1329, 1987.

[Anand *et al.*, 1989] K. J. S. Anand, D. Phil, and D. B. Carr, "The neuroanatomy, neurophysiology, and neurochemistry of pain, stress, and analgesia in newborns and children," *Pediatric Clinics of North America*, vol. 36, pp. 795–822, August 1989.

[Ananthapadmanabha and Yegnanarayana, 1979] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," vol. 27, pp. 309–319, August 1979.

[Anderson-Huntington and Rosenblith, 1976] R. B. Anderson-Huntington and J. F. Rosenblith, "Central nervous system damage as a possible component of unexpected deaths in infancy," *Developmental Medicine and Child Neurology*, vol. 18, pp. 480–492, August 1976.

[Anderson, 1993] T. R. Anderson, "Comparison of auditory models for speaker independent phoneme recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 231–234, IEEE, 1993.

[Andrews *et al.*, 1989] M. S. Andrews, R. D. DeGroat, and J. Picone, "Robust cepstral based pitch determination," in *Asilomar Conference on Signals, Systems, and Computers*, (Washington, D.C.), pp. 744–748, Computer Society Press of the IEEE, 1989.

[Andrews *et al.*, 1990a] M. S. Andrews, R. D. DeGroat, and J. Picone, "A high performance MUSIC based pitch extractor," in *Asilomar Conference on Circuits, Systems, and Computers*, (Washington, D.C.), pp. 669–673, Computer Society Press of the IEEE, 1990.

[Andrews *et al.*, 1990b] M. S. Andrews, J. Picone, and R. D. DeGroat, "Robust pitch determination via SVD based cepstral methods," (Albuquerque, New Mexico), pp. 253–256, IEEE, April 3-6 1990.

[Atal and Rabiner, 1976] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," vol. 24, pp. 201–212, June 1976.

[Barnard *et al.*, 1991] E. Barnard, R. A. Cole, M. P. Vea, and F. A. Alleva, "Pitch detection with a neural net classifier," vol. 39, pp. 298–307, February 1991.

[Beemer *et al.*, 1984] F. A. Beemer, H. F. de France, I. J. Rosina-Angelista, L. J. Gerards, B. P. Cats, and R. Guyt, "Familial partial monosomy 5p and trisomy 5q; three cases due to paternal pericentric inversion 5 (p151q333)," *Clinical Genetics*, vol. 26, pp. 209–215, September 1984.

[Bengio and Mori, 1988] Y. Bengio and R. D. Mori, "Use of neural networks for the recognition of place of articulation," pp. 103–106, April 1988.

[Bengio *et al.*, 1992] Y. Bengio, R. De Mori, G. Flammia, and R. Kompe, "Phonetically motivated acoustic parameters for continuous speech recognition using artificial neural networks," *Speech Communication*, vol. 11, pp. 261–271, June 1992.

[Benini *et al.*, 1993] F. Benini, C. C. Johnston, D. Faucher, and J. V. Aranda, "Topical anesthesia during circumcision in newborn infants," *Journal of the American Medical Association*, vol. 270, pp. 850–853, August 1993.

[Boashash, 1992a] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal - Part 1: Fundamentals," *Proceedings of the IEEE*, vol. 80, pp. 520–538, April 1992.

[Boashash, 1992b] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal - Part 2: Algorithms and applications," *Proceedings of the IEEE*, vol. 80, pp. 540–568, April 1992.

[Bodenhausen and Waibel, 1991] U. Bodenhausen and A. Waibel, "Learning the architecture of neural networks for speech recognition," (Toronto, Canada), pp. 117–120, IEEE, 1991.

[Bosma *et al.*, 1965] J. F. Bosma, H. M. Truby, and J. Lind, "Cry motions of the newborn infant," *Acta Paedriatica Scandinavia Supplement*, vol. 163, pp. 61–92, 1965.

[Boukydis, 1985] C. F. Z. Boukydis, "Perception of infant crying as an interpersonal event," in *Infant Crying: Theoretical and Research Perspectives*, ch. 9, pp. 187–215, New York, New York: Plenum Press, 1985.

[Bourlard and Morgan, 1993] H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*. Hingham, MA: Kluwer Academic Press, 1993.

[Bourlard and Wellekens, 1987] H. Bourlard and C. Wellekens, "Multilayer perceptrons and automatic speech recognition," in *IEEE First International Conference on Neural Networks*, pp. 407–416, 1987.

[Bourlard and Wellekens, 1989] H. Bourlard and C. Wellekens, "Speech dynamics and recurrent neural networks," (Glasgow, Scotland), pp. 33–36, IEEE, May 1989.

[Breiman, 1984] L. Breiman, *Classification and Regression Trees*. Wadsworth Statistics/Probability Series, Belmont, CA: Wadsworth International Group, 1984.

[Bridle, 1991] J. S. Bridle, "Neural networks or hidden markov models for automatic speech recognition: Is there a choice?," in *Speech Recognition and Understanding: Recent Advances, Trends and Applications* (P. Laface, ed.), Springer-Verlag, 1991.

[Brown and Puckette, 1993] J. C. Brown and M. S. Puckette, "A high resolution fundamental frequency determination based on phase changes of the Fourier transform," *The Journal of the Acoustical Society of America*, vol. 94, pp. 662–667, August 1993.

[Charpentier, 1986] F. J. Charpentier, "Pitch detection using the short-term phase spectrum," (Tokyo, Japan), pp. 113–116, IEEE, May 1986.

[Cheng and O'Shaughnessy, 1989] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," vol. 37, pp. 1805–1815, December 1989.

[Chernos *et al.*, 1992] J. E. Chernos, S. B. Fowlow, and D. M. Cox, "Cri du chat syndrome due to meiotic recombination in a pericentric inversion 5 carrier," *Clinical Genetics*, vol. 41, pp. 266–269, May 1992.

[Chung and Algazi, 1985] S. Chung and V. R. Algazi, "Improved pitch detection algorithm for noisy speech," pp. 407–410, IEEE, 1985.

[Colombi *et al.*, 1993] J. M. Colombi, T. R. Anderson, S. K. Rogers, D. W. Ruck, and G. T. Warhola, "Auditory model representation and comparison for speaker recognition," in *1993 IEEE International Conference on Neural Networks*, pp. 1914–1917, IEEE, 1993.

[Colton and Steinschneider, 1980] R. Colton and A. Steinschneider, "Acoustic relationships of infant cries to the sudden infant death syndrome," in *Infant Communication: Cry and Early Speech* (T. Murry and J. Murry, eds.), ch. 8, pp. 183–208, Houston, TX: College-Hill Press, 1980.

[Colton *et al.*, 1985] R. H. Colton, A. Steinschneider, L. Black, and J. Gleason, "The newborn infant cry: Its potential implications for development and SIDS," in *Infant Crying: Theoretical and Research Perspectives* (B. M. Lester and C. F. Z. Boukydis, eds.), ch. 6, pp. 119–138, New York, NY: Plenum Press, 1985.

[Corwin *et al.*, 1992] M. J. Corwin, B. M. Lester, C. Sepkoski, S. McLaughlin, H. Kayne, and H. L. Golub, "Effects of in utero cocaine exposure on newborn acoustical cry characteristics," *Pediatrics*, vol. 89, pp. 1199–1203, June 1992.

[Darwin, 1872] C. Darwin, *The Expression of the Emotions in Man and Animals*. London: J. Murray, 1872.

[Dautrich *et al.*, 1983] B. A. Dautrich, L. R. Rabiner, and T. B. Martin, "On the effects of varying filter bank parameters on isolated word recognition," vol. 31, pp. 793–807, August 1983.

[Davis and Mermelstein, 1980] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," vol. 28, pp. 357–366, August 1980.

[Davis, 1959] H. Davis, "Excitation of auditory receptors," in *Handbook of Physiology: Section I. Neurophysiology (Volume 1)* (J. Field, ed.), Baltimore, MD: Williams and Wilkins, 1959.

[Dayhoff, 1990] J. E. Dayhoff, *Neural Network Architectures: An Introduction*. New York, NY: Van Nostrand Reinhold, 1990.

[De Mori and Omologo, 1993] R. De Mori and M. Omologo, "Normalised correlation features for speech analysis and pitch extraction," in *Visual Representation of Speech Signals*, ch. 31, John Wiley & Sons Ltd., 1993.

[Demichelis *et al.*, 1989] P. Demichelis, L. Fissore, P. Laface, G. Micca, and E. Piccolo, "On the use of neural networks for speaker independent isolated word recognition," (Glasgow, Scotland), pp. 314–317, IEEE, May 1989.

[Donovan and Leavitt, 1985a] W. L. Donovan and L. A. Leavitt, "Simulating conditions of learned helplessness: the effects of interventions and attributions," *Child Development*, vol. 56, pp. 594–603, June 1985.

[Donovan and Leavitt, 1985b] W. L. Donovan and L. A. Leavitt, "Physiology and behaviour: Parents' response to the infant cry," in *Infant Crying: Theoretical and Research Perspectives* (B. M. Lester and C. F. Z. Boukydis, eds.), ch. 11, pp. 241–259, New York, New York: Plenum Press, 1985.

[Donovan and Leavitt, 1989] W. L. Donovan and L. A. Leavitt, "Maternal self-efficacy and infant attachment: integrating physiology, perceptions, and behavior," *Child Development*, vol. 60, pp. 460–472, April 1989.

[Donovan *et al.*, 1990] W. L. Donovan, L. A. Leavitt, and R. O. Walsh, "Maternal self-efficacy: illusory control and its effect on susceptibility to learned helplessness," *Child Development*, vol. 61, pp. 1638–1647, October 1990.

[Donovan, 1981] W. L. Donovan, "Maternal learned helplessness and physiologic response to infant crying," *Journal of Personality and Social Psychology*, vol. 40, pp. 919–926, May 1981.

[Donzelli *et al.*, 1994] G. Donzelli, G. Rapisardi, M. Moroni, S. Zani, B. Tomasini, A. Ismaelli, and P. Bruscaglioni, "Computerized cry analysis in infants affected by severe protein energy malnutrition," *Acta Paediatrica*, vol. 83, pp. 204–11, February 1994.

[Dubnowski *et al.*, 1976] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time digital hardware pitch detector," vol. 24, pp. 2–8, February 1976.

[Duifhuis *et al.*, 1982] H. Duifhuis, L. F. Willems, and R. J. Sluyter, "Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception," vol. 71, pp. 1568–1580, June 1982.

[Fahlman and Lebiere, 1991] S. E. Fahlman and C. Lebiere, "The cascade-correlation learning architecture," Tech. Rep. CMU-CS-90-100, School of Computer Science, Carnegie Mellon University, August 1991.

[Fahlman, 1988] S. E. Fahlman, "Faster-learning variations on back-propagation: An empirical study," in *1988 Connectionist Models Summer School* (T. J. Sejnowski, G. E. Hinton, and D. S. Touretzky, eds.), San Mateo, CA: Morgan Kauffman, 1988.

[Fairbanks, 1942] G. Fairbanks, "An acoustical study of the pitch of infant hunger wails," *Child Development*, vol. 13, no. 3, pp. 227–232, 1942.

[Fallside *et al.*, 1990] F. Fallside, H. Lucke, T. P. Marsland, P. J. O'Shea, M. S. J. Owen, R. W. Prager, A. J. Robinson, and N. H. Russell, "Continuous speech recognition for the TIMIT database using neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 445–448, 1990.

[Feijóo and Hernández, 1985] S. Feijóo and C. Hernández, "Quantification of dysphony with allowance for inter-utterance variation," pp. 423–426, IEEE, May 1985.

[Finnegan, 1985] L. P. Finnegan, "Effects of maternal opiate abuse on the newborn," *Federation Proceedings*, vol. 44, pp. 2314–2317, April 1985.

[Flaherty and Roe, 1993] M. J. Flaherty and D. B. Roe, "Orthogonal transformations of stacked feature vectors applied to HMM speech recognition," *IEE Proceedings. Part I, Communications, Speech, and Vision*, vol. 140, pp. 121–126, April 1993.

[Flatau and Gutzmann, 1906] T. S. Flatau and H. Gutzmann, "Die stimme des säuglings," *Archiv für Laryngologie and Rhinologie*, vol. 64, p. 119, 1906.

[Friedman, 1978] D. H. Friedman, "Multidimensional pseudo-maximum-likelihood pitch estimation," vol. 26, pp. 185–196, June 1978.

[Frodi, 1985] A. Frodi, "When empathy fails: Aversive infant crying and child abuse," in *Infant Crying: Theoretical and Research Perspectives* (B. M. Lester and C. F. Z. Boukydis, eds.), ch. 12, pp. 263–278, New York, New York: Plenum Press, 1985.

[Fuller and Horii, 1988] B. F. Fuller and Y. Horii, "Spectral energy distribution of four types of infant vocalizations," *Journal of Communication Disorders*, vol. 2, no. 3, pp. 111–121, 1988.

[Fuller, 1991] B. F. Fuller, "Acoustic discrimination of three types of infant cries," *Nursing Research*, vol. 40, pp. 336–340, May-June 1991.

[Geckinli and Yavuz, 1977] N. C. Geckinli and D. Yavuz, "Algorithm for pitch extraction using zero-crossing interval sequence," vol. 25, pp. 559–564, December 1977.

[Ghitza, 1986] O. Ghitza, "Auditory nerve respresentation as a front-end for speech recognition in a noisy environment," *Computer, Speech, and Language*, vol. 1, pp. 109–130, 1986.

[Ghosh and Reilly, 1994] S. Ghosh and D. L. Reilly, "Credit card fraud detection with a neural-network," in *Proceedings of the Hawaii International Conference on System Sciences*, (Los Alamitos, CA), pp. 621–630, IEEE Computer Society Press, 1994.

[Gladding, 1978] S. T. Gladding, "Empathy, gender, and training as factors in the identification of normal infant cry-signals," *Perceptual and Motor Skills*, vol. 57, pp. 267–270, August 1978.

[Gold *et al.*, 1987] B. Gold, R. P. Lippmann, and M. L. Malpass, "Some neural net recognition results on isolated words," in *Proceedings ICNN*, (San Diego), pp. 427–434, June 1987.

[Goldstein *et al.*, 1978] J. L. Goldstein, A. Gerson, P. Srulovicz, and M. Furst, "Verification of the optimal probabilistic basis of aural processing of pitch of complex tones," vol. 63, pp. 486–497, 1978.

[Goldstein, 1973] J. L. Goldstein, "An optimum processor for the central formation of pitch complex tones," vol. 54, pp. 1496–1516, 1973.

[Golub and Corwin, 1982] H. L. Golub and M. J. Corwin, "Infant cry: a clue to diagnosis," *Pediatrics*, vol. 69, pp. 197–201, February 1982.

[Golub and Corwin, 1985] H. L. Golub and M. J. Corwin, "A physioacoustic model of the infant cry," in *Infant Crying: Theoretical and Research Perspectives*, ch. 3, New York, New York: Plenum Press, 1985.

[Gong and Haton, 1987] Y. Gong and J.-P. Haton, "Time domain harmonic matching pitch estimation using time-dependent speech modeling," vol. 35, pp. 1386–1400, October 1987.

[Goryn and Kaveh, 1991] D. Goryn and M. Kaveh, "Conjugate gradient learning algorithms for multilayer perceptrons," in *Proceedings of the 32nd Midwest Symposium on Circuits and Systems Part 2*, (Champaign, IL), pp. 736–739, IEEE, 1991.

[Grunau and Craig, 1987] R. V. Grunau and K. D. Craig, "Pain expression in neonates: facial action and cry," *Pain*, vol. 28, pp. 395–410, March 1987.

[Grunau *et al.*, 1990] R. V. E. Grunau, C. C. Johnston, and K. D. Craig, "Neonatal facial and cry responses to invasive and non-invasive procedures," *Pain*, vol. 42, pp. 295–305, 1990.

[Hadjistavropoulos *et al.*, 1994] H. D. Hadjistavropoulos, K. D. Craig, R. V. E. Grunau, and C. C. Johnston, "Judging pain in newborns - facial and cry determinants," *Journal of Pediatric Psychology*, vol. 19, pp. 485–491, August 1994.

[Hadjitodorov *et al.*, 1994] S. Hadjitodorov, B. Boyanov, T. Ivanov, and N. Dalakchieva, "Text-independent speaker identification using neural nets and AR-vector models," *Electronics Letters*, vol. 30, pp. 838–840, May 1994.

[Hanna, 1992] S. A. Hanna, "Frequency-domain maximum likelihood pitch determination approach," *International Journal of Electronics*, vol. 73, pp. 1185–1199, December 1992.

[Hataoka *et al.*, 1990] N. Hataoka, A. Amano, T. Aritsuka, and A. Ichikawa, "Large vocabulary speech recognition using neural-fuzzy and concept networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (New York, NY, USA), pp. 513–16, IEEE, 1990.

[Haykin, 1994] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New York, NY: Macmillan College Publishing Company, 1994.

[Hecht-Nielsen, 1990] R. Hecht-Nielsen, *Neurocomputing*. Reading, Massachusetts: Addison-Wesley Publishing Company, Inc., 1990.

[Hedelin and Huber, 1990] P. Hedelin and D. Huber, "Pitch period determination of aperiodic speech signals," (Albuquerque, New Mexico), pp. 361–364, IEEE, April 3-6 1990.

[Hermansky *et al.*, 1977] H. Hermansky, B. A. Hanson, and H. Fujisaka, "Linear predictive coding of speech in modified spectral domains," in *Proceedings of the Conference on Digital Processing of Signals in Communications*, (Loughborough, U.K.), pp. 55–62, Institution of Electronic and Radio Engineers, September 6-9 1977.

[Hess, 1976] W. J. Hess, "A pitch-synchronous digital feature extraction system for phoneme recognition of speech," vol. 24, pp. 14–24, February 1976.

[Hess, 1983] W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices*. Berlin: Springer-Verlag, 1983.

[Hollien, 1980] H. Hollien, "Developmental aspects of neonatal vocalizations," in *Infant Communication: Cry and Early Speech* (T. Murry and J. Murry, eds.), ch. 2, pp. 21–55, Houston, TX: College-Hill Press, 1980.

[Hopfield, 1984] J. J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons," *Proceedings of the National Academy of Science USA*, vol. 80, pp. 3088–3092, May 1984.

[Huang *et al.*, 1988] W. Huang, R. P. Lippman, and B. Gold, "A neural net approach to speech recognition," pp. 99–102, April 1988.

[Illingsworth, 1980] R. S. Illingsworth, "The development of communication in the first year and factors which affect it," in *Infant Communication: Cry and Early Speech* (T. Murry and J. Murry, eds.), ch. 1, pp. 4–19, Houston, TX: College-Hill Press, 1980.

[Imazumi, 1986] S. Imazumi, "Acoustic measurement of pathologic voice qualities for medical purposes," pp. 677–680, 1986.

[Indefrey *et al.*, 1985] H. Indefrey, W. Hess, and G. Seeser, "Design and evaluation of double-transform pitch determination algorithms with nonlinear distortion in the frequency domain," pp. 415–418, IEEE, May 1985.

[Jagoda and Renner, 1990] A. Jagoda and G. Renner, "Infant botulism: case report and clinical update," *American Journal of Emergency Medicine*, vol. 8, pp. 318–320, July 1990.

[Jin and Chung, 1992] G. Jin and L. H. Chung, "Multilayer perceptron postprocessor to hidden Markov modeling for speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 263–266, IEEE, 1992.

[Johnston and O'Shaughnessy, 1988] C. C. Johnston and D. O'Shaughnessy, "Acoustical attributes of infant pain cries: discriminating features," in *Proceedings of the Vth World Congress on Pain* (R. Dubner, G. F. Gebhart, and M. R. Bonds, eds.), (Hamburg, Germany), pp. 336–340, August 1988.

[Johnston and Strada, 1986] C. C. Johnston and M. E. Strada, "Acute pain response in infants: a multidimensional description," *Pain*, vol. 24, pp. 373–382, March 1986.

[Johnston *et al.*, 1993] C. C. Johnston, B. Stevens, K. D. Craig, and R. V. Grunau, "Developmental changes in pain expression in premature, full-term, two- and four-month-old infants," *Pain*, vol. 52, pp. 201–208, February 1993.

[Johnston, 1989] C. C. Johnston, "Pain assessment and management in infants," *Pediatrician*, vol. 16, no. 1, pp. 16–23, 1989.

[Juntunen *et al.*, 1978] K. Juntunen, P. Sirvio, and K. Michelsson, "Cry analysis in infants with severe malnutrition," *European Journal of Pediatrics*, vol. 128, pp. 241–246, July 1978.

[Kadambe and Boudreaux-Bartels, 1991] S. Kadambe and G. F. Boudreaux-Bartels, "A comparison of a wavelet functions for pitch detection and speech signals," pp. 449–452, IEEE, April 14-17 1991.

[Kadambe and Boudreaux-Bartels, 1992] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Transactions on Information Theory*, vol. 38, pp. 917–924, March 1992.

[Kadambe and Srinivasan, 1994] S. Kadambe and P. Srinivasan, "Applications of adaptive wavelets for speech," *Optical Engineering*, vol. 33, pp. 2204–2211, July 1994.

[Kasuya *et al.*, 1983] H. Kasuya, Y. Kobayashi, and T. Kobayashi, "Characteristics of pitch period and amplitude perturbations in pathologic voice," (Boston, Mass.), pp. 1372–1375, IEEE, 1983.

[Keating, 1980] P. Keating, "Patterns of fundamental frequency and vocal registers," in *Infant Communication: Cry and Early Speech* (T. Murry and J. Murry, eds.), ch. 9, pp. 209–233, Houston, TX: College-Hill Press, 1980.

[Kennedy III and Kuehn, 1989] J. G. Kennedy III and D. P. Kuehn, "Neuroanatomy of speech," in *Neural Bases of Speech, Hearing and Language* (D. P. Kuehn, M. L. Lemme, and J. M. Baumgartner, eds.), ch. 5, pp. 111–145, Boston, MA: College-Hill Press, 1989.

[Kepuska and Gowdy, 1989] V. Z. Kepuska and J. N. Gowdy, "Phonemic speech recognition system based on a neural network," in *SOUTHEASTCON '89 Proceedings. Energy and Information Technologies in the Southeast*, vol. 2, (New York, NY, USA), pp. 770–5, IEEE, 1989.

[Kinsner and Langi, 1993] W. Kinsner and A. Langi, "Speech and image signal compression with wavelets," in *1993 IEEE Wescanex Conference on Communications, Computers and Power in the Modern Environment*, (Saskatoon, Canada), pp. 368–375, IEEE, 1993.

[Kohonen *et al.*, 1984] T. Kohonen, K. Mäkisara, and T. Saramäki, "Phonotopic maps – insightful representation of phonological features for speech recognition," in *Proceedings of the Seventh International Conference on Pattern Recognition*, (Montreal, Canada), pp. 182–185, July 1984.

[Kohonen, 1988] T. Kohonen, *Self-Organization and Associate Memory*. Series in Information Sciences, Berlin: Springer-Verlag, second ed., 1988.

[Korner and Grobstein, 1966] A. Korner and R. Grobstein, "Visual alertness as related to soothing in neonates: Implications for maternal stimulation and early deprivation," *Child Development*, vol. 37, pp. 867–876, 1966.

[Kuah *et al.*, 1994] K. Kuah, M. Bodruzzaman, and S. Zein-Sabatto, "Neural network-based text independent voice recognition system," in *Conference Proceedings - IEEE SOUTHEASTCON*, pp. 131–135, IEEE, 1994.

[Lahat *et al.*, 1987] M. Lahat, R. J. Niederjohn, and D. A. Krubsack, "A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech," vol. 35, pp. 741–750, June 1987.

[Laver *et al.*, 1982] J. Laver, S. Hiller, and R. Hanson, "Comparative performance of pitch detection algorithms on dysphonic voices," pp. 192–195, IEEE, 1982.

[Leighton, 1992] R. R. Leighton, *The Aspirin/MIGRAINES Neural Network Software: User's Manual Release 6.0*. The MITRE Corporation, October 1992.

[Lemme *et al.*, 1989] M. L. Lemme, D. P. Kuehn, and J. M. Baumgartner, "Studying the nervous system: Communication science perspective," in *Neural Bases of Speech, Hearing, and Language* (D. P. Kuehn, M. L. Lemme, and J. M. Baumgartner, eds.), ch. 1, pp. 1–24, Boston, MA: College-Hill Press, 1989.

[Lester and Boukydis, 1985] B. M. Lester and C. F. Z. Boukydis, *Infant Crying: Theoretical and Research Perspectives*. New York, New York: Plenum Press, 1985.

[Lester *et al.*, 1989] B. M. Lester, L. T. Anderson, C. F. Boukydis, C. T. Garcia-Coll, B. Vohr, and M. Peucker, "Early detection of infants at risk for later handicap through acoustic cry analysis," *Birth Defects: Original Article Series*, vol. 25, no. 6, pp. 99–118, 1989.

[Lester, 1976] B. M. Lester, "Spectrum analysis of the cry sounds of well-nourished and malnourished infants," *Child Development*, vol. 47, pp. 237–241, March 1976.

[Lester, 1984] B. M. Lester, "A biosocial model of infant crying," in *Advances in Infancy Research* (L. P. Lipsitt, ed.), vol. 3, pp. 167–212, Norwood, New Jersey: Ablex Publishing Corporation, 1984.

[Leung and Zue, 1988] H. C. Leung and V. W. Zue, "Some phonetic recognition experiments using neural nets," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 422–425, 1988.

[Leung, 1989] H. C. Leung, *The Use of Artificial Neural Networks for Phonetic Recognition*. PhD thesis, Massachusetts Institute of Technology, 1989.

[Li *et al.*, 1993] T. Li, F. Luyuan, and K. Q.-Q. Li, "Hierarchical classification and vector quantization with neural trees," *Neurocomputing*, vol. 5, pp. 119–139, April 1993.

[Lieberman *et al.*, 1971] P. Lieberman, K. S. Harris, P. Wolff, and L. H. Russell, "Newborn infant cry and nonhuman primate vocalization," *Journal of Speech and Hearing Research*, vol. 14, pp. 718–727, December 1971.

[Lind *et al.*, 1970] J. Lind, V. Vuorenkoski, G. Rosberg, T. J. Partanen, and O. Wasz-Höckert, "Spectrographic analysis of vocal response to pain stimulus in infants with Down's syndrome," *Developmental Medicine and Child Neurology*, vol. 12, pp. 478–486, 1970.

[Lippmann and Gold, 1987] R. P. Lippmann and B. Gold, "Neural-net classifiers useful for speech recognition," in *IEEE First International Conference on Neural Networks*, vol. 4, pp. 417–422, 1987.

[Lippmann and Singer, 1993] R. P. Lippmann and E. Singer, "Hybrid neural-network/HMM approaches to wordspotting," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 565–568, IEEE, 1993.

[Lippmann, 1987] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, vol. 4, pp. 4–22, April 1987.

[Lundh, 1986] P. Lundh, "A new baby-alarm based on tenseness of the cry signal," *Scandinavian Audiology*, vol. 15, no. 4, pp. 191–196, 1986.

[Lunji *et al.*, 1993] Q. Lunji, K. Soo-Ngee, and Y. Haiyun, "Pitch determination of noisy speech using wavelet transform in time and frequency domains," pp. 337–340, 1993.

[Maikler, 1991] V. E. Maikler, "Effects of a skin refrigerant/anesthetic and age on the pain responses of infants receiving immunizations," *Research in Nursing and Health*, vol. 14, pp. 397–403, December 1991.

[Makhoul, 1973] J. Makhoul, "Spectral analysis of speech by linear prediction," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, pp. 140–148, June 1973.

[Maksym, 1973] J. N. Maksym, "Real-time pitch extraction by adaptive prediction of the speech waveform," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, pp. 149–154, June 1973.

[Markel and Gray Jr., 1976] J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech*. Berlin: Springer-Verlag, 1976.

[Markel, 1972a] J. D. Markel, "Digital inverse filtering - a new tool for formant trajectory estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. 20, pp. 129–137, June 1972.

[Markel, 1972b] J. D. Markel, "The sift algorithm for fundamental frequency estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. 20, pp. 367–377, December 1972.

[Markel, 1973] J. D. Markel, "Application of a digital inverse filter for automatic formant analysis," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, pp. 154–160, June 1973.

[Martin, 1982] P. Martin, "Comparison of pitch detection by cepstrum and spectral comb analysis," pp. 180–183, IEEE, 1982.

[Martínez-Alfaro and Contreras-Vidal, 1991] H. Martínez-Alfaro and J. L. Contreras-Vidal, "A robust real-time pitch detector based on neural networks," (Toronto, Canada), pp. 512–523, IEEE, April 14-17 1991.

[Mat, 1992] The Math Works Inc., Natic, Mass., *MATLAB High-Performance Numeric Computation and Visualization Software: User's Guide*, 1992.

[Matausek and Batalov, 1980] M. R. Matausek and V. S. Batalov, "A new approach to the determination of the glottal waveform," vol. 28, pp. 616–622, December 1980.

[McAulay and Qualtieri, 1986] R. J. McAulay and T. F. Qualtieri, "Speech analysis/synthesis based on a sinusoidal representation," vol. 34, pp. 744–754, August 1986.

[McAulay and Quatieri, 1990] R. J. McAulay and T. F. Quatieri, "Pitch detection and voicing detection based on a sinusoidal speech model," (Albuquerque, New Mexico), pp. 249–252, IEEE, April 3-6 1990.

[McCandless, 1974] S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," vol. 22, pp. 135–141, April 1974.

[McCulloch and Pitts, 1943] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, 1943.

[Medan and Yair, 1989] Y. Medan and E. Yair, "Pitch synchronous spectral analysis scheme for voiced speech," vol. 37, pp. 1321–1328, September 1989.

[Medan *et al.*, 1991] Y. Medan, E. Yair, and D. Chazon, "Super resolution pitch determination of speech signals," vol. 39, pp. 40–48, January 1991.

[Michelsson and Sirvio, 1976] K. Michelsson and P. Sirvio, "Cry analysis in congential hypothyroidism," *Folia Phoniatrica*, vol. 28, no. 1, pp. 40–47, 1976.

[Michelsson and Wasz-Höckert, 1980] K. Michelsson and O. Wasz-Höckert, "The value of cry analysis in neonatology and early infancy," in *Infant Communication: Cry and Early Speech* (T. Murry and J. Murry, eds.), ch. 7, pp. 152–182, Houston, TX: College-Hill Press, 1980.

[Michelsson *et al.*, 1974] K. Michelsson, P. Sirvio, M. Koivisto, and O. Wasz-Höckert, "Comparison of pain cry in neonates with and without feeding tube," *Developmental Medicine and Child Neurology*, vol. 16, p. 397, June 1974.

[Michelsson *et al.*, 1977a] K. Michelsson, P. Sirvio, and O. Wasz-Höckert, "Pain cry in full-term asphyxiated newborn infants correlated with late findings," *Acta Paediatrica Scandinavica*, vol. 66, pp. 611–616, September 1977.

[Michelsson *et al.*, 1977b] K. Michelsson, P. Sirvio, and O. Wasz-Höckert, "Sound spectrographic cry analysis of infants with bacterial meningitis," *Developmental Medicine and Child Neurology*, vol. 19, pp. 309–315, June 1977.

[Michelsson *et al.*, 1980] K. Michelsson, N. Tuppurainen, and P. Aula, "Cry analysis of infants with karyotype abnormality," *Neuropediatrics*, vol. 11, pp. 365–376, November 1980.

[Michelsson *et al.*, 1984] K. Michelsson, H. Kaskinen, R. Aulanko, and A. Rinne, "Sound spectrographic cry analysis of infants with hydrocephalus," *Acta Paediatrica Scandinavica*, vol. 73, pp. 65–68, January 1984.

[Michelsson *et al.*, 1990] K. Michelsson, A. Rinne, and S. Paajanen, "Crying, feeding and sleeping patterns in 1 to 12-month-old infants," *Child: Care, Health and Development*, vol. 16, pp. 99–111, March-April 1990.

[Michelsson, 1980] K. Michelsson, "Cry characteristics in sound spectrographic analysis," in *Infant Communication: Cry and Early Speech* (T. Murry and J. Murry, eds.), ch. 4, pp. 85–105, Houston, TX: College-Hill Press, 1980.

[Miyoshi *et al.*, 1986] Y. Miyoshi, K. Yamato, M. Yanagida, and O. Kakusho, "Analysis of speech signals of short pitch period by the sample-selective linear prediction," pp. 1245–1248, 1986.

[Morgan and Scofield, 1991] D. P. Morgan and C. L. Scofield, *Neural Networks and Speech Processing*. The Kluwer international series in engineering and computer science, Norwell, Massachusetts: Kluwer Academic Publishers, 1991.

[Murayama *et al.*, 1991] K. Murayama, R. S. Greenwood, K. W. Rao, and A. S. Aylsworth, "Neurological aspects of del(1q) syndrome," *American Journal of Medical Genetics*, vol. 40, pp. 488–492, September 1991.

[Murray, 1985] A. D. Murray, "Aversiveness is in the mind of the beholder: Perception of infant crying by adults," in *Infant Crying: Theoretical and Research Perspectives* (B. M. Lester and C. F. Z. Boukydis, eds.), ch. 10, pp. 217–240, New York, New York: Plenum Press, 1985.

[Murry and Murry, 1980] T. Murry and J. Murry, eds., *Infant Communication: Cry and Early Speech*. Houston, TX: College-Hill Press, 1980.

[Nadeu *et al.*, 1991] C. Nadeu, J. Pascual, and J. Hernando, "Pitch determination using the cepstrum of the one-sided autocorrelation sequence," (Toronto, Canada), pp. 3677–3680, IEEE, April 14-17 1991.

[Nakamura and Sawai, 1992] S. Nakamura and H. Sawai, "Performance comparison of neural network architectures for speaker-independent phoneme recognition," *Systems and Computers in Japan*, vol. 23, no. 14, pp. 72–83, 1992.

[Nguyen *et al.*, 1990] T. K. P. Nguyen, R. P. Lippman, B. Gold, and D. P. Paul, "A physiologically motivated front-end for speech recognition," *International Joint Conference on Neural Networks*, pp. II–503–508, June 1990.

[Nikias and Mendel, 1993] C. L. Nikias and J. M. Mendel, "Signal processing with higher-order spectra," *IEEE Signal Processing Magazine*, vol. 10, pp. 10–37, July 1993.

[Niles *et al.*, 1989] L. Niles, H. Sliverman, G. Tajchman, and M. Bush, "How limited training data can allow a neural network to outperform an 'optimal' statistical classifier," (Glasgow, Scotland), pp. 17–20, IEEE, May 1989.

[Noll, 1967] A. M. Noll, "Cepstrum pitch determination," vol. 41, no. 2, pp. 293–309, 1967.

[Oppenheim and Schafer, 1975] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1975.

[Oppenheim, 1970] A. V. Oppenheim, "Speech spectrogram using the fast Fourier transform," *IEEE Spectrum*, pp. 57–62, August 1970.

[O'Shaughnessy, 1987] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Reading, Massachusetts: Addison-Wesley Publishing Company, 1987.

[Ostrea Jr. *et al.*, 1975] E. M. Ostrea Jr., C. J. Chavez, and M. E. Strauss, "A study of factors that influence the severity of neonatal narcotic withdrawal," *Addictive Diseases: an International Journal*, vol. 2, no. 1–2, pp. 187–199, 1975.

[Ostwald and Murry, 1985] P. F. Ostwald and T. Murry, "The communicative and diagnostic significance of infant sounds," in *Infant Crying: Theoretical and Research Perspectives* (B. M. Lester and C. F. Z. Boukydis, eds.), ch. 7, pp. 139–158, New York, NY: Plenum Press, 1985.

[Ostwald *et al.*, 1968] P. F. Ostwald, R. Phibbs, and S. Fox, "Diagnostic use of infant cry," *Biologia Neonatorum*, vol. 13, no. 1, pp. 68–82, 1968.

[Parmelee, 1962] A. Parmelee, "Infant crying and neurological diagnosis," *Journal of Pediatrics*, vol. 61, pp. 801–802, 1962.

[Partanen *et al.*, 1967] T. J. Partanen, O. Wasz-Höckert, V. Vuorenkowski, K. Theorell, E. H. Valanne, and J. Lind, "Auditory identification of pain cry signals of young infants in pathological conditions and its sound spectrographic basis," *Annales Paediatriae Fenniae*, vol. 13, no. 2, pp. 56–63, 1967.

[Petroni *et al.*, 1994a] M. Petroni, A. S. Malowany, C. C. Johnston, and B. J. Stevens, "A new, robust vocal fundamental frequency determination method for the analysis of infant cries," in *Proceedings of the 1994 IEEE Seventh Symposium on Computer-Based Medical Systems*, (Winston-Salem, NC), pp. 223–228, IEEE, June 10-12 1994.

[Petroni *et al.*, 1994b] M. Petroni, A. S. Malowany, C. C. Johnston, and B. J. Stevens, "A crosscorrelation-based method for improved visualization of infant cry vocalizations," in *1994 Canadian Conference on Electrical and Computer Engineering*, (Halifax, Nova Scotia), pp. 453–456, September 1994.

[Petroni *et al.*, 1995] M. Petroni, A. S. Malowany, C. C. Johnston, and B. J. Stevens, "Classification of infant cry vocalizations using artificial neural networks (ANNs)," (Detroit, MI), IEEE, May 1995.

[Porter *et al.*, 1986] F. L. Porter, R. H. Miller, and R. E. Marshall, "Neonatal pain cries: Effect of circumcision on acoustic features and perceived urgency," *Child Development*, vol. 57, pp. 790–802, 1986.

[Prescott, 1975] R. Prescott, "Infant cry sound; developmental features," *Journal of the Acoustical Society of America*, vol. 75, pp. 1186–1191, May 1975.

[Proakis and Manolakis, 1988] J. G. Proakis and D. G. Manolakis, *Introduction to Digital Signal Processing*. New York, New York: Macmillan Publishing Company, 1988.

[Rabiner and Schafer, 1975] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs: Prentice-Hall Inc., 1975.

[Rabiner *et al.*, 1969] L. R. Rabiner, R. W. Schafer, and C. M. Rader, "The chirp z-transform algorithm and its application," *The Bell System Technical Journal*, pp. 1249–1292, May-June 1969.

[Rabiner *et al.*, 1976] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance of several pitch detection algorithms," vol. 24, pp. 399–418, October 1976.

[Rabiner, 1977] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," vol. 25, pp. 24–33, February 1977.

[Rabiner, 1989] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications," *Proceedings of the IEEE*, vol. 77, pp. 257–286, February 1989.

[Reddy and Swamy, 1984] N. S. Reddy and M. N. S. Swamy, "High-resolution formant extraction from linear prediction phase spectra," vol. 32, pp. 1136–1144, December 1984.

[Robb *et al.*, 1989] M. P. Robb, J. H. Saxman, and A. A. Grant, "Vocal fundamental frequency characteristics during the first two years of life," vol. 85, pp. 1708–1717, April 1989.

[Robinson *et al.*, 1993] T. Robinson, L. Almeida, J.-M. Boite, H. Bourlard, F. Fallside, M. Hochberg, D. Kershaw, P. Kohn, Y. Konig, S. Renals, M. Saerens, J. P. Neto, N. Morgan, and C. Wooters, "A neural network based, speaker independent, large vocabulary, continuous speech recognition system: The WERNICKE project," in *Proceedings of the European Conference on Speech Technology*, 1993.

[Robinson, 1992] T. Robinson, "A real-time recurrent error propagation network word recognition system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, (New York, NY, USA), pp. 617–20, IEEE, 1992.

[Ross *et al.*, 1974] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," vol. 22, pp. 353–362, October 1974.

[Sakoe *et al.*, 1989] H. Sakoe, R. Isotani, K. Yoshida, K. Iso, and T. Watanabe, "Speaker-independant word recognition using dynamic programming neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 29–32, May 1989.

[Scofield *et al.*, 1988] C. L. Scofield, D. L. Reilly, C. Elbaum, and L. N. Cooper, "Pattern class degeneracy in an unrestricted storage density memory," in *Neural Information Processing Systems* (D. Z. Anderson, ed.), pp. 674–682, Denver, CO: American Institute of Physics, 1988.

[Seneff, 1978] S. Seneff, "Real-time harmonic pitch detector," vol. 26, pp. 358–365, August 1978.

[Seneff, 1984] S. Seneff, "Pitch and spectral estimation of speech based on auditory synchrony model," pp. 36.2.1–36.2.4, 1984.

[Shamma, 1987] S. A. Shamma, "Neural networks for speech processing and recognition," in *IEEE First International Conference on Neural Networks*, vol. 4, pp. 397–405, 1987.

[Sherman, 1927] M. Sherman, "The differentiation of emotional responses in infants," *Journal of Comparative Psychology*, vol. 7, pp. 265–284, 1927.

[Simpson, 1990] P. K. Simpson, *Artificial Neural Systems: Foundations, Paradigms, Applications and Implementations.* Neural Networks: Research and Applications, New York: Pergamon Press, 1990.

[Slaney and Lyon, 1990] M. Slaney and R. F. Lyon, "A perceptual pitch detector," (Albuquerque, NM), pp. 357–360, IEEE, 1990.

[Slaney, 1994] M. Slaney, "Auditory toolbox: A Matlab toolbox for auditory modelling work," Tech. Rep. 45, Apple Computer Inc., 1994.

[Sluyter *et al.*, 1982] R. J. Sluyter, H. J. Kotmans, and T. A. C. M. Claasen, "Improvements of the harmonic-sieve pitch extraction scheme and an appropriate method for voiced-unvoiced detection," pp. 188–191, IEEE, 1982.

[Sondhi, 1968] M. M. Sondhi, "New methods of pitch extraction," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, pp. 262–266, June 1968.

[Stallard and Juberg, 1981] R. Stallard and R. C. Juberg, "Partial monosomy 7q syndrome due to distal interstitial deletion," *Human Genetics*, vol. 57, no. 2, pp. 210–213, 1981.

[Stark and Nathanson, 1973] R. E. Stark and S. N. Nathanson, "Spontaneous cry in the newborn infant; sounds and facial gestures," *Symposium on Oral Sensation and Perception*, no. 4, pp. 323–352, 1973.

[Stark and Nathanson, 1975] R. E. Stark and S. N. Nathanson, "Unusual features of cry in an infant dying suddenly and unexpectedly," in *Development of upper respiratory anatomy and function* (J. F. Bosma and J. Showacre, eds.), pp. 233–249, Bethesda, MD: NIH, 1975.

[Stevens *et al.*, 1994] B. J. Stevens, C. C. Johnston, and L. Horton, "Factors that influence the behavioral pain responses of premature infants," *Pain*, vol. 59, pp. 101–109, October 1994.

[Tenold *et al.*, 1974] J. L. Tenold, D. H. Crowell, R. H. Jones, T. H. Daniel, D. F. McPherson, and A. N. Popper, "Cepstral and stationary analyses of full-term and premature infants' cries," vol. 56, pp. 975–980, September 1974.

[Tetko *et al.*, 1994] I. V. Tetko, V. Tanchuk, N. P. Chentsova, S. V. Antonenko, G. Poda, V. Kukhar, and A. I. Luik, "HIV-1 reverse transcriptase inhibitor design using artificial neural networks," *Journal of Medicinal Chemistry*, vol. 37, pp. 2520–2526, August 1994.

[Thodén and Koivisto, 1980] C.-J. Thodén and M. Koivisto, "Acoustic analysis of the normal pain cry," in *Infant Communication: Cry and Early Speech* (T. Murry and J. Murry, eds.), ch. 6, pp. 124–151, Houston, TX: College-Hill Press, 1980.

[Thodén and Michelsson, 1979] C. J. Thodén and K. Michelsson, "Sound spectrographic cry analysis in Krabbe's disease," *Developmental Medicine and Child Neurology*, vol. 21, pp. 400–402, June 1979.

[Thodén *et al.*, 1985] C.-J. Thodén, A.-L. Järvenpää, and K. Michelsson, "Sound spectrographic cry analysis of pain cry in prematures," in *Infant Crying : Theoretical and Research Perspectives*, ch. 5, New York, New York: Plenum Press, 1985.

[Truby and Lind, 1965] H. M. Truby and J. Lind, "Cry sounds of the newborn infant," in *Newborn Infant Cry* (J. Lind, ed.), 1965.

[Turner *et al.*, 1978] H. D. Turner, E. M. Brett, R. J. Gilbert, A. C. Ghosh, and H. J. Liebeschuetz, "Infant botulism in England," *Lancet*, vol. 1, pp. 1277–1278, June 1978.

[van Camp, 1993] D. van Camp, *A Users Guide for the Xerion Neural Network Simulator Version 3.1*. Department of Computer Science, University of Toronto, Toronto, Canada, May 1993.

[Veeneman and BeMent, 1984] D. E. Veeneman and S. L. BeMent, "Automatic glottal inverse filtering," pp. 36.5.1–36.5.4, IEEE, 1984.

[Vuorenkoski *et al.*, 1970] V. Vuorenkoski, M. Kaunisto, P. Tjernlund, and L. Vesa, "Cry detector. A clinical apparatus for surveillance of pitch and activity in the crying of a newborn infant," *Acta Paediatrica Scandinavica - Supplement*, vol. 206, p. 103, 1970.

[Waibel *et al.*, 1987] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," Tech. Rep. TR-1-0006, ATR Interpreting Telephony Research Laboratories, October 1987.

[Waibel *et al.*, 1988] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition: Neural networks vs. hidden Markov models," pp. 107–110, April 1988.

[Waibel *et al.*, 1989] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," vol. 37, pp. 328–339, 1989.

[Wasz-Höckert *et al.*, 1968] O. Wasz-Höckert, J. Lind, V. Vuorenkoski, T. Partanen, and E. Valanne, *The Infant Cry: A Spectrographic and Auditory Analysis*. Lavenham, U.K.: Spastics International Medical Publications, 1968.

[Wasz-Höckert *et al.*, 1985] O. Wasz-Höckert, K. Michelsson, and J. Lind, "Twenty-five years of Scandinavian cry research," in *Infant Crying: Theoretical and Research Perspectives*, ch. 4, New York, New York: Plenum Press, 1985.

[Watrous and Shastri, 1987] R. L. Watrous and L. Shastri, "Learning phonetic features using connectionist networks: An experiment in speech recognition," in *Proceedings of the IEEE First International Conference on Neural Networks*, vol. 4, (San Diego, CA), pp. 619–627, IEEE, June 1987.

[Weisenfeld *et al.*, 1981] A. Weisenfeld, C. Zander Malatesta, and L. De Loach, "Differential parental responses to familiar and unfamiliar infant distress signals," *Infant Behavior and Development*, vol. 4, no. 3, p. 281, 1981.

[Weiss and Kulikowski, 1991] S. M. Weiss and C. A. Kulikowski, *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. San Mateo, CA: Morgan Kaufmann Publishers Inc., 1991.

[Werbos, 1974] P. J. Werbos, *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University, 1974.

[Wise *et al.*, 1976] J. D. Wise, J. R. Caprio, and T. W. Parks, "Maximum likelihood pitch estimation," vol. 24, pp. 418–423, October 1976.

[Wu and Chan, 1993] J. Wu and C. Chan, "Isolated word recognition by neural network models with cross-correlation coefficients for speech dynamics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 1174–1185, November 1993.

[Xie *et al.*, 1993] Q. Xie, R. K. Ward, and C. A. Laszlo, "Determining normal infants' level-of-distress from cry sounds," in *Proceedings of the 1993 Canadian Conference on Electrical and Computer Engineering*, (Vancouver, B.C.), pp. 1094–1096, September 14–17 1993.

[Yamauchi *et al.*, 1993] K. Yamauchi, M. Fukuda, and K. Fukushima, "Speech recognition system consisting of auditory feature extracting cells and velocity-controlled delay-lines part II: Recognition module," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, pp. 259–262, IEEE, 1993.

[Ye *et al.*, 1990] H. Ye, S. Wang, and F. Robert, "A PCMN neural network for isolated word recognition," *Speech Communication*, vol. 9, pp. 141–53, April 1990.

[Zell *et al.*, 1994] A. Zell, N. Mache, R. Hübner, G. Maimer, M. Vogt, K.-U. Herrmann, M. Schmalzl, T. Sommer, A. Hatzigeorgiu, S. Döring, and D. Posselt, *Stuttgart Neural Network Simulator User's Manual, Version* 3.2. University of Stuttgart, Institute for Parallel and Distributed High Performance Systems, Stuttgart, Germany, 1994.

[Zeskind and Lester, 1978] P. S. Zeskind and B. M. Lester, "Acoustic features and auditory perception of the cries of newborns with prenatal and perinatal complications," *Child Development*, vol. 49, pp. 580–589, 1978.

[Zeskind *et al.*, 1985] P. S. Zeskind, J. Sale, M. L. Maio, L. Huntington, and J. R. Weisman, "Adult perceptions of pain and hunger cries: a synchrony of arousal," *Child Development*, vol. 56, pp. 549–554, June 1985.

[Zeskind, 1985] P. S. Zeskind, "Developmental perspective of infant crying," in *Infant Crying: Theoretical and Research Perspectives* (B. M. Lester and C. F. Z. Boukydis, eds.), ch. 8, pp. 159–185, New York, New York: Plenum Press, 1985.

[Zhu *et al.*, 1993] C. Zhu, L. Li, C. Guan, and Z. He, "A study of LVQ-based architectures for robust speech recognition," in *Proceedings of the World Congress on Neural Networks*, (Hillsdale, New Jersey), pp. IV–177–180, INNS, Lawrence Erlbaum Associates, 1993.

[Zwicker, 1961] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," vol. 33, p. 248, February 1961.