CO-CHANNEL SPEAKER SEPARATION

THESIS

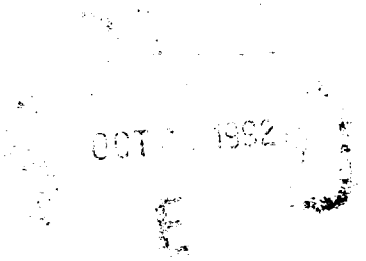
Thomas S. Andrews
Captain, USAF

AFIT/GE/ENG/92S-07

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GE/ENG/92S-07



CO-CHANNEL SPEAKER SEPARATION

THESIS

Thomas S. Andrews
Captain, USAF

AFIT/GE/ENG/92S-07

Approved for public release; distribution unlimited

AFIT/GE/ENG/92S-07

CO-CHANNEL SPEAKER SEPARATION

THESIS

Presented to the Faculty of the School of Engineering

of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the

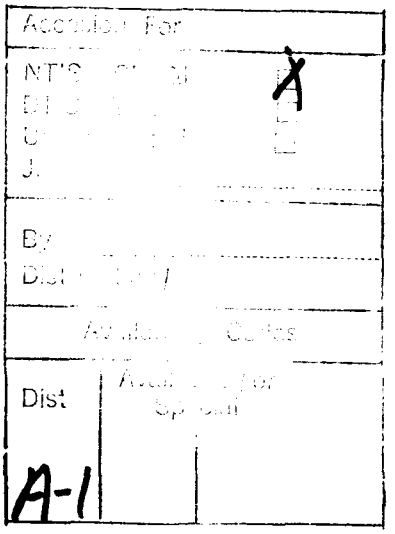
Requirements for the Degree of

Master of Science in Electrical Engineering

Thomas S. Andrews, B.S.E.E.

Captain, USAF

September, 1992



Approved for public release; distribution unlimited

Acknowledgments

Throughout my AFIT graduate course work, I did not have the opportunity to learn or study speech signals. So the task of "co-channel speaker separation" for my thesis topic was interesting and very challenging. I was attracted to this research since the problem has not completely been solved, and I attempted to help solve a small portion of the problem. Solving the co-channel speaker separation problem will have important benefits not only to the Air Force, but to the civilian community as well.

I could not have done this work alone. I want to thank my good friends Capt Jim Geurts and Capt Mark Dewitt not only their excellent technical suggestions, but also for their good humor and support they provided me along the way. I'd also like to thank Dave Morgan and Leonard Lee at Lockheed-Sanders for their Matlab code that was used as an integral basis of my work. I am grateful for the suggestions, encouragement, and direction provided by Major (Dr.) Steve Rogers and Dr. Mark Oxley. I would especially like to thank my advisor Capt (Dr.) Dennis Ruck for instilling in me the desire to excel, and for his patience and technical guidance. I could not have accomplished any of my computer work without the help of Mr. Dan Zambon and Mr. Dave Doak. They kept the Sun Sparc networks "up" and solved many system oriented problems that I encountered (or created). I feel fortunate to have had the opportunity to work with such an enthusiastic and knowledgeable group of people, and I will favorably reminisce my 15 months at AFIT.

This has also been a trying 15 months for me and my family. I want to thank my wife Annette and my son Joshua. I look forward to being a "real" husband and father to them in the very near future. I could not have accomplished any of my work without Annette and Joshua's support, encouragement, and understanding.

Thomas S. Andrews

Table of Contents

List of Figures

List of Tables

Abstract

In the co-channel speaker separation problem, the goal is to recover two separate speech signals from a monaural channel which contains the sum of the two speech signals. A new methodology is developed that if given that a segment of co-channel speech is separated into a "stronger" and "weaker" segment, the correct assignment of these separated segments to the appropriate talker can be made using a Linear Predictive Coding (LPC) based minimum-prediction residual computation. The uniqueness of the developed technique is that no *a priori* information is required of the co-channel speech signal. The information needed to appropriately assign these separated segments from the co-channel speech signal are "clean" speech that is separate from the co-channel speech signal that are used to compute model LPC vectors. This "clean" speech is derived from the same channel that the co-channel speech signal is derived from. This technique has shown the ability to correctly assign the given "stronger" and "weaker" segments to the appropriate talker at signal-to-signal ratios down to equal power levels. The resulting separated speech is clearly understandable, and the interfering talker's speech signal is effectively eliminated.

CO-CHANNEL SPEAKER SEPARATION

I. Introduction

In a communications system, the goal is to transfer information from point A to point B intelligibly. In the transfer of this information, as it applies to this thesis, the transmitted signal is a speech signal and it invariably becomes corrupted by noise and/or other interfering speech signals. Thus, the aim of a communications system designer is to minimize the interference and maximize the intelligibility of the received speech signal. The focus of this research is the enhancement of the intelligibility of the received corrupted speech signal.

1.1 Background

A speech signal becomes corrupted by noise and by other speech signals when these corrupted speech signals simultaneously occupy the same frequency band. Broadcast communications occur at a carrier frequency, and the speech signal is modulated on this carrier frequency. Another speech signal from a different transmitter at the same carrier frequency may interfere with the speech signal of interest.

As previously mentioned, several speech signals may occupy the same frequency space (bandwidth) at the same time. This bandwidth co-occupation may occur unintentionally via cross-talk in a communications system, or it may occur if the speech signals of interest were combined (corrupted) before transmission through the communications system. This corrupted speech signal, regardless of its origin, is referred to as a **co-channel speech signal.**

The corruption of a speech signal and subsequent co-channel speech signal is evident in many Air Force applications. Signals Intelligence (SIGINT) operators may encounter a co-channel speech signal in normal intercept operations or an air traffic controller could receive communications from two or more aircraft on the same frequency simultaneously. Land-line

cable communication links may experience cross-talk interference which would create a co-channel speech signal. Improper sampling in a time-division multiplexed circuit may also lead to a co-channel speech signal. For this application, as it applies to this thesis, the origin of the co-channel signal is not of importance, rather the resulting baseband co-channel speech signal is analyzed.

1.2 Problem Statement and Scope

A co-channel speech signal contains the *desired speech signal*, a *corrupting speech signal*, and noise. The process by which the desired speech signal is separated from the corrupting speech signal within a co-channel signal is known as the **co-channel speaker separation problem**. The terms *signal, speaker*, and *talker* are interchangeable in this thesis. The term *desired signal* is synonymous with either *desired talker* or *target speaker*, and the term *corrupting signal* is synonymous with *interfering talker*. "Clean" or "clear" speech denotes a single talker's speech signal that has a high signal-to-noise ratio (SNR).

The **co-channel speaker separation problem** is defined as the desire to extract the *target speaker* from the co-channel signal and make this target speaker more *intelligible*. Perfect separation of the target speaker from the corrupting signals in a co-channel interference situation is virtually impossible, but the post processing is a measure of the success of any co-channel speaker separation system. The academic co-channel speaker problem involves only two speakers: the target voice and the interfering voice and the co-channel signal has a high SNR. A more practical situation of a co-channel signal would include the addition of significant noise to the speech waveforms. The thrust of the co-channel speaker separation problem lies in the separation and recovery of these two (or more) vocal tract signals which are superimposed on a monophonic recording.

This thesis will explore the portion of the co-channel speaker separation problem involved with the assignment of processed speech segments to the appropriate talker. Specific limitations are outlined below under **Assumptions**. A given speech segment is processed and divided into two different parts. These two parts are differentiated by one that includes the

energy of the detected pitch and harmonics (termed "stronger" segment), and the other includes the energy of the segment not contained in the pitch frequency and harmonics (termed "weaker" segment). The focus of this research effort will be the development of a spectral assignment methodology/algorithm based on a linear predictive coding (LPC) distortion metric.

LPC coefficients are calculated from the separated segment of speech containing the stronger energy, and a resulting distortion metric is computed with a precalculated set of model LPC vectors (from both talkers) in order to base a decision rule on the assignment of the separated segments of speech. The speech used to create the set of precalculated parameters is independent of the speech in the co-channel signal. This independence is significant because no *a priori* knowledge of the co-channel signal is required. The only *a priori* information required in this co-channel separation process is "clean" speech from the desired talker or "clean" speech of both talkers if the recovery of both talkers is desired (or feasible).

Formal intelligibility measurements of the post-processed co-channel speech signals is not explored in depth in this thesis. Rather, the applicability of the proposed processing techniques will be explored and their feasibility, usefulness, and limitations will be discussed. Some measures of intelligibility of speech signals may be found in Parsons (21).

The co-channel speaker separation problem poses difficult signal processing problems. One problem lies in the non-stationary property of speech signals and another problem is the lack of unique features in the co-channel speech signal that identifies one talker from another. The major feature used to separate the two talkers is the pitch of their voiced speech Assuming a segment of co-channel speech can be separated, the assignment of the segments to the appropriate talker is difficult.

Co-channel speaker separation has useful applications but the state of the solution to the co-channel speaker separation problem is still in the feasibility/development stage. This research effort will attempt to improve the current co-channel speaker separation capabilities by applying an LPC based distortion metric in the spectral assignment portion of the process.

1.3 Assumptions

The co-channel speaker separation problem addressed in this thesis is based on the following ground rules:

- The signals of interest are analog voiced speech at baseband.
- The speech signals are additively combined to make the co-channel signal.
- Noise in the co-channel signal is modelled as additive white Gaussian noise (AWGN) and the noise is statistically independent of the speech signals.
- An *a priori* data set of clear speech for each talker is given. This *a priori* clear speech set is **independent** of the speech in the co-channel signal. By **independent**, the *a priori* data set is spoken from the same speaker, but the sentences are different from the sentence spoken in the co-channel speech signal.
- The co-channel speech signal exists on a monaural channel.
- There are only two separate speakers in the co-channel signal.
- The co-channel signal-to-noise ratio (SNR) is sufficiently large so that the noise interference is negligible.

1.4 General Approach

The general approach taken in this research is broken down to a four steps. The first step is to create a speech signal sample set from the TIMIT (2) speech database and from recorded speech signals (discussed in Appendix A). This step also involves the creation of the co-channel signal and the "clean" speaker subsets (used for precalculating the model LPC vectors). In the second step, software algorithms are developed to process the speech signals. Partial use of the algorithms described by L. Lee and Morgan (11) will be augmented by code developed as part of this research effort (Chapter III). As a third step, the co-channel speech signals are processed with varying signal-to-signal ratios (SSRs). The developed techniques are compared against previous co-channel signal processing methodologies. The results are analyzed in the final step, and improvements/limitations to the algorithms are discussed.

1.5 Resources

The resources required for this research include the digitized speech data, processing software, and a digital computer. These items are all available at AFIT. The speech signals used in this thesis originate from the TIMIT database or recorded speech signals. Matlab will be used to implement the co-channel speaker separation algorithm on the Sparc-2 Sun workstations. Further descriptions of the speech database files and software and hardware tools used in this thesis are provided in the appendices.

1.6 Organization

This chapter provided a brief description of the co-channel speaker separation problem, the research objectives, assumptions, resources needed, and the general approach taken in this thesis effort. Chapter II provides the background information and current aspects of other co-channel speaker separation techniques which build the foundation for the algorithms developed in the next chapter. Chapter III discusses the co-channel speaker separation algorithms developed in this thesis. In Chapter IV, the results of testing the various co-channel speaker signals on the developed algorithms will be discussed. Limitations to the algorithms will be identified, and alternate processing algorithms will be discussed and implemented. The last chapter provides conclusions and recommendations for further research.

II. Literature Review

2.1 Overview

This chapter discusses the previous research and methodologies in co-channel speaker separation. A brief review of the technical issues surrounding a co-channel speech signal are presented. A historical review of the different methodologies that have been developed will be presented, culminating with the current state-of-the-art techniques. The chapter concludes with a synopsis of the techniques employed in this thesis.

2.2 Co-Channel Speaker Separation Technical Issues

Most of the co-channel speech separation algorithms that have been developed have implemented similar signal processing nodes. The differences between the applications lie in the approach taken in the particular individual processing steps. A typical co-channel speaker separation process is illustrated in Figure 2.1.

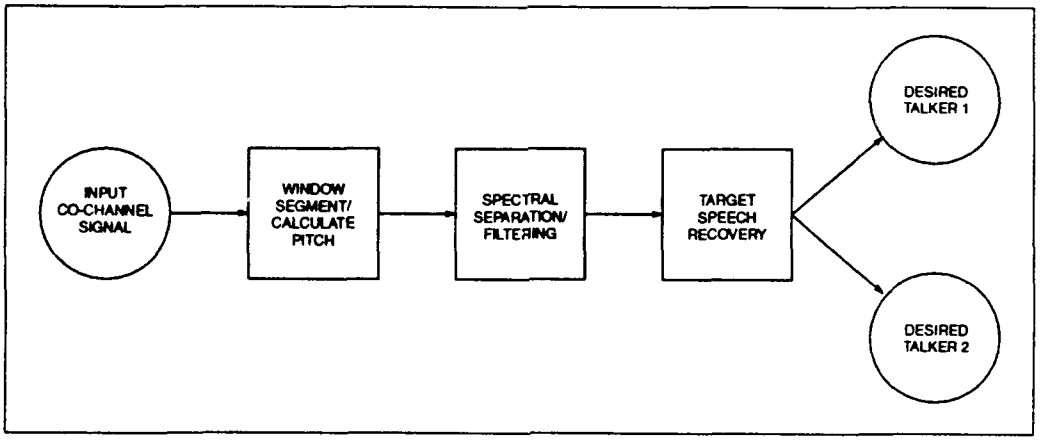


Figure 2.1. A Typical Co-Channel Speaker Separation Process

A description of various signal processing techniques that are considered in many co-channel speaker separation algorithms include portions of the following:

1. Sampling and digitizing the analog speech waveform (at or above the Nyquist criteria).
2. Segmenting the co-channel signal into small stationary analysis windows. The window size is selected to be as large as possible while keeping the assumption of the segmented speech signal as stationary.
3. Determining the pitch and harmonic values in a voiced segment of speech.
4. Filtering, sampling, or digital signal processing the segment of speech in order to separate the energy of the two talkers, and/or suppressing the unwanted interfering speech signals.
5. Recovering/synthesizing the desired talker's speech waveform.
6. Using speaker identification techniques, where applicable.

The co-channel speech signal examined in this thesis contains three signals: talker *X*, talker *Y*, and noise. The signal-to-noise ratio (SNR) is a measurement of the signal power of talker *X* and talker *Y* (singularly or combined) to the noise power. A similar measurement called the signal-to-signal ratio (SSR) is the measure of the signal power of talker *X* to talker *Y*, and likewise the voiced-to-voiced ratio (VVR) is the measure of the signal power of only the voiced regions of talker *X* to talker *Y* (11). In this thesis the SNR and SSR will be used. The SNR is usually measured from a speech signal if the speech signal has portions that are silence (noise only). The additive white Gaussian noise (AWGN) power can be calculated during this silence (noise only). The SNR found by the following equations:

$$\text{Signal Power} = \frac{\text{Signal Power + Noise Power}}{\text{Total Signal Power}} - \frac{\text{Noise Power}}{\text{Total Signal Power}} \qquad (2.1)$$

$$\text{SNR} = \frac{\text{Signal Power}}{\text{Noise Power}} \qquad (2.2)$$

Since the two speech waveforms in the co-channel speech signal occupy the same bandwidth, little or no pre-processing can be done to the co-channel speech signal to separate the two talkers. The input signal may be low-pass filtered at around 4 kHz (typical of standard telephone bandwidth) to remove any high frequency noise or speech components. This low-pass filtering also helps prevent aliasing in digitizing the co-channel signal and removes unwanted high frequency noise and interfering waveforms. The speech signal is usually digitized prior to the co-channel processing algorithm. Once the co-channel speech signal is digitized, it naturally lends itself to discrete time applications and processing. In digitizing the speech waveform consideration of the Nyquist criteria must be adhered to. Also employing a digitizer with a high number of quantization levels and ensuring the speech waveform spans the dynamic range of the quantizer will enable the speech waveform to be digitized and reconstructed (A/D $\rightarrow$ D/A) with minimal distortion.

The next application in co-channel speech processing typically involves segmenting the digitized speech signal with a window function other than a basic "rect" function. The windowing function, e.g. Hamming, Hanning, Kaiser, or raised-cosine, is used to segment the co-channel speech signal into frames. These window functions each have the characteristics of smooth transitions to zero at the ends, and the heights of the sidelobes are be diminished. These types of window functions are also employed to provide the frequency response that has generalized linear phase (19).

A speech signal in, general, is a non-stationary, random process. A random process is stationary if its statistics do not vary with time. However, some short segments ($\sim$ 20-70 msec) in speech can be considered locally stationary, while other similar length segments will remain non-stationary (21) regardless of the window size. Most speech recognition systems assume the speech signal is stationary for periods up to 70 msec (21). Assuming a windowed segment of speech is stationary, linear signal processing techniques can be employed on these windowed segments.

A segmented speech signal can be classified into three categories: voiced, unvoiced, or silence. During typical speech, the energy of any talker varies with the utterance being spoken.

The energy levels of the two talkers vary and cross at irregular intervals between segments in the co-channel speech signal. The voiced portions of speech have a characteristic periodic structure and have higher energy than unvoiced or silent segments. Unvoiced portions of speech have no structure, are noise-like, and have been modeled as AWGN.

Most co-channel speaker separation algorithms use a fixed window size and overlap adjacent data window frames. For example a segmented window might be 50 msec long and be stepped by 10 msec. For this example if the sampling rate was 10 kHz the first segment contains samples 1-500, the second segment would contain samples 101-600, etc. In this thesis a fixed sample window of 50 msec is used, this window is incremented by 10 msec, and all segments of speech are assumed stationary. It is acknowledged that this assumption (on the stationarity of the windowed segment) will not hold during the entire co-channel processing, especially in segments which are predominately unvoiced, or in transitions between voiced and unvoiced speech. It is assumed that any processing of unvoiced segments will not be detrimental to the outcome since the resulting waveform will be characteristically unvoiced and the perception of the processed speech signal will not be severely degraded.

Once the co-channel signal a been windowed, typically the time-domain signal is Fourier transformed so the spectral components can be analyzed. The detection of the pitch in this segment of co-channel speech can be performed. The knowledge of the pitch (and its harmonics) enables frequency domain signal processing techniques to be used to separate the energy of the two talkers.

The periodic structure of a voiced portion of speech is related to the "pitch." It is the determination of the fundamental pitch frequency that has received significant attention in co-channel speech analysis. Many clever pitch selection, prediction, calculation, and tracking algorithms have been developed (6). Some of the significant pitch tracking algorithms will be discussed in the following section.

Linear predicative coding (LPC) techniques have also been used in co-channel speaker separation algorithms (9, 12). LPC estimates have been used in virtually all phases of co-channel speaker separation algorithms. LPC estimators have been used in pitch tracking,

speaker identification, calculation of distortion metrics, and formant prediction. LPC techniques have also been used to synthesize/reconstruct speech.

The previously described techniques can now be used to dissect a separated segment of co-channel speech. Time domain overlap and add techniques are usually applied to the windowed and stepped signal. Post-filtering would occur that off-sets any pre-filtering that might have been used on the input co-channel speech signal.

This section discussed the technical issues concerning the co-channel speaker separation problem, and some typical signal processing applications. The next section will discuss in detail the specific co-channel speaker separation techniques that have been developed.

2.3 Co-Channel Speaker Separation Sub-process Techniques

2.3.1 **Pitch Detectors.** As noted previously, some form of pitch calculation has been used in most co-channel speaker separation algorithms. The pitch detectors have been used primarily to calculate the pitch of the present segment, plus track relative continuity with the previously calculated pitch values. The following paragraphs provide a description of some of the important pitch detection algorithms that have been developed and implemented in solving the co-channel speaker separation problem.

The maximum likelihood (ML) pitch detection algorithm has been the most widely used pitch detector (6, 11, 17, 28). The ML pitch detector works on the basis of a likelihood function for a given segment of speech. That is, the peak of the likelihood function is found for all values of the pitch in the range of interest. The advantage of the ML algorithm is that it is relatively immune to noise, can detect non-integer pitch values (by interpolating the autocorrelation function), and the estimator is based in the autocorrelation domain. This is significant since the two speakers in the co-channel signal are added in the time domain and hence are added in the autocorrelation domain. The major drawback of the ML pitch detector is in the computational burden, since an autocorrelation must be calculated for each segment for each pitch value in the range of pitch values selected of the co-channel signal ($\sim$ 250 autocorrelation calculations per segment) (28). The range of pitch values usually varies from

50 Hz - 400 Hz. McAulay examined the ML pitch detection theory and concluded that the average squared difference function is optimum and robust when the voiced speech waveform is modeled as quasi-periodic with the periodicity extending over two periods in the segment of interest (15). McAulay also claimed the ML pitch estimator was roughly equivalent to the cepstral method and successful in strong noise environments.

L. Lee and Morgan extended the ML pitch detector (described above) to work with the two speakers in the co-channel signal (11). Their algorithm detects the dominant pitch value in the segment. This pitch value is used further in their algorithm for the separation of both talkers.

The modified covariance (MC) pitch detection algorithm is a linear prediction algorithm which uses a forward and backward prediction to estimate the prediction coefficients (poles) of a system (8, 11, 13, 17). The MC method is based on minimization with respect to all the prediction coefficients. The MC method does not guarantee a stable linear prediction filter, although most of the time it will yield a stable filter. When the MC technique is applied for spectral estimation this instability condition is not a problem. For the pitch detection application a high-order predictor is used so the spectral peaks between 50-400 Hz (typical range of male/female pitch values) can be located. Naylor and Porter tested a MC modified pitch detector on a -12 dB co-channel signal. Their results were compared with *a priori* pitch tracks, and their method was able to estimate the pitch of the two talkers. The only problem encountered in this test occurred when the pitch tracks of the two talkers crossed, or when the energy of the target speaker was very low.

The cepstral (homomorphic) pitch detection algorithm begins by windowing a segment of speech, then a Fourier transform is taken on the segment, and the resulting spectrum's log magnitude is taken. Then, the inverse Fourier transform is computed, and a peak picking algorithm detects the pitch above a certain threshold (22). On clean speech, this method works quite well. Paul showed that the homomorphic pitch detector was prone to error in the voiced/unvoiced transition regions. Paul compared his homomorphic pitch detector to one developed by Gold-Rabiner. Paul found that his method and the Gold-Rabiner method

performed similarly for male speakers, but the homomorphic algorithm made less errors for female speakers. Paul also examined the performance of the homomorphic algorithm to the Gold-Rabiner algorithm with speech corrupted by noise. He found that noise degraded the Gold-Rabiner algorithm more than the homomorphic algorithm. A SNR level of about 10 dB made the Gold-Rabiner method yield bad "chopped-up" pitch while the homomorphic method was unaffected at this SNR level.

Dick suggested a method to compute the pitch based on the complex correlation (3). This method took advantage of the pitch value and several of its harmonics. The complex correlation is found by taking a Fast Fourier Transform (FFT) of a windowed segment of speech data, and then taking the magnitude squared. The negative frequency components of the transform are set to zero, and the square root of the positive frequency components are computed. The resulting data is inverse fast Fourier transformed and the result is the complex correlation. The pitch is determined from selecting the peak of this function. This method was derived for the best fit for the implementation of a comb filter on co-channel speech. Dick claimed his method was robust against noise and distortion.

Quatieri and Daniesewicz developed a pitch tracking routine that is fundamentally different from the above techniques (23). Their technique involves tracking both fundamental frequencies in the co-channel speech signal. One of their primary assumptions was that the segment of speech was vocalic, and the pitch tracks between segments did not vary significantly. Given the harmonic assumption of the segment, two fundamental frequencies can be tracked in time by using calculated estimates on each analysis frame as initial estimates on subsequent frames. An iterative method of steepest descent is used for updating the pitch estimates for each frame.

Another pitch tracking algorithm was developed by Min *et al* (16). This method used a look-forward and look-backward technique to track the previous pitch values and to predict the future pitch values. A unique feature of this algorithm was the measurement of the pitch values for both speakers. This algorithm was in the refinement stage, but Min *et al* claimed this technique was useful and effective (no quantifying results provided). They later reported

that this pitch tracking technique had difficulty computing the pitch during times where the pitch tracks crossed or during extended periods of silence (1).

Medan *et al* developed a pitch tracking algorithm that is based on a definition of a local pitch period interval. The method claims to have a high degree of accuracy (a real number vice an integer) in calculating the pitch. Two successive frames are compared, and the mean square error between them is minimized. This minimization is equivalent to minimizing the cross-correlation between these segments. The technique was tested on sentences with AWGN added to levels of 10 dB and 3 dB SNR. They claimed their technique performed well in this test and a graph depicting the pitch tracks illustrated their claim.

Naylor detailed an extensive comparison between four pitch tracking algorithms: ML, modified cepstrum, harmonic matching, and auditory synchrony model (18). He concluded that the ML pitch tracking algorithm performed superior to these techniques when used in the co-channel speaker separation problem. He found that the harmonic matching algorithm was comparable to the ML algorithm, with the ML algorithm having a slight advantage in terms of error standard deviation, and the harmonic matching algorithm had an implementation advantage when used in the harmonic magnitude suppression algorithm (discussed below) (18). The ML pitch detector has been found to be able to tolerate 12 dB more noise than the cepstral method (18, 28). The ML pitch detector and the modified covariance pitch detector were reviewed by L. Lee and Morgan. They concluded that the ML pitch detector was superior to the modified covariance algorithm since the ML estimator is an unbiased estimator and the ML estimator is noise invariant. The ML pitch detector used by L. Lee and Morgan is utilized in this thesis.

2.3.2 **Segment Classification.** Some co-channel processing algorithms have employed segment classification (5, 14, 15, 26, 29). An algorithm developed by Smyth uses the auto-correlation function of a windowed segment to determine if the segment is voiced or unvoiced. Another method of classifying speech was discussed by McAulay. McAulay developed a robust detector that applied statistical decision theory to models of speech and

background noise to synthesize an optimum (minimum probability) of error classifier. Hanson and Wong used a voicing detection scheme to replace unvoiced segments with appropriately scaled AWGN. The Pitch Delay Spectral Recovery (PDSR) algorithm developed by L. Lee and Morgan claims to be robust enough so no segment classification is necessary (11).

2.3.3 **Comb Filtering.** Comb filtering techniques have historically shown only limited success in co-channel separation applications (3, 4, 11). Dick reported that the initial hope for the comb voice processor was that it would be enough by itself to make intelligible a secondary voice, which prior to processing had been masked by a primary voice. This turned out not to be the case. Dick's approach used frequency warping to vary the width of the peaks and valleys in the comb filter. The probable cause of the failure of this technique is that there are significant amounts of time, perhaps 20 to 30 percent, when the primary speaker is not producing sound, but the secondary speaker is. L. Lee and Morgan reviewed the comb filtering techniques and concluded that further research into these discrete time specially designed adaptive comb filters for co-channel processing was fruitless. The reasoning behind this conclusion was that while these filters have appropriate frequency response the phase response is not linear and their impulse response is too long. This long impulse response would work well on stationary signals, but breaks down in speech signal processing where short segments are required.

2.4 Co-Channel Speaker Separation Algorithms

Several complete co-channel processing algorithms have been developed. Historically, the first attempt to solve the co-channel speaker separation problem was the comb filtering method developed by Shields in 1970. Frazier *et al* in 1975 followed this work with an algorithm that enhanced co-channel speech by a variable comb filter that passed the desired talker's pitch and harmonics. Frazier *et al* claimed the procedure provided enhancement of speech which was degraded by another speech signal and background noise (4). They acknowledged some limitations in their approach. These limitations included problems in

separating the speech during voiced/unvoiced transitions, determining an optimum impulse response length, or lack of pitch detection.

Parsons in 1976 developed an algorithm to process co-channel speech by means of harmonic selection (20). In this approach, the input speech signal must be periodic, which restricts it to vowels and vowel-like sounds. The basic process was to identify the two talker's harmonic trains, and create a set of peak tables where the parameters of every peak were recorded and assigned to either of the two talkers. The reconstructed speech was developed by using the stored information in the peak tables which belonged to the desired talker. The results of Parson's efforts were that for *vocalic* speech the intelligibility varied from fair to excellent. The intelligibility was noted to be worst when the recovered voice is the weaker and best when both voices were strong (as expected). Additionally, the recovered speech had a remarkable naturalness and the voice is recognizable. For *non-vocalic* speech, the algorithm was faced with data it was not designed to process. The lack of uniform phonation in the input unvoiced parts of the co-channel signal did not allow for peak separation (there were no peaks) or resynthesis of the speech. The results however were remarkably good, and the unwanted voice was not completely suppressed, but was reduced to a murmur. The intelligibility of the post-processed speech was fair to good and had poor naturalness, but the target talker's voice was recognizable.

Dick developed a co-channel speech separation algorithm in 1980 that used adaptive lattice filtering techniques as a means of suppressing a primary voice (3). His algorithm used lattice filtering, comb filtering, and LPC analysis. His efforts demonstrated that lattice filtering was impractical for co-channel speaker separation, and listening tests showed that this method did not separate the co-channel speakers as hoped.

Hanson and Wong in 1983 developed the Harmonic Magnitude Suppression (HMS) technique as a solution to the co-channel speech problem (5). They also developed a well-defined formal subjective intelligibility test procedure to evaluate their results. They claimed the HMS technique could be implemented in real time with sufficient signal processing hardware/software. The HMS algorithm was applied to voiced speech segments only. No

processing was performed on unvoiced segments. They found that the HMS technique significantly improved intelligibility for SSRs of minus six dB (target speaker to interfering speaker) or better. L. Lee and Morgan reviewed the HMS algorithm, and noted that this technique recovered too much of the interfering talker. Additionally, they discussed the type of windowing function employed could impede the recovery if the pitch was too low and the harmonic distances are smaller than the frequency response of the Hamming window used.

C. K. Lee and Childers applied multi-signal minimum-cross-entropy spectral analysis (MCESA) to the co-channel speaker separation problem (10). Their process involves making an initial estimate of the spectrum of each talker. Once this estimate is made, spectral tailoring using MCESA is done to refine the initial estimates of the spectrum of each talker. They used the HMS algorithm to estimate the spectrum of the two talkers. Their results show that an intelligible estimate of the desired speech was achieved for a male and a female talker with combined at SSRs down to -18 dB.

In 1987, Naylor published a report on an "Interference Reduction Model" (IRM) (18). His effort focused on developing techniques for suppressing the interfering (louder) talker for use in automatic speech or speaker recognition systems. Naylor's baseline system was based on the HMS algorithm, and processed only segments of speech that were voiced. No processing was performed on segments that were unvoiced. Naylor discussed a flexible harmonic placement modification to the HMS algorithm. This modification proved to be quite useful and allowed for more general placement of the individual harmonic pulses. Naylor developed an alternative processing algorithm for unvoiced segments. The unvoiced frames were low-pass filtered at f_{cutoff}=1800 Hz, and resulted in minimal voiced target speaker suppression, and considerable unvoiced interference suppression. Another significant accomplishment by Naylor included the development of a listening test station.

C. Rogers *et al* developed a co-channel speaker separation process called the automated two speaker separation system using a neural network (1, 16). Their process was based on accurate pitch detection and frame size determination, a speech detection scheme, selection and assigning of spectral components to the appropriate talker, and low-pass filtering. This

process was unique in that the speech detection process determined if one, or more than one talker was present in the segment of interest. Here a neural network frame classifier was used to extract information to determine the number of speakers and their voicing states in each segment. They implemented two feed-forward multi-layer perceptrons that were trained by back-propagating errors based on Fourier coefficients. They claimed this neural network made choices concerning the spectral components for each voice was much closer to those decisions made by experienced human operators in manual control of the system. Their results showed that a neural network can make accurate judgements as to the nature (number of talkers, segment is voiced or unvoiced) of a co-channel speech waveform. This neural network based information could be used to enhance co-channel separation algorithms over information derived from rule-based systems.

Stubbs and Summerfield developed two algorithms for co-channel speaker separation in 1988 (27). The first algorithm was a development of Parsons' harmonic selection algorithm, and the second algorithm operates on the cepstrum of speech. They clearly state that their algorithms have limited applicability and are not to be construed as complete co-channel separation algorithm, but they might lend themselves as enhancing speaker separation strategies. The cepstrum strategy attempts to exploit the fact that the interfering talker is a harmonic signal (they limited their algorithms to voiced sections), and the interfering talker can be filtered out in the cepstral domain. The harmonic selection technique used the harmonic peaks in the frequency domain that constitute a voice and uses this information to reconstruct the voice. This technique achieves separation of the talkers by exploiting knowledge of the spectro-temporal characteristics of the target voice.

Zissman *et al* devised a co-channel speaker separation algorithm which utilized the voicing states of the target and interfering talker, and applied suppression to selective regions of the interfering talker depending on the voicing state of the segment (29). Their results identify the regions in the co-channel speech signal where interference suppression would prove beneficial. They claimed that in speech segments where the voicing state of the target talker was anything, and the interfering talker's voicing state was voiced, the post processed

intelligibility of these segments improved dramatically. Their algorithm may not improve the intelligibility when the target talker was voiced and the interfering talker had any voicing state.

An alternate co-channel speaker separation algorithm based on a sinusoidal model for speech was developed by Quatieri and Danisewicz (23). Their approach used these sinusoidal models for the suppression of the interfering talker. In this technique, as in Naylor's HMS technique, only voiced segments of speech are considered. The primary distinction between this technique and previous techniques is that a least-squares estimation of the sine-wave parameters of both the target and interfering speaker are calculated. Once these sine-wave parameters are estimated, a sinusoidal based speech analysis/synthesis system is used to reconstruct the speech of the desired talker. This technique examines each segment of data and estimates of the sine-wave amplitudes by performing short-time Fourier transform magnitude analysis and peak-picking to determine an estimate of the fundamental frequency. *A priori* pitch information provided good separation up to -16 dB SSR. With no *a priori* data, good separation was achieved at approximately equal signal levels (0 dB).

Quatieri and Danisewicz's approach operates on the basis that a sinusoidal model for speech can be obtained for each of the two speech signals in the co-channel signal. A unique feature in this process is the concept of a time-evolution of the sine wave parameters between segments during the processing. This time-evolution occurs across frames where the frequencies of both talkers are closely spaced. The estimates of the sine wave assume the frequencies are harmonically related

Two approaches were introduced by Quatieri and Danisewicz for talker separation: peak picking and frequency sampling. The peak picking methodology selects the largest peaks from the summed spectra and this data is used to reconstruct the larger of the two waveforms. The waveform estimate is subtracted from the combined waveform to form an estimate of the lower signal. (This approach is similar to L. Lee and Morgan). A trade-off between the length of the analysis window and the resolution of the harmonic peaks was examined. The longer the analysis window, the narrower the peaks will be and greater separation can be achieved between the two talkers having close frequencies. The frequency sampling approach assumed

a priori knowledge of the sine wave parameters of both talkers. The method used to separate the two talkers begins by parallel processing the spectrum. One path recovers the energy of one talker from sampling the spectrum at the known frequencies. The other path samples the spectrum at the frequencies between the known frequencies, thus recovering the energy of the other talker. This sampling method is similar to comb filtering techniques. Their analysis window varied from 20-50 msec. A large portion of their work used *a priori* knowledge of the least-squares estimates of the parameters in the sinusoidal models of each talker. Overall, this approach showed an interesting application to co-channel suppression. However, their approach is limited to only a small subset of vocalic co-channel signals. It was suggested that if adequate pitch detection/estimation could be accomplished, their suppression technique is viable (23).

L. Lee and Morgan developed a co-channel speaker separation algorithm termed the Pitch Delay Spectral Recovery (PDSR) (11). The PDSR has the goal to separate and reconstruct both the target and interfering talker. Their co-channel speech signal was comprised from a male and female talker from the TIMIT data base. They combined these speech signals at SSRs of 0, 3, 6, 10, and 20 dB. The ML pitch detection algorithm (28) was also used to estimate the pitch of the windowed segment of data. This pitch information was input into a speaker recovery algorithm. The unique feature of this speaker recovery algorithm is that it used a lag spectral estimate to differentiate between the target and interfering talker. Finally, a spectral assignment algorithm (using an LPC filter) reassigns/reconstructs the speech segment for both the target and interfering talker. Their algorithm operated effectively in SSR between -18 to 18 dB and produced the recovered talkers which are intelligible to human listeners. The speaker recovery algorithm was the crux of their efforts, and they reported good results in their experiments. However, as noted in their report, their spectral assignment algorithm was questionable and perhaps a better assignment algorithm could be devised. This uncertainty in the assignment of the separated segments of co-channel speech is the motivation behind this thesis.

This chapter discussed the issues concerning the co-channel speaker separation algorithm and provided a brief review of the historical efforts that have been developed to help solve the co-channel speaker separation algorithm. The next chapter introduces the spectral assignment processing algorithm developed in this thesis.

III. Spectral Assignment Algorithm Development

In a typical co-channel speech signal, the relative energy of each talker will vary depending upon the utterance being spoken, and the respective energy levels for talker *X* may be higher than talker *Y* at one instance, and this condition can and often reverses itself in the co-channel signal. The basic idea in co-channel speaker separation is that if the signal energy of the two talkers in the co-channel signal can be identified and isolated, both talkers' speech tracks may be recovered (to a degree). Even if the energies of both talkers could be separated, an uncertainty still exists when the time comes to properly assign the separated energy to the appropriate talker. The hypothesis for this thesis is an attempt to solve this uncertainty in the assignment of the separated energy to the appropriate talker by using an LPC based distortion metric.

This chapter will discuss the methodology used to separate the energy of the two talkers and the subsequent assignment of the separated frames to the correct talker. Care should be taken not to confuse a co-channel speech segment with a separated co-channel speech segment. The co-channel speech segment contains all the energy, and has two talkers plus noise. A *separated* co-channel speech segment (hopefully) contains the energy of *one* talker plus some noise. Thus a single co-channel speech segment is processed and split resulting in two speech segments: one containing the energy of talker *X*, and the other containing the energy of talker *Y*.

Two spectral assignment methodologies are discussed: Pitch Delay Spectral Recovery (PDSR) (*a priori* pitch error based) and an LPC based distortion metric. Both of these algorithms were implemented using Matlab.

3.1 Spectral Separation Algorithm.

The spectral separation algorithm used in this effort to separate the co-channel speech signal was developed by L. Lee and Morgan (11). The PDSR algorithm has three parts: the maximum likelihood (ML) pitch detector, a speaker recovery (spectral separation) algorithm,

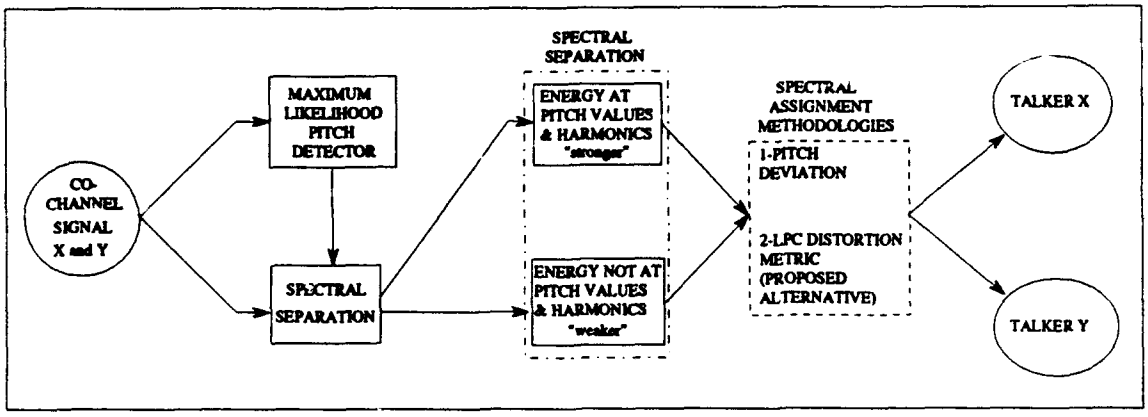


Figure 3.1. PDSR Algorithm

and a frame assignment algorithm. Figure 3.1 shows the basic signal flow used to separate the two talkers in the PDSR.

In the PDSR algorithm, the process to recover the energies of both talkers begins with the ML pitch detector. The ML pitch detector calculates the pitch in a segment of co-channel speech. This pitch value is derived from the stronger talker's energy but it can be associated with either talker. This pitch value is vital to the spectral separation process.

The next step in the PDSR is to use the calculated pitch value to separate the energy of the co-channel segment into two parts. These two parts are termed the "stronger" and "weaker" segments. The PDSR algorithm works on cyclo-stationary signal analysis and the concept of a lag spectrum. The value of the pitch computed in the i^{th} frame of the co-channel speech signal is used to determine the amount of lag to create the lag spectrum. The energy in that "lag" frame will be aligned at the pitch values (and harmonics) of the stronger talker, and the energy will be added together constructively, while the underlying frequencies not at this pitch value and harmonics will be added together destructively.

Thus given the amount of lag to add to the i^{th} frame of co-channel signal based on the i^{th} calculated pitch value, the signal was delayed by this lag factor and *added* to itself to recover the stronger talker, and the signal was delayed and *subtracted* from itself to recover the weaker

talker. The segment containing the energy from the stronger talker is made by sampling the lag spectrum at the pitch value and its harmonics. All other spectral components are "thrown away". The segment for the weaker talker is determined in a similar fashion, *except* the spectral components at the pitch and harmonics are "thrown away" and the remaining spectral samples are kept.

This spectral separation does not completely separate the energies of the two talkers, but the majority of the co-channel energy is separated. For example, if the pitch in frame i of talker X was 75 Hz the harmonics would be at 150, 225, 300 Hz, etc. If talker Y's pitch value was 100 Hz, the harmonics would be 200, 300, 400 Hz, etc. From this example the spectral components at 300 Hz overlap and these harmonics would not be separated, and would subsequently be assigned to only one talker. Thus, at the output of the speaker recovery algorithm, are two segments of speech, one containing the stronger talker's energy and the other containing the energy of the weaker talker (11).

This procedure also has varying degrees of success depending on the voicing states of talkers X and Y. That is, the voicing states of talker X and talker Y can be any combination of voiced, unvoiced, or silence, and these co-channel speech segments will vary throughout. When both talkers in the co-channel speech signal are unvoiced, separating the two unvoiced speech signals by the PDSR algorithm is virtually impossible. This is because there are no spectral properties in an unvoiced segment of speech. The two "summed" and "weaker" separated speech segments will both have spectra that is unvoiced. The only uncertainty involved in this case would be the relative energy levels of the "stronger" and "weaker" segments. No adjustment is made in the PDSR to vary the energy level of separated unvoiced segments of co-channel speech data. When the target talker enters a period of silence, the unwanted talker's spectral energy can simply be suppressed; however, detecting the silent portions for only one talker is virtually impossible. When both talkers are silent (trivial case), no processing is required.

This section discussed how the PDSR algorithm separates a segment of co-channel speech signal into the "stronger" and "weaker" segments. The next section will introduce the

two spectral assignment algorithms that will be used in this thesis to assign these separated segments of co-channel speech to the appropriate talker.

3.2 Spectral Assignment Based on Pitch Values.

Once the energy in a segment of co-channel speech has been separated, the proper assignment of these separated segments is the next challenge. Two spectral assignment algorithms were tested in this thesis. Both of these algorithms provide a decision making device which is based on calculated speech signal parameters. The algorithms discussed are: *a priori* minimum pitch distance, and an LPC based distortion metric (both *a priori* and non-*a priori* training data). These algorithms follow the PDSR spectral separation process outlined in Figure 3.1.

The spectral assignment logic used is the PDSR algorithm which uses the individual "clean" speech signals of both talkers in the co-channel speaker signal to compute their two individual pitch tracks. The spectral assignment decision logic is based on the minimum percent deviation of the computed co-channel pitch value to the *a priori* pitch values from talker *X* and talker *Y* respectively. In the algorithm during the i^{th} frame processed, the co-channel pitch is calculated and compared to the *a priori* individual pitch values, and a decision is made on assignment of the i^{th} separated frame of co-channel speech. The assignment decision is made on each i^{th} segment of co-channel data (not deferred), and is based on the minimum value found by the following equation:

$$\text{Decision:} \quad \min\left\{\left|\frac{\text{pitch } X_i\text{-pitch cochan}_i}{\text{pitch } X_i}\right|, \left|\frac{\text{pitch } Y_i\text{-pitch cochan}_i}{\text{pitch } Y_i}\right|\right\} \tag{3.1}$$

If the original speakers are talker *X* and talker *Y*, the individual pitch tracks would be calculated and assigned as p1 (pX) and p2 (pY) respectively. Suppose the co-channel signal was created and talker *X* had a +10 dB advantage over talker *Y*. The co-channel pitch track is calculated and labeled pC. The input data into the PDSR is the co-channel speech signal,

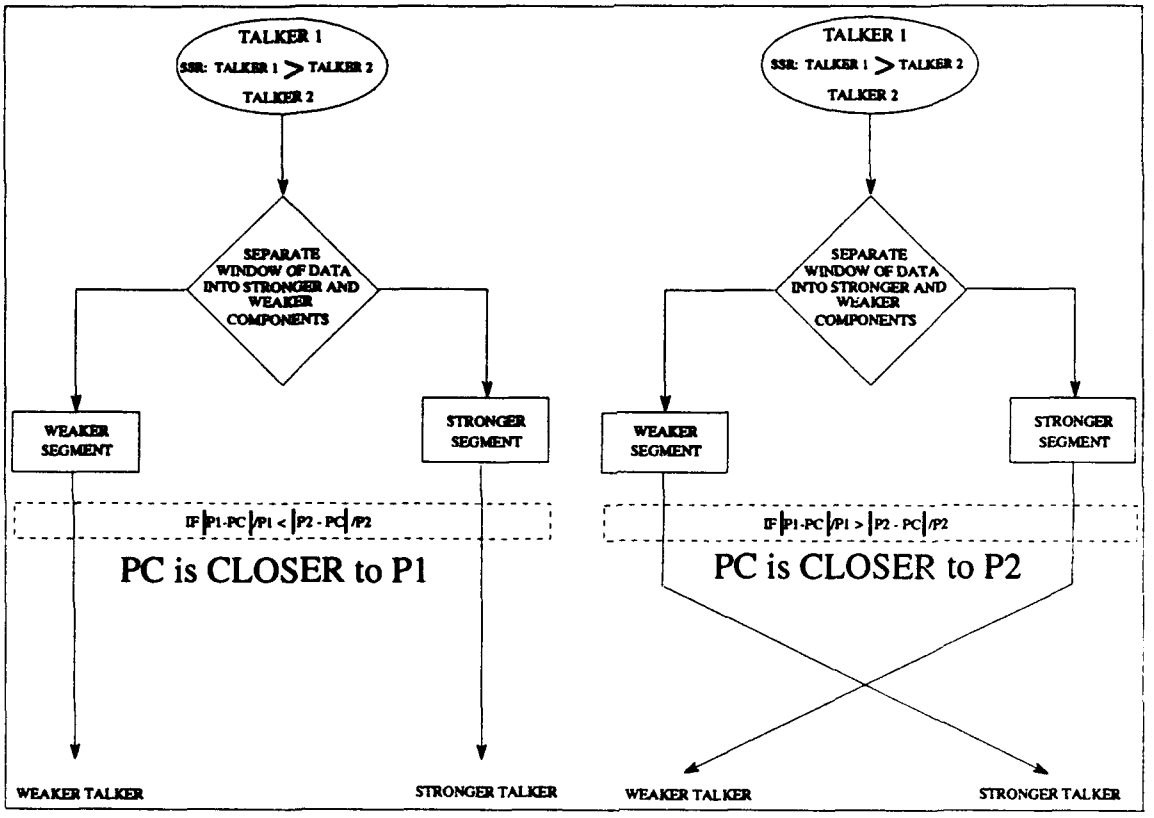


Figure 3.2. PDSR Pitch Deviation Frame Assignment Logic

and the previously calculated pitch tracks; p1 (pX), p2 (pY), and pC. During processing in the PDSR algorithm a window of co-channel speech is divided into two segments labeled "out2", and "out1". The "out2" segment contains the "summed/stronger" speech created by sampling the pitch and harmonic values. At this point, a decision is required on which talker is to be assigned the out2 and out1 segment. The "minimum pitch"decision is based on the minimum deviation from pC_i to $p1_i$ and $p2_i$. If the i^{th} frame's minimum deviation was closer to p1, the "out2" segment would be assigned to t1, and by default the "out1" segment would be assigned to t2. In the other case, if the i^{th} frame's minimum deviation was closer to p2, the "out2" segment would be assigned to t2, and the "out1" segment would be assigned to t1. These two assignment cases are illustrated in Figure 3.2. Shown in Figure 3.2 are the two decision choices for the co-channel case when talker 1 has a SSR advantage over talker 2.

Thus for each frame of co-channel signal processed, the minimum deviation of the i^{th} frame co-channel pitch value to the *a priori* individual i^{th} pitch values determines the assignment of the "'stronger" and "weaker" segments. As noted, this methodology requires the "clean" talker's *a priori* pitch tracks to properly assign the separated frames of co-channel speech signal. A methodology that is not based on *a priori* information would be more practical and realistic. The following section introduces an assignment methodology that is not based on any *a priori* information that is contained in the co-channel speech signal.

3.3 Spectral Assignment from an LPC Based Distortion Metric.

An alternate approach to the pitch deviation assignment algorithm is an assignment methodology based on an LPC distortion metric. LPC techniques have been used in speaker identification processes, speech data reduction, and other speech processing applications. In this thesis, LPC techniques are being employed in a similar fashion as speaker identification. That is, given the two separated segments of speech can the appropriate speaker be "identified" from this information and hence have an appropriate assignment of the two separated segments? This question leads to the quantification of the LPC based distortion metric discussed in the following sections.

As mentioned, LPCs are frequently used as a distinctive feature in speaker recognition. In this thesis they are employed to determine the similarity between a test speech signal and a model speech signal. Several LPC based "distance" measures have appeared in the literature. Parsons (7, 21) provide a general overview of the Itakura-Saito measure and the Itakura minimum-prediction residual.

The Itakura minimum-prediction residual is used in this thesis. In computing the minimum-prediction residual, the autocorrelation function, r, of a segment of the test signal is first calculated. Additionally, the LPC coefficients of this segment are computed. In this thesis the LPC coefficients are found by the Levinson-Durbon recursion algorithm (21). From the autocorrelation function, a Toeplitz matrix, R, is made. A Toeplitz matrix has the property where all the elements in each diagonal are equal. The Toeplitz matrix is used in the residual

calculation and will be referred to as the autocorrelation matrix. For a segment of speech the minimum-prediction residual error, D, is found by computing the autocorrelation matrix R, and the LPC vector $\vec{a}$, as shown in equation 3.2.

$$D = \vec{a}^T R \vec{a} \tag{3.2}$$

The prediction error in equation 3.2 has been used as a measure of the difference (or similarity) between two signals: a test signal and a model signal. The residual can be thought of in a similar fashion to a "Euclidean distance," where the smaller the distortion the "closer" the test signal is to the model signal. The error, D, is calculated between a test LPC vector, $\vec{a}$, and the test signal's autocorrelation matrix, R, and the model LPC vector, $\vec{b}$. Itakura (7) showed that for a test LPC vector $\vec{a}$, the test signal's autocorrelation matrix, R, and a model LPC vector, $\vec{b}$, the following relation holds:

$$D = \frac{\vec{b}^T R \vec{b}}{\vec{a}^T R \vec{a}} \geq 1 \tag{3.3}$$

The minimum-prediction residual shown in equation 3.3 is used in this thesis as the basis for the spectral assignment algorithm developed. This section discussed the minimum-prediction residual that will be used to assign the separated segments of co-channel signal to the appropriate talker. The next section will discuss the particular processing used to calculate D.

3.4 Calculating the Minimum-Prediction Residual

The minimum-prediction residual, D, in equation 3.3 provides a scalar quantity that measures the "similarity" between a test signal and a model signal. The precomputed model

parameters are the LPC vectors $[1 - a(1) - a(2) - \cdots - a(p)]$ where p is the order of the predictor. The model LPC vector is derived from model speech signals. In this thesis, the model speech signals are derived from two sources. The first source of model speech is the same "clean" speech that was used to make the co-channel speech signal. This *a priori* model speech will be used as a "proof-of-concept" for the spectral assignment from the LPC based distortion metric. If this "proof-of-concept" proves fruitful, the second source of model speech used to precompute the model LPC vectors are derived from the same talkers in the co-channel signal, *except* the utterances which are spoken are **independent** from the co-channel speech signals. Since the second source of model data is independent of the co-channel speech data, the resulting co-channel separation processing does not require any *a priori* information in the co-channel speech.

The model LPC vectors are computed using the same processing (i.e. normalization, prefiltering, windowing, and window-stepping, etc.) that was used to process the co-channel speech signal. During training (i.e. computing the model LPC vectors), for talker *X* there may be N precomputed LPC vectors and M LPC vectors for talker *Y*. Once these LPC vectors are precomputed, they are stored and used in the co-channel separation algorithm.

In the PDSR algorithm, the co-channel signal is separated into the "stronger" and "weaker" segments. The "stronger" segment is used as the test signal. From this "stronger" segment the autocorrelation matrix R (p X p) and the LPC vector $\vec{a}(0, 1, \cdots, p)$ are computed. Given the R and $\vec{a}$ data, the residual, equation 3.3, is computed for each N and M stored LPC model vectors. The minimum value of these $N + M$ distortion computations is used to find the decision. Once the minimum-prediction residuals for each N and M model data sets are computed, and minimum of these residuals are found, and the "stronger" and "weaker" segments are assigned to talker *X* or *Y* accordingly. This assignment process is illustrated in Figure 3.3.

Classification of the voicing states was made in the "clean" speech during the precalculation of the LPC vectors. A voiced/unvoiced speech detector was implemented based on the work by Rabiner and Schafer (25) and Rabiner and Sambur (24). This voiced/unvoiced

speech detector works on the energy in a speech segment and a threshold value. Rabiner *et al.* used a definition of "energy" as being the sum of the magnitudes of the samples in a signal. The magnitudes of the samples were used instead of the square of the samples since a small number of samples with a high amplitude will not be emphasized by the summing operation. The threshold value, T, is computed by taking a percentage of the maximum value of this energy function minus the noise energy. The procedure to obtain the threshold follows.

Each speech signal in the data base has at least 1000 samples before the actual spoken speech signal begins. These first 1000 samples can be used to calculate the noise energy. The noise floor is calculated using equation 3.4:

$$\sigma = \sum_{k=0}^{P-1} | s[k] | \tag{3.4}$$

where $P = 801$ is the number of samples in a frame

Thus σ is the noise floor or "silence" energy in the speech signal. The energy per frame in the speech signal is then computed. For this thesis, a sampling rate of 16 kHz was used, and the window size was 50 msec, or 800 samples. These frames were stepped by 10 msec overlapping intervals of 160 samples long. The "energy" value in segment i of the speech signal is shown in equation 3.5:

$$E_i[n] = \sum_{k=0}^{P-1} | s_i[k] | \tag{3.5}$$

where $P = 801$ is the number of samples in the i^{th} frame

Equation 3.5 is computed for each i^{th} frame in the model speech signal. The threshold value is found by equation 3.6:

$$T = z(\max(E) - \sigma) - \sigma \tag{3.6}$$

In equation 3.6, the parameter z drives the threshold value T. In this thesis z was selected to be 10% of the maximum from the energy function in equation 3.5. Thus, during the processing of the training speech signal if the energy E_i in the i^{th} frame is less than the threshold T, $E_i < T$, this frame of model data is disregarded and the next frame is processed.

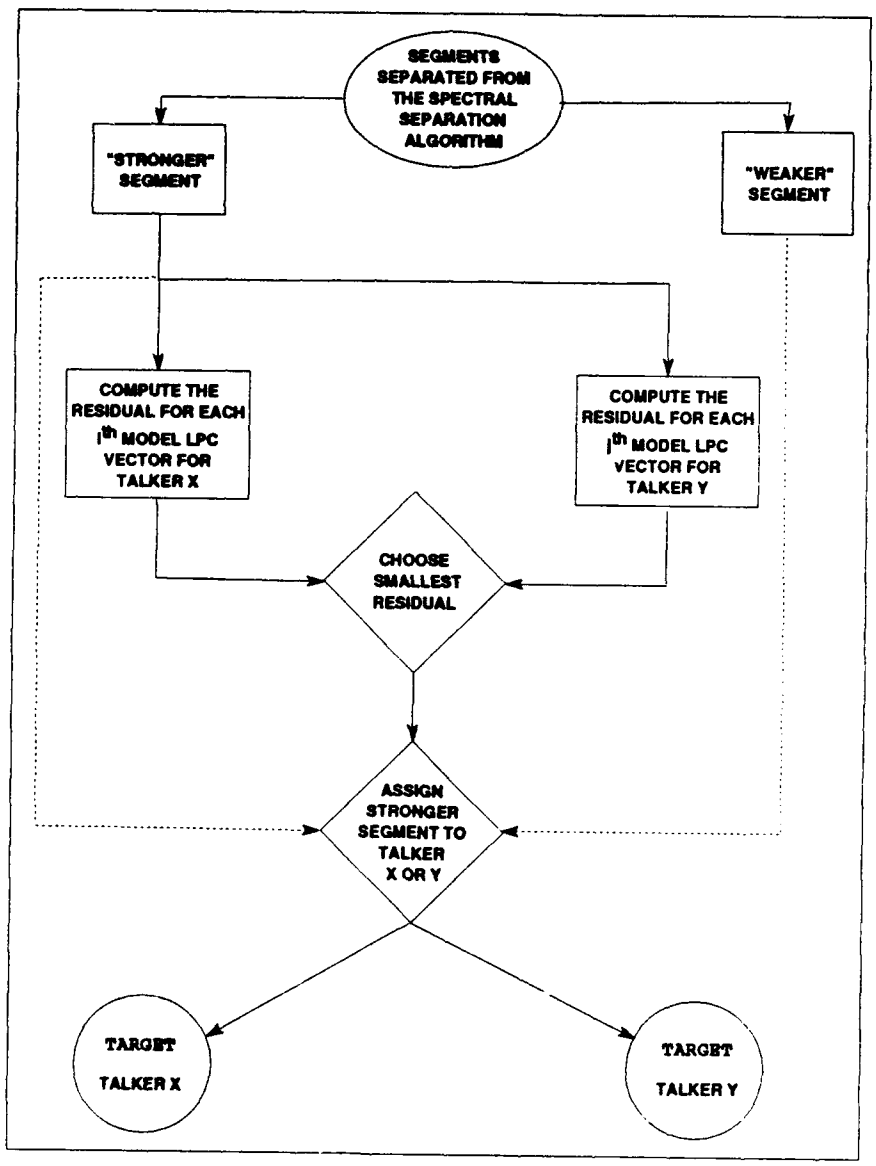


Figure 3.3. Block Diagram of the LPC Based Distortion Metric Spectral Assignment Algorithm

3.5 Summary

This chapter discussed the process by which a frame of co-channel signal is separated into "stronger" and "weaker" energy segments and subsequently assigned to the appropriate talker. In this thesis, the process by which the co-channel speech signal is separated is given, and the method by which the separated segments are assigned to the "correct" target talker is the focus of the research. Two spectral assignment algorithms were discussed, the PDSR (pitch distance) and the LPC based minimum-prediction residual error. The LPC based methodology is proposed as an alternate solution of the pitch distance algorithm in the problem of appropriately assigning the separated co-channel speech signal. The next chapter will discuss the testing performed on the spectral assignment algorithms presented in this chapter and discuss the results of the experiments.

IV. Co-Channel Test Signals/Experimentation Results

This chapter discusses the spectral assignment algorithms that were tested against differing types of co-channel speech signals. The PDSR algorithm discussed in Chapter III was used to separate the co-channel speech signals into "stronger" and "weaker" speech segments. The primary focus of this chapter is to evaluate the success of the spectral assignment algorithms in their assignment of these "stronger" and "weaker" segments to the appropriate talker. The chapter begins with a discussion of the test speech signals used to make co-channel speech signals and the computation of the model LPC vectors. Additionally, a description of the test procedures employed is provided, and the chapter concludes with a discussion of the test results.

4.1 Co-Channel Test Speech Signal Databases.

The speech data used in this thesis was derived from two sources: the pre-recorded/CD-ROM TIMIT speech database, and speech signals that were recorded on the Sun Sparc-2 workstation. A description of the TIMIT speech files and the recording hardware/software are provided in Appendix A. Also provided in Appendix A is a description of the different digital data formats of the respective speech signals. These two sources of speech signals are convenient to use and analyze because they both are sampled at 16 kHz and are each 16 bit linearly quantized. The unprocessed "clean" single talker speech data files are used to make the test co-channel speech signals and are also used to precompute the sets of model LPC vectors.

Each speech signal, regardless of the source, was normalized before any processing. That is, the DC component (mean) was subtracted off, and the resulting signal was normalized by the standard deviation. The following relation was used to normalize the speech data files:

Individual Speech File	SNR (dB)
TIMIT/Female	44.9
TIMIT/Male	25.2
Recorded/Talker 1	18.3
Recorded/Talker 2	17.9

Table 4.1. SNR of Individual Speech Signals

$$\text{Normalized Signal} = \frac{[\text{Raw Signal} - \text{DC (mean)}]}{\text{Standard Deviation}} \tag{4.1}$$

Once the individual input speech signals were normalized, the test co-channel speech signals were generated at three signal-to-signal ratios (SSR): +5, 0, and -5 dB. The appropriate gain factor for the SSR was multiplied to the input speech signals and the co-channel speech signals were made by simple point-by-point addition.

The SNR of the respective individual speech signals was computed using equations 2.1 and 2.2. The speech signals were examined and the power was calculated in the segments that contained noise only, and in the segments that had signal plus noise. The noise power was subtracted from the signal plus noise sections leaving the signal power only. The SNR was easily calculated from the remaining noise only and signal only data. Table 4.1 lists the SNR values calculated for the individual speech files from the TIMIT database and the recorded speech files.

The first set of co-channel test speech signals was derived from the TIMIT database. From the TIMIT speech database two talkers (one sentence each) were arbitrarily chosen. One male and one female talker was chosen to favorably bias the testing, and the individual sentences were chosen that had similar (short) lengths. The female talker (fdhc0), spoke the sentence: *You saw them always together those years, (si2189).* The male talker (mrgt0), spoke the sentence: *He would not carry a brief case, (si2080).* The talker's initials and sentence

code numbers shown are provided as a reference to the TIMIT database. These sentences have no special phonemes or word choice to distinguish them. The shorter sentence was padded with zeros to make the individual speech signals the same length. The only bias in choosing these sentences were (a hoped) general male/female talker distinction (pitch values), and short lengths to speed up the processing.

The other set of test speech signals were recorded in the computer lab using the Ariel/ProPort A/D converter on the Sun Sparc-2 workstation. No special noise reducing techniques were used, and this is evident in the differing SNR values listed in Table 4.1. One advantage to recording the speech signals in the lab was the ability to manually set the record length, so that both recorded speech signals had the same number of samples.

The recorded speakers are designated talker 1, and talker 2. The following sentences were spoken: *Why were you away a year Roy?* (talker 1), and *While you were away in Walla Walla,* (talker 2). These sentences are similar to the co-channel test sentences used by Quatieri (23). These sentences have the distinction of having relatively continuous pitch tracks. The two talkers were both males and are differentiated by their pitch tracks; talker 2's pitch is lower than talker 1's.

Given these two sets of individual "clean" speech signals, the suite of co-channel test signals were created. In order to maintain continuity, each individual speech signal was normalized prior to making the co-channel speech signals. Six co-channel speech signals (three each from the TIMIT data base and recorded signals) were created with SSRs of +5, 0, and -5 dB; and they constitute the co-channel test signals analyzed in this thesis.

4.2 Speech Data used to Precompute the Model LPC Vectors.

The spectral assignment algorithm proposed in this thesis and discussed in Section 3.3 requires precomputation of model LPC vectors from model speech data. For each talker in the TIMIT database there are eight sentences that are suggested for use in speech testing. As such, with one sentence selected for use in the co-channel signal, that left seven sentences spoken by the same talker for other speech applications. These extra seven sentences for the

male and female talkers are used to precompute the model LPC vectors and eventually the minimum-prediction residual, reference equation 3.3 in Section 3.3.

Additionally, these extra seven TIMIT sentences were arbitrarily chosen to be spoken and recorded in the lab by talker 1 (sentences spoken by fdhc0) and talker 2 (sentences spoken by mrgt0). The 14 sentences used to precompute the model LPC vectors for the TIMIT speakers and the recorded talkers are:

Female (fdhc0) and talker 1:

The misquote was retracted with an apology. (sx119)

Michael colored the bedroom wall with crayons. (sx209)

This brochure is particularly informative for a prospective buyer. (sx290)

Shaving cream is a popular item on Halloween. (sx299)

They own a big house in the remote countryside. (sx389)

But such cases were, in the past, unusual. (si929)

Visually, these approximated what he was feeling within himself. (si1559)

Male (mrgt0) and talker 2:

Are your grades higher or lower than Nancy's? (sx10)

Project development was proceeding too slowly. (sx100)

Serve the coleslaw after I add the oil. (sx190)

The oasis was a mirage. (sx280)

That noise problem grows more annoying each day. (sx370)

By that, one feels that magnetic forces are as general as electrical forces. (si820)

Meats: the radiation processing of meat has received extensive investigation. (si1450)

The first attempt to compute the model LPC vectors used the entire model speech signals and processed them through the same normalization, pre-filtering, windowing, and frame steping that the co-channel signal would experience. Preliminary results from this methodology did not prove successful.

The problem in this methodology involved the co-channel spectral separation process. In the co-channel spectral separation process, there exists a large difference in values of the computed model LPC vectors between the original input signal and the separated "stronger" segment. The difference in the LPC vector values for each frame of an input signal is shown in Figure 4.1. The difference in LPC vectors was found by equation 4.2.

$$\text{Diff. in LPC Vectors} = \text{LPC Vectors Calc. by Co-Chan Pre-Processing} - \text{LPC Vectors Calc. on "Stronger" Seg} \quad (4.2)$$

The significant differences in the LPC vectors shown in Figure 4.1 caused severe errors in the computation of the minimum-prediction residual, and subsequent erroneous assignment of the "stronger" and "weaker" segments of speech occurred. As a result of the above errors in appropriately assigning the "stronger" and "weaker" segments, the model LPC vectors were computed from processing the "clean" speech signals through the co-channel spectral separation algorithm and using the "stronger" segment to compute the LPC vectors. The precomputed model LPC vectors were obtained by concatenating the seven sentences together and processing this signal through the training software. The following procedure was finally developed and used to precompute the model LPC vectors for each talker in the co-channel test signals:

1. Concatenate the "clean" model speech signals into one signal.
2. Compute the pitch of each of the "clean" model signals with the maximum-likelihood (ML) pitch detector.

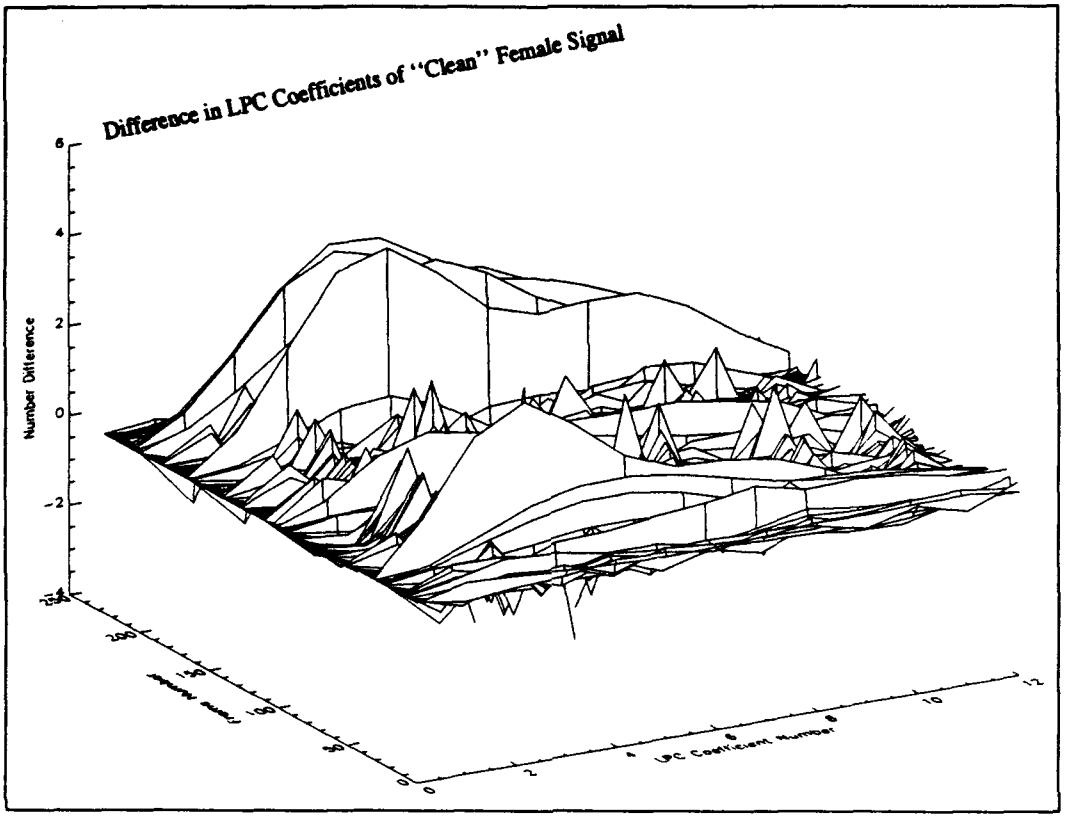


Figure 4.1. Difference in LPC Vectors

3. Normalize the signal by subtracting off the mean and dividing by the standard deviation.
4. Pre-filter the model signal with the following transfer function: $H(z) = 5/(1 - 0.95z)$
5. Segment the model signal using a 6th order Kaiser window, 50 msec long. Step the frames forward by 10 msec increments.
6. Calculate the energy in each frame. Calculate the noise floor. Calculate the threshold at 10% of the maximum energy value plus noise floor.
7. Process the model signals through the co-channel separation algorithm.
8. Compute the model LPC vectors from the "stronger" separated segment from the co-channel processing.

Data File	Contents
FemaleLPC1	Model LPC Vectors from Female's Speech in Co-Channel Signal
MaleLPC1	Model LPC Vectors from Male's Speech in Co-Channel Signal
Talker1LPC1	Model LPC Vectors from Talker1's Speech in Co-Channel Signal
Talker2LPC1	Model LPC Vectors from Talker2's Speech in Co-Channel Signal
FemaleLPC7	Model LPC Vectors from 7 Female Excluded Sentences
MaleLPC7	Model LPC Vectors from 7 Male Excluded Sentences
Talker1LPC7	Model LPC Vectors from Recorded 7 Female Excluded Sentences
Talker2LPC7	Model LPC Vectors from Recorded 7 Male Excluded Sentences

Table 4.2. Model LPC Vector Data Files Generated

9. Discard any LPC vector that was calculated in frames where the energy in the frame fell below the threshold value.
10. Store the remaining LPC vectors for each talker.

At the conclusion of precalculating the model LPC vectors, the different sets of model LPC vectors listed in Table 4.2 were generated.

4.3 **Test Procedure.**

The test procedure employed in this thesis was to process the six co-channel speech signals through the one spectral separation algorithm, and then test the success of each spectral assignment methodologies.

The following test procedure was used in this thesis:

1. Precompute the *a priori* pitch tracks for all test signals, both model and co-channel signals.
2. Create the six co-channel test signals.
3. Calculate the model LPC vectors listed in Table 4.2 for each model signal in the co-channel signals.

4. Process the co-channel test signals through the co-channel separation algorithm that uses the "pitch deviation" method for the assignment of the "stronger" and "weaker" segments. The results of the "pitch deviation" methodology serve as a baseline for comparison with the other spectral assignment techniques.

5. Process the co-channel test signals through the co-channel separation algorithm that uses the minimum-prediction residual (using the **same** speech in the co-channel signal to precompute the LPC vectors) for the assignment of the separated co-channel speech segments.

6. Process the co-channel test signals through the co-channel separation algorithm that uses the minimum-prediction residual (using the seven **independent** sentences to precompute the LPC vectors) for the assignment of the separated co-channel speech segments.

4.4 Test Results.

The following subsections describe the results of the different spectral assignment techniques tested in this thesis. An informal personal listening test will judge the intelligibility of each post-processed co-channel signal. Graphical results pertaining to the frame assignments will be developed that will show the effectiveness of the separation methodologies. Spectrogram plots of the "clean" and post-processed separated speech signals are also provided in Appendix B. These spectrogram plots show graphically how well the energies in the separated co-channel signals were recovered from the input co-channel speech signals. As an example of the spectrograms, Figures 4.2, 4.3, 4.4, and 4.5 are the spectrogram plots of the four "clean" individual speaker files used in this thesis.

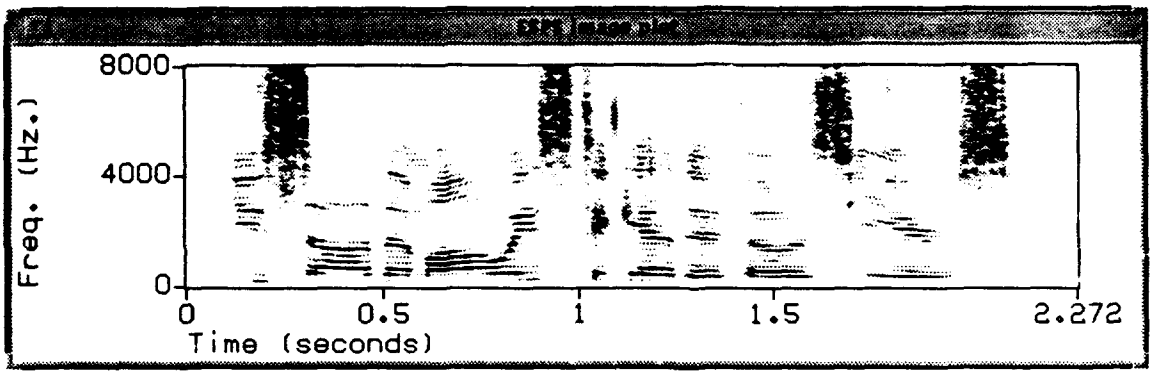


Figure 4.2. Spectrogram of "Clean" Female Speech Signal

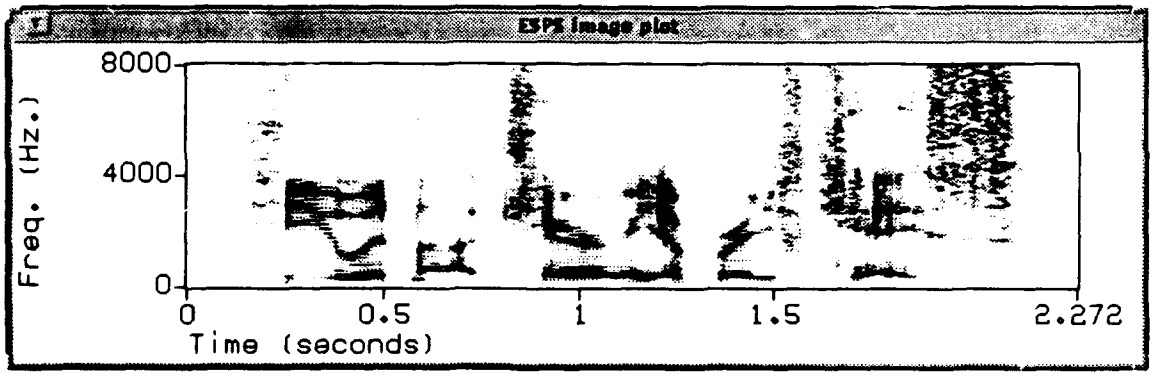


Figure 4.3. Spectrogram of "Clean" Male Speech Signal

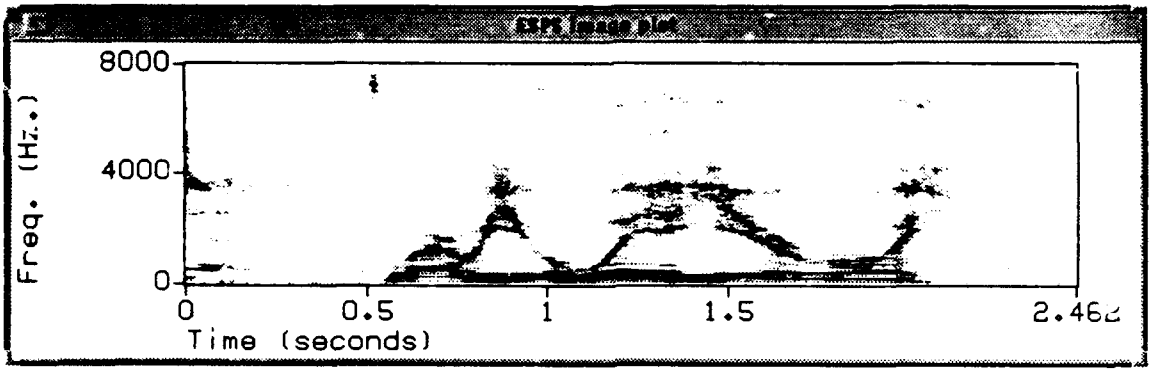


Figure 4.4. Spectrogram of "Clean" Talker 1 Speech Signal

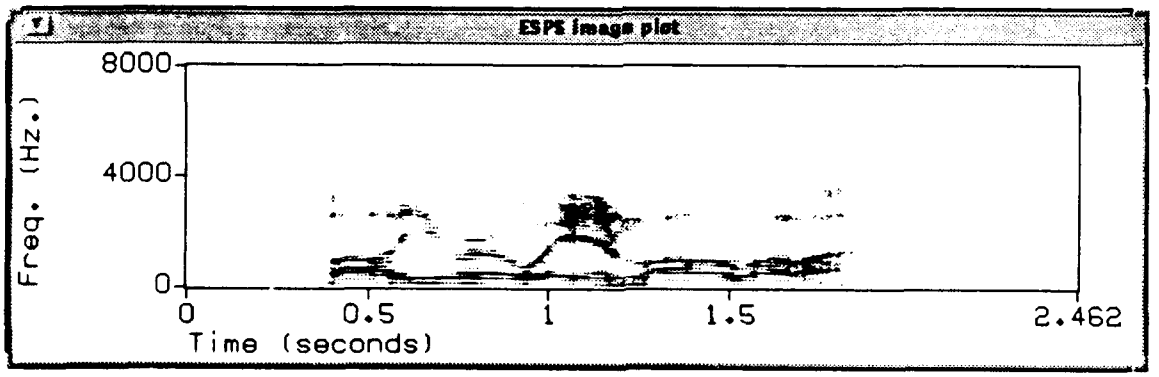


Figure 4.5. Spectrogram of "Clean" Talker 2 Speech Signal

4.4.1 **Pitch Deviation Method of Assigning Separated Segments.** As a baseline comparison, the first spectral assignment algorithm examined is the "pitch deviation" algorithm that was developed and supplied by L. Lee and Morgan (11). Their algorithm was implemented using Matlab on the Sun Sparc-2 workstation. The "pitch deviation" methodology requires the two *a priori* "clean" speech signals (specifically their pitch tracks) that were used to create the co-channel speech signal.

The first step in the "pitch deviation" methodology is the calculation of the two individual pitch tracks using the ML pitch detector of the "clean" speech signals. Figure 4.6 shows a plot of these individual pitch tracks and the time domain signals from the sentences recorded by talker 1 and talker 2.

You can see from Figure 4.6 that the pitch tracks are continuous during the time the signal is voiced. The erratic pitch values at the beginning and end of the speech signals are obtained from the noise portion of the time signals shown in Figure 4.6. Also note that talker 1's pitch track varies from 125-150 Hz, while the talker 2's pitch track varies around 100-125 Hz. There exists an apparent anomaly in talker 2's pitch track around sample 16000. An expanded graph around this sample point is shown in Figure 4.7. Talker 2's pitch track was continuous before and after this section, and the time sequence shows definite structure, but the ML pitch detector failed to calculate an appropriate pitch value for this section of speech. Further examination of this segment of speech shows that the correct pitch for this segment should be in the range of 100-110 Hz. Other than this apparent anomaly, the ML pitch detector appeared to provide accurate pitch detection in subsequent use.

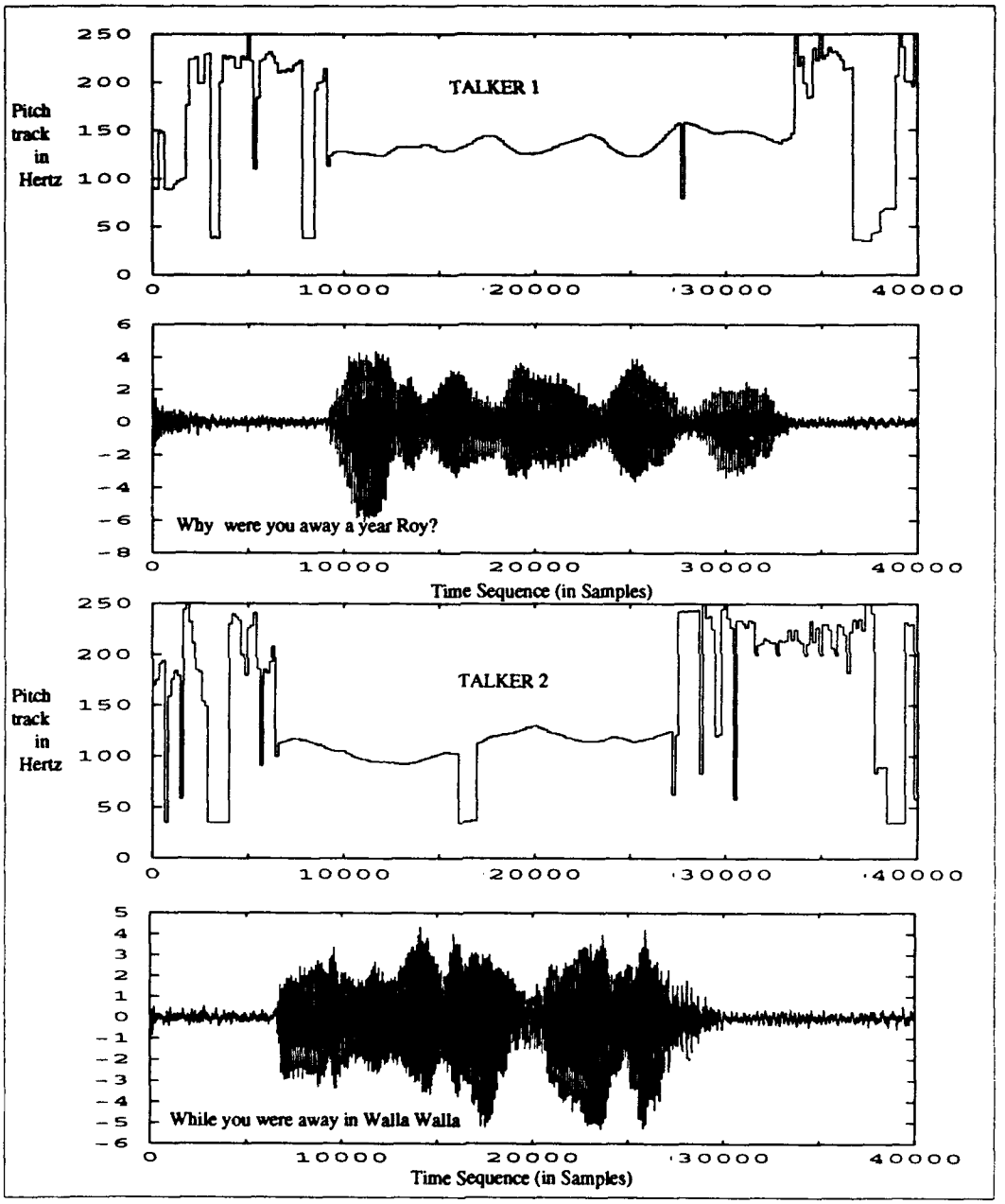


Figure 4.6. Time and Pitch Tracks - Recorded Talker 1 and Talker 2

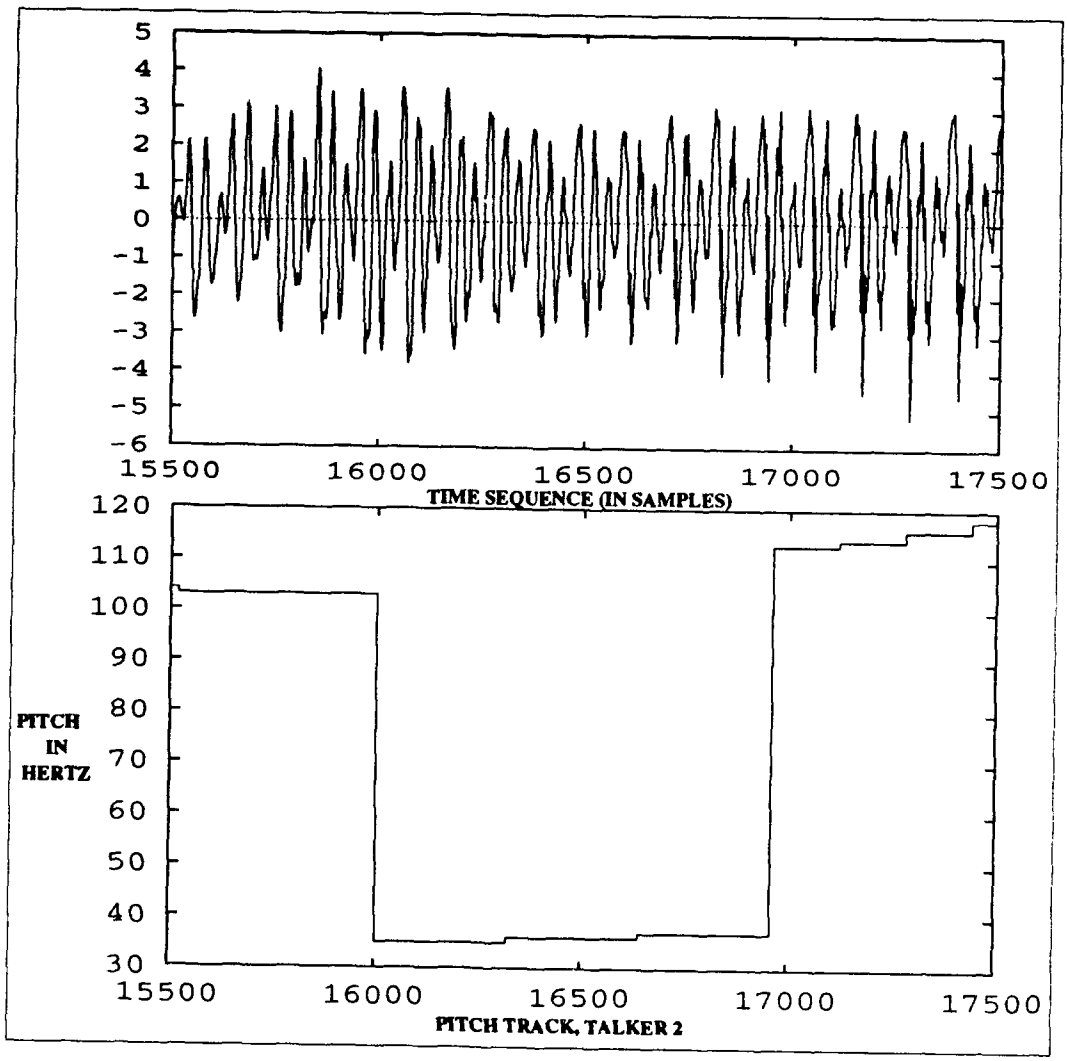


Figure 4.7. Irregularities in ML Pitch Tracker, Recorded Talker 2

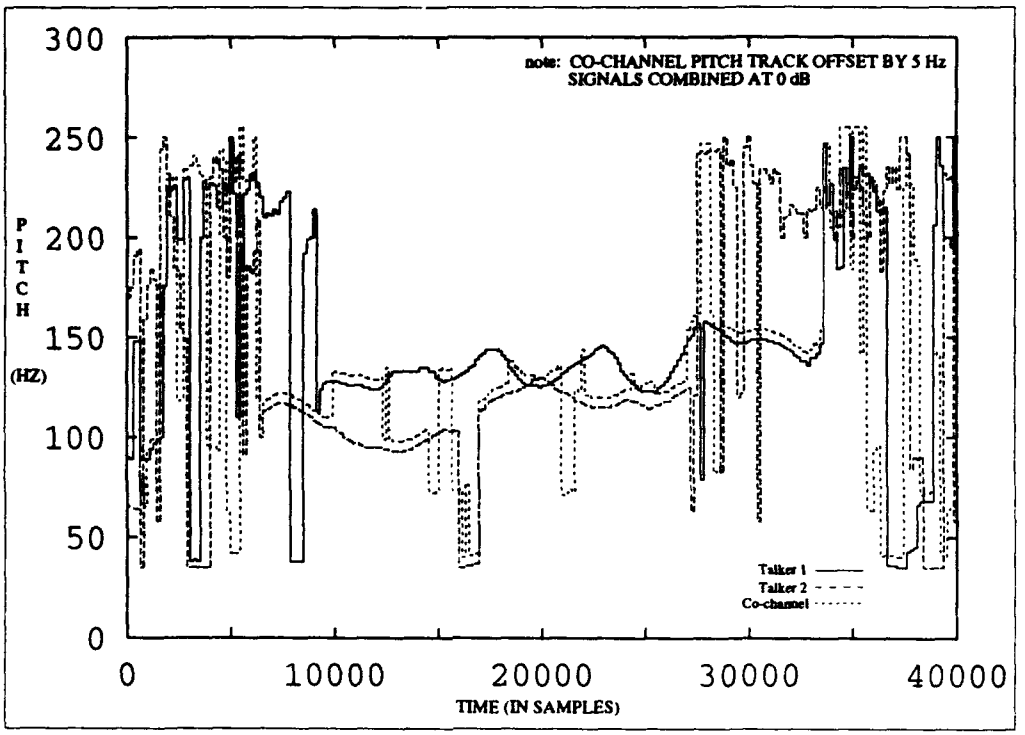


Figure 4.8. Recorded Co-Channel Pitch Tracks, 0 dB SSR

Next a test co-channel signal was created from talker 1 and talker 2 at 0 dB SSR, and the pitch track of this co-channel signal was calculated. A plot of the 0 dB co-channel pitch track, along with the two individual pitch tracks of talker 1 and talker 2 are provided in Figure 4.8. As noted in the figure, the computed co-channel pitch track value was usually one of the two talker's individual pitch tracks. However, the co-channel pitch track did have periods where the calculated pitch value did not track with either talker's pitch track. These portions of the co-channel signal were manually examined, and the pitch was determined. The results of this investigation revealed the ML pitch detector was accurate in its pitch prediction. It just happened that the time domain addition of these speech tracks resulted in a different pitch calculation.

Input Signals SSR Talker 1/Talker 2	Target Post Processed Signals:	
	Talker 1	Talker 2
+5 dB	Good	Marginal
0 dB	Good	Good
-5 dB	Marginal	Good

Table 4.3. Listening Test of Recorded Signals through the "Pitch Deviation"

The three test cases of recorded co-channel signals (+5, 0, & -5 dB SSR) for talker 1 and talker 2 were processed through the spectral separation algorithm and the recovered output speech signals were obtained using the "pitch deviation" method. Personal listening of the results of the "pitch deviation" method are shown in Table 4.3. Three levels of intelligibility for the personal listening test were assigned: good, marginal, and poor. A post-processed speech signal was deemed "good" if the target talker was clearly understandable, and the interfering talker was barely noticeable. A post-processed speech signal marked "marginal" means the target talker was mostly intelligible, but the interfering talker could also be heard. A "poor" post-processed speech signal was not intelligible and was worse than the original co-channel signal.

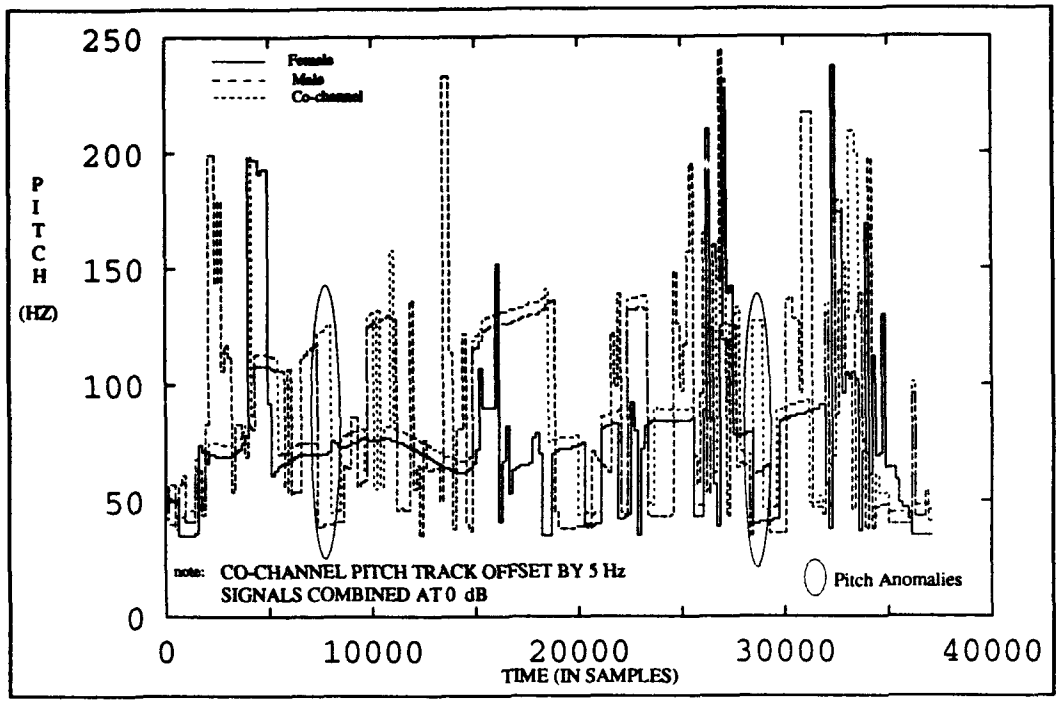


Figure 4.9. TIMIT Co-Channel Pitch Tracks, 0 dB SSR

The TIMIT female/male signals were combined to make test co-channel signals at +5, 0, and -5 dB SSR. The individual pitch tracks for both the female and male talkers were calculated. The co-channel pitch track at 0 dB SSR was also calculated. These three pitch tracks are shown in Figure 4.9. As shown in Figure 4.9 the co-channel pitch track followed one of the individual talker's pitch tracks for most of the test signal. Only a small portion of the graph showed areas where the co-channel pitch value obtained a value that was vastly different than either the female or male's calculated pitch value.

These three TIMIT test co-channel signals at +5, 0, and -5 dB SSR were processed through the "pitch deviation" algorithm. The results of a personal listening test is provided in Table 4.4. These results will be used to compare with the results obtained from the next section.

Input Signals SSR Female/Male	Target Post Processed Signals:	
	Female	Male
+5 dB	Good	Marginal
0 dB	Good	Good
-5 dB	Marginal	Good

Table 4.4. Listening Test of TIMIT Signals through the "Pitch Deviation"

The "pitch deviation" metjhod assigning the separated segments co-channel data proved successful. However, a poor or erroneous pitch value (noted in Figures 4.8 and 4.9 will greatly affect the spectral separation algorithm's ability to separate the two talker's energy–and ultimately degrade the post-processed speech signals. Further investigation in these erroneously calculated pitch values may be warranted–if the hope is to have better spectral separation. None the less, since this algorithm is purely academic co-channel speech processing, further research into algorithms that will yield better pitch distance calculations, is unwarranted.

The listening results provided in Tables 4.3 and 4.4 for the "pitch deviation" method coincide with the expected results reported by L. Lee and Morgan. These results will be used as a baseline to compare with the results obtained in the LPC based assignment algorithms discussed in the following sections. This section provided the results of the "pitch deviation" method of assigning the separated segments of the co-channel signal. These results are to be interpreted as a baseline for comparison with the results described in the next sections from the LPC based distortion method. Spectrogram plots of the resulting separated speech signals listed in Tables 4.3 and 4.4 are provided in Appendix B.

4.4.2 **LPC Based Distortion Method:** ***A Priori*** **Sentences.** This LPC based distortion algorithm uses the minimum-prediction residual distortion metric from equation 3.3 in order to base the decision rule of assigning the "stronger" and "weaker" separated segments of speech from the co-channel signal. The distinction for the *a priori* algorithm in this section is the "clean" speech signals that were used to create the co-channel signals are also used to precompute the model LPC vectors. This methodology was purely implemented as a *proof-of-concept* for the LPC based distortion spectral assignment algorithm.

Comparisons are made between this methodology and the "pitch deviation" methodology described in Section 4.4.1. The particular differences noted between these two methodologies are the decisions made to assign the i^{th} "stronger" and "weaker" segments. No analysis was done on the separated segments of speech to determine the amount of energy associated with the particular talker. If this analysis was done, it is probable that the true decision of the assignment of the separated segments to the appropriate talker could be made. The absolute measure of the success of the spectral assignment methodologies would be an extensive listening test–which is not performed in this thesis.

By using the same *a priori* individual speech signals as in the co-channel test signals to precompute the model LPC vectors for the computation of the minimum-prediction residuals, this algorithm should yield the smallest attainable distortion values during the process. Figure 4.10 shows the minimum-prediction residual values computed for each frame for both talkers in the 0 dB SSR TIMIT co-channel signal. The plot is interesting in that it shows a relatively small residual error was found for each talker, regardless of whether the energy in the "stronger" frame was from either talker. That is, no large numerical values occurred in computing the residual between the test signal and the male or female model data.

The three TIMIT test co-channel speech signals were processed through the *a priori* LPC based spectral assignment algorithm. Figure 4.11 shows the decision points made by the *a priori* LPC residual method and the "pitch deviation" method. From this figure, the *a priori* LPC residual assignment decisions closely tracked the "pitch deviation" assignment decisions

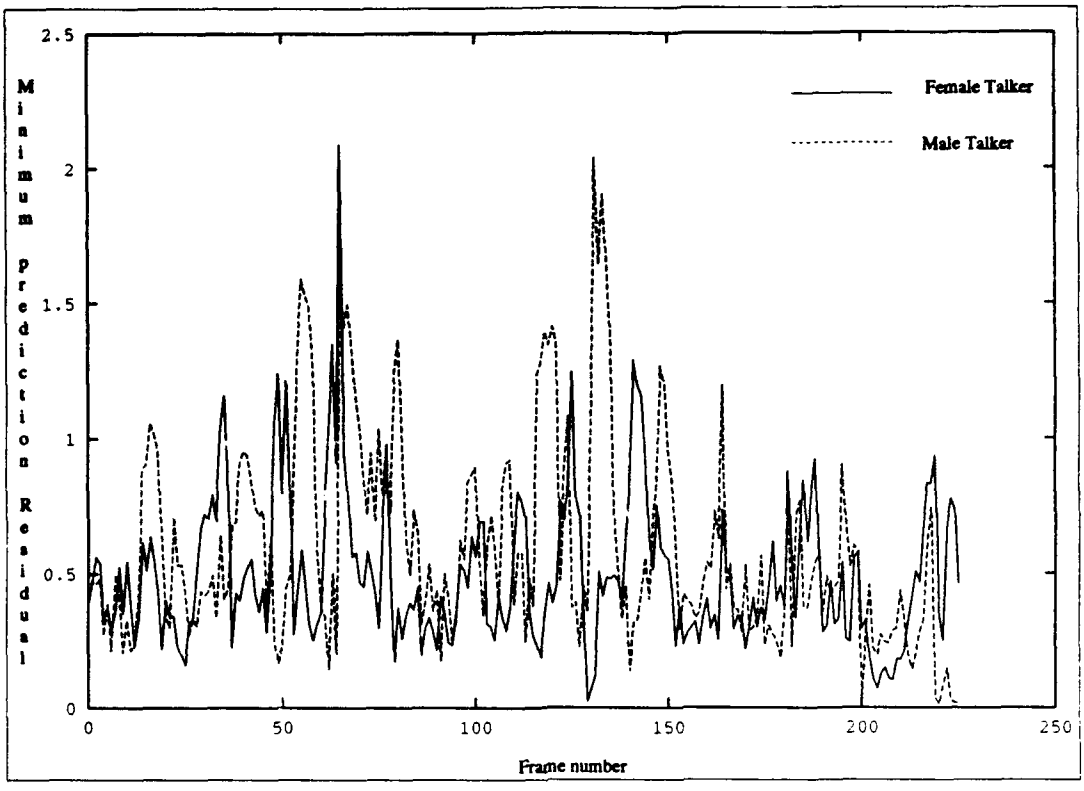


Figure 4.10. Minimum-Prediction Residual Values at 0 dB SSR TIMIT Co-Channel Signal

for all SSR's tested. It should be noted that several frames in the beginning and end contain noise only.

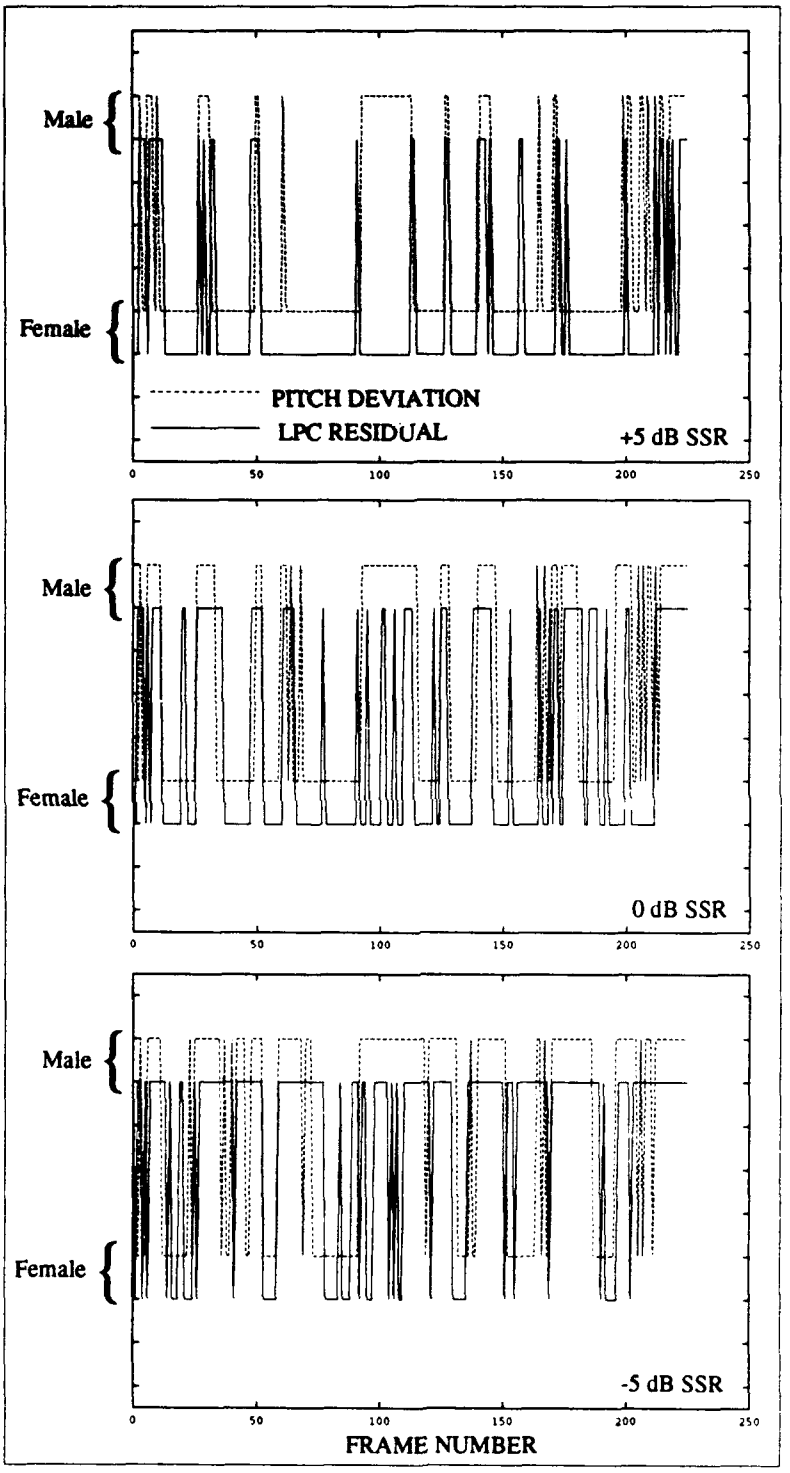


Figure 4.11. Frame Assignment for +5, 0, & -5 dB SSR (Female/Male) TIMIT Co-Channel Signals

For the *a priori* LPC distortion spectral assignment case, the model LPC vectors came from the same talkers as in the co-channel signal, and an expected linear relationship should exist between the test frame processed, and the i^{th} model LPC vector used to compute the minimum distortion. Figure 4.12 shows which precomputed i^{th} LPC vector that was used from the model database of each talker to provide the minimum-prediction residual for the 5 dB SSR case (female/male). The graphs in Figure 4.12 show an expected linear relationship between the co-channel frame tested and the LPC vector used to compute the minimum LPC residual. The graphs illustrate the LPC residual method is accurately computing the minimum distortion between the separated test co-channel signal and the model data. This linear relationship is most notable in the female talker's graph since she had a 5 dB SSR advantage over the male talker. The male talker's graph also showed a slight linear relation. The slight linear relation in the male talker's graph is attributed to the fact that even though the male talker was 5 dB below the female talker, some segments in the co-channel signal had predominant energy that was attributed to the male talker, and hence some "stronger" segments were assigned to the male talker. A similar linear relationship exists for the minus 5 dB SSR case.

Figure 4.13 shows which precomputed i^{th} LPC vector that was computed for each talker to provide the minimum-prediction residual for the 0 dB SSR case. A fair linear relation exists for both talkers in this case. This fair linear relation is expected since the signal energies are equal and vary. Additionally, in the training of the LPC vectors, the lower energy frames were discarded and the absence of these frames provides "holes" in the expected linear relationship.

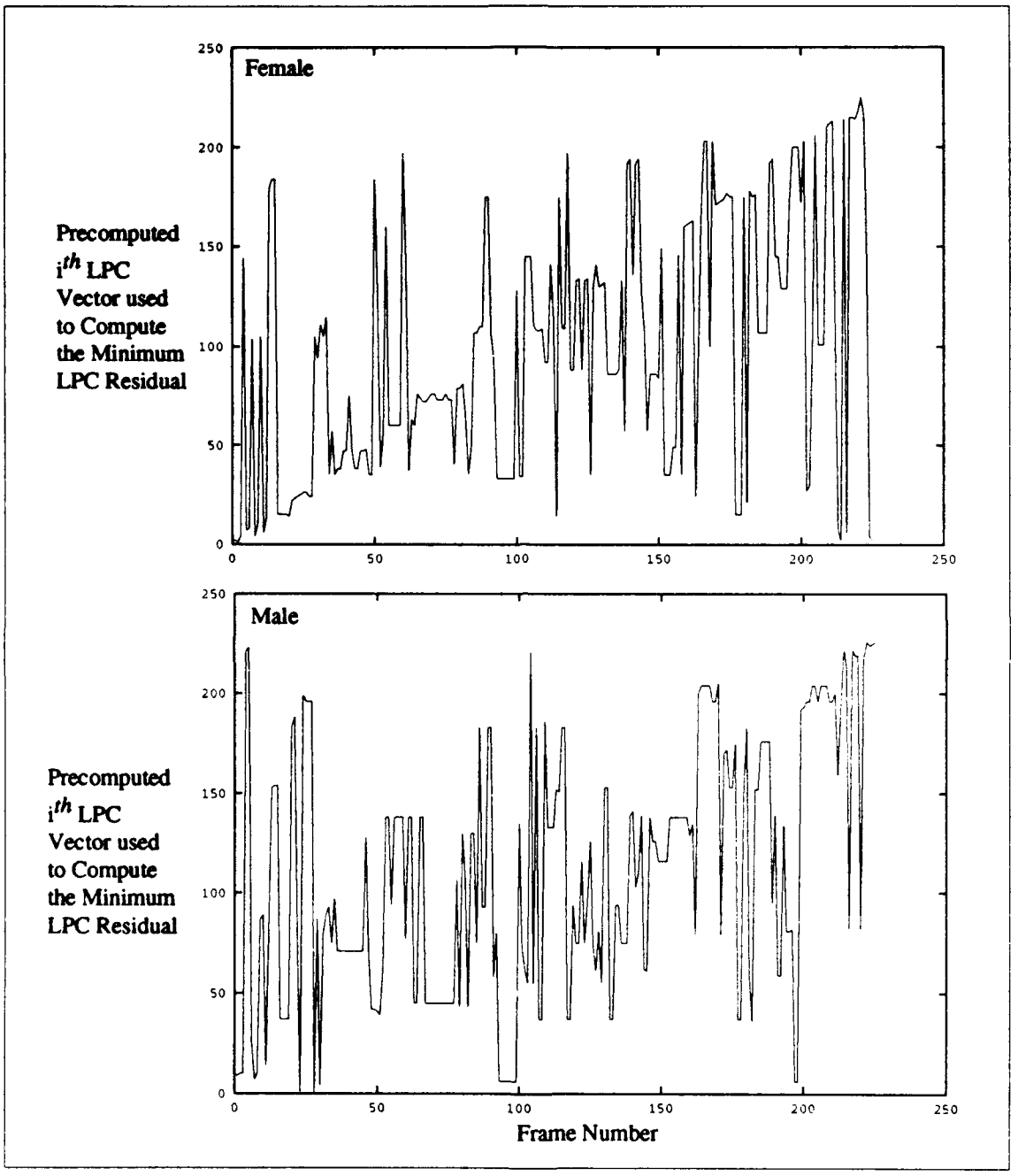


Figure 4.12. LPC Vectors Selected in Computing the Minimum-Prediction Residual, TIMIT +5 dB SSR

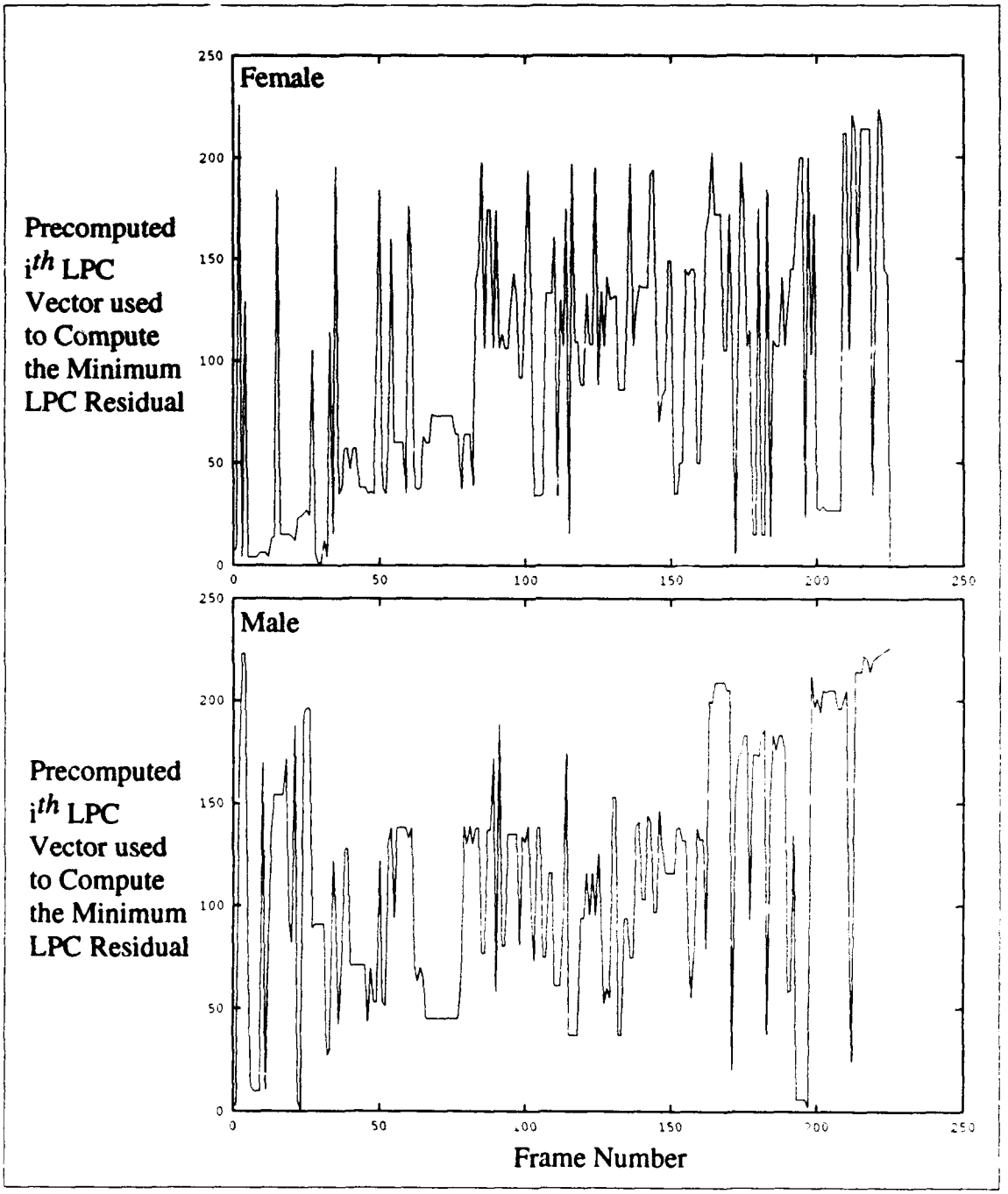


Figure 4.13. LPC Vectors Selected in Computing the Minimum-Prediction Residual, TIMIT 0 dB SSR

Input Signals SSR Female/Male	Target Post Processed Signals:	
	Female	Male
+5 dB	Good	Marginal
0 dB	Good	Good
-5 dB	Marginal	Good

Table 4.5. Listening Test of TIMIT Signals through the *A Priori* LPC Based Assignment Algorithm

Personal informal listening tests were performed on the three TIMIT co-channel test signals. The results are provided in Table 4.5. These results show that the *a priori* LPC based asssignment algorithmn provided good recovery for equal power SSR signals and above. these results are not as good as the "pitch deviation" method, but they confirm the *proof-of-concept* for the LPC based residual assignment algorithm is successful. Spectrogram plots for the recovered speech signals listed in Table 4.5 are provided in Appendix B.

Next, the three recorded test co-channel signals were processed by the *a priori* LPC residual spectral assignment method. Figure 4.14 shows the assignment selections for the "pitch deviation" method and the *a priori* LPC residual method for each co-channel frame processed. From the graphs shown in Figure 4.14, the LPC residual method tracked very closely with the "pitch deviation" assignment methodology. Only in a few instances were frames not assigned to the same person for both methodologies.

Figure 4.15 show the linear relationship between the i^{th} LPC vector chosen to compute the minimum-residual, and the segment of co-channel speech processed. The linear relationship is very noticeable for Talker 1 (who had a 5 dB advantage), and the linear relationship is hardly noticeable for talker 2.

Figure 4.16 again shows the expected linear relationship between the i^{th} model LPC vector and the segment of co-channel of speech processed, this time for the 0 dB SSR case. This case shows the least linear relationship, and this is probably caused by the LPC vectors

that were discarded during the pre-computation, and the fact that the recorded signals had around a 18 dB SNR.

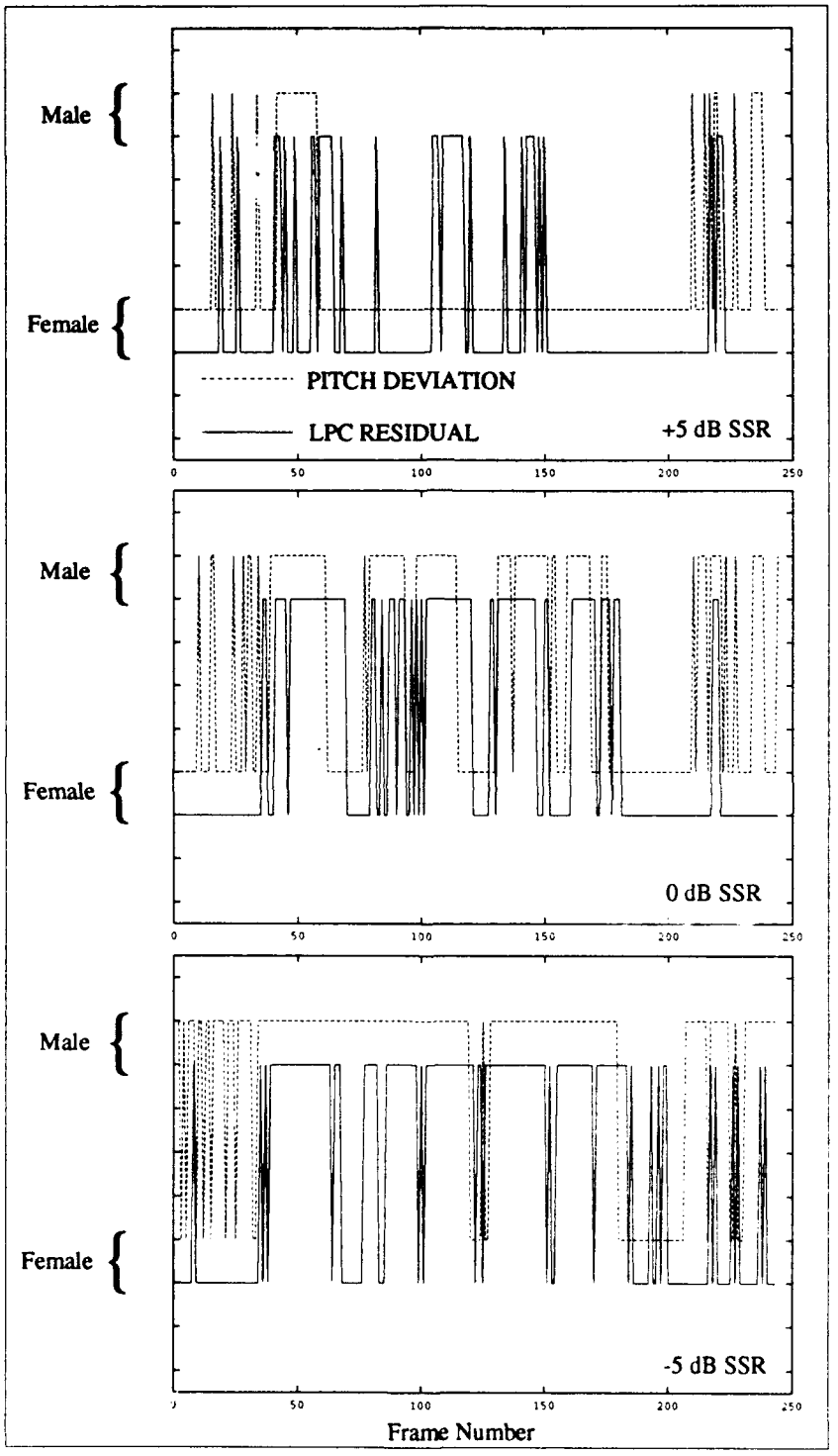


Figure 4.14. Frame Assignment for +5, 0, & -5 dB SSR (Talker 1/Talker 2) Recorded Co-Channel Signals, *A Priori* LPC Based Method

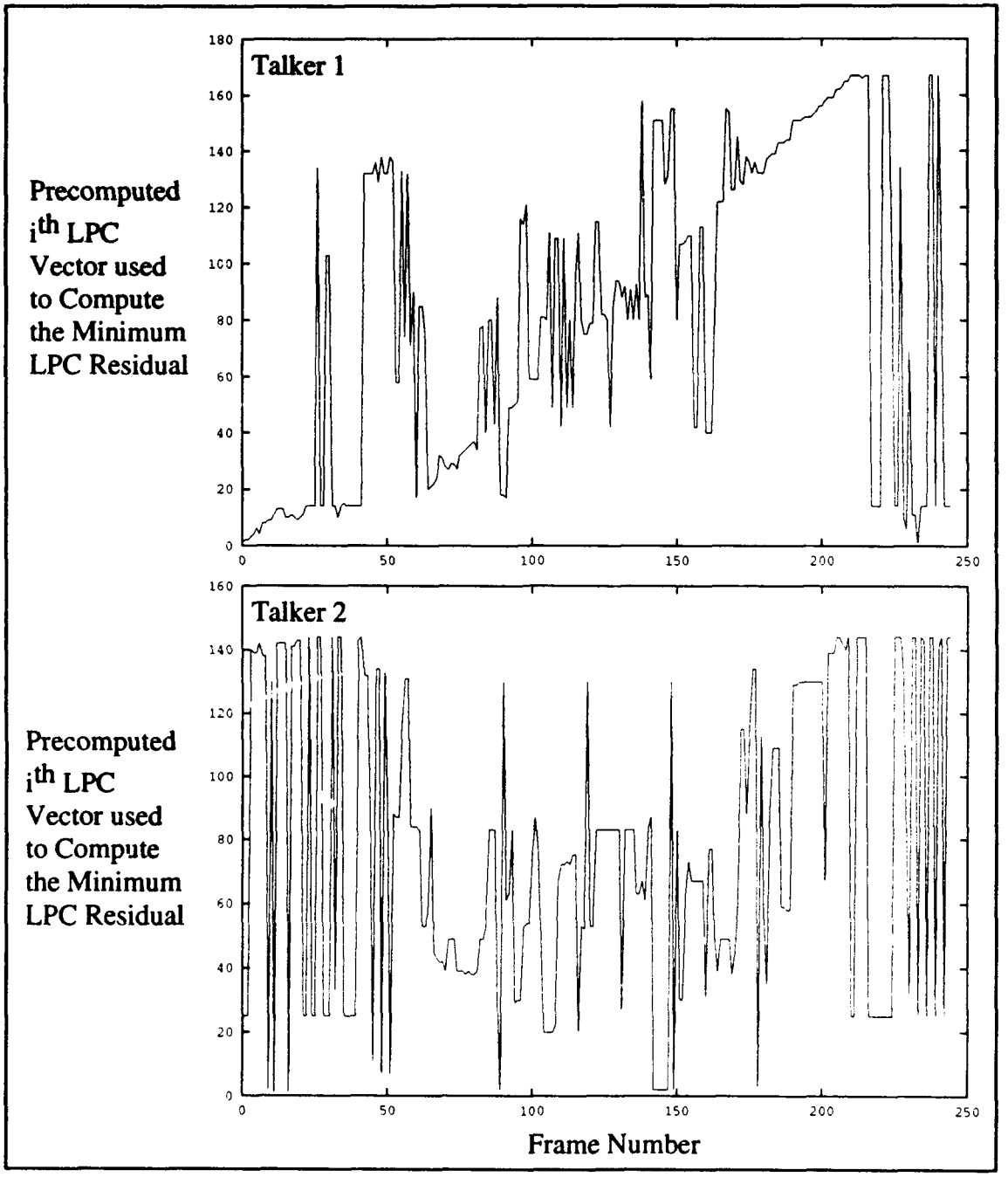


Figure 4.15. LPC Vectors Selected in Computing the Minimum-Prediction Residual, Recorded +5 dB SSR

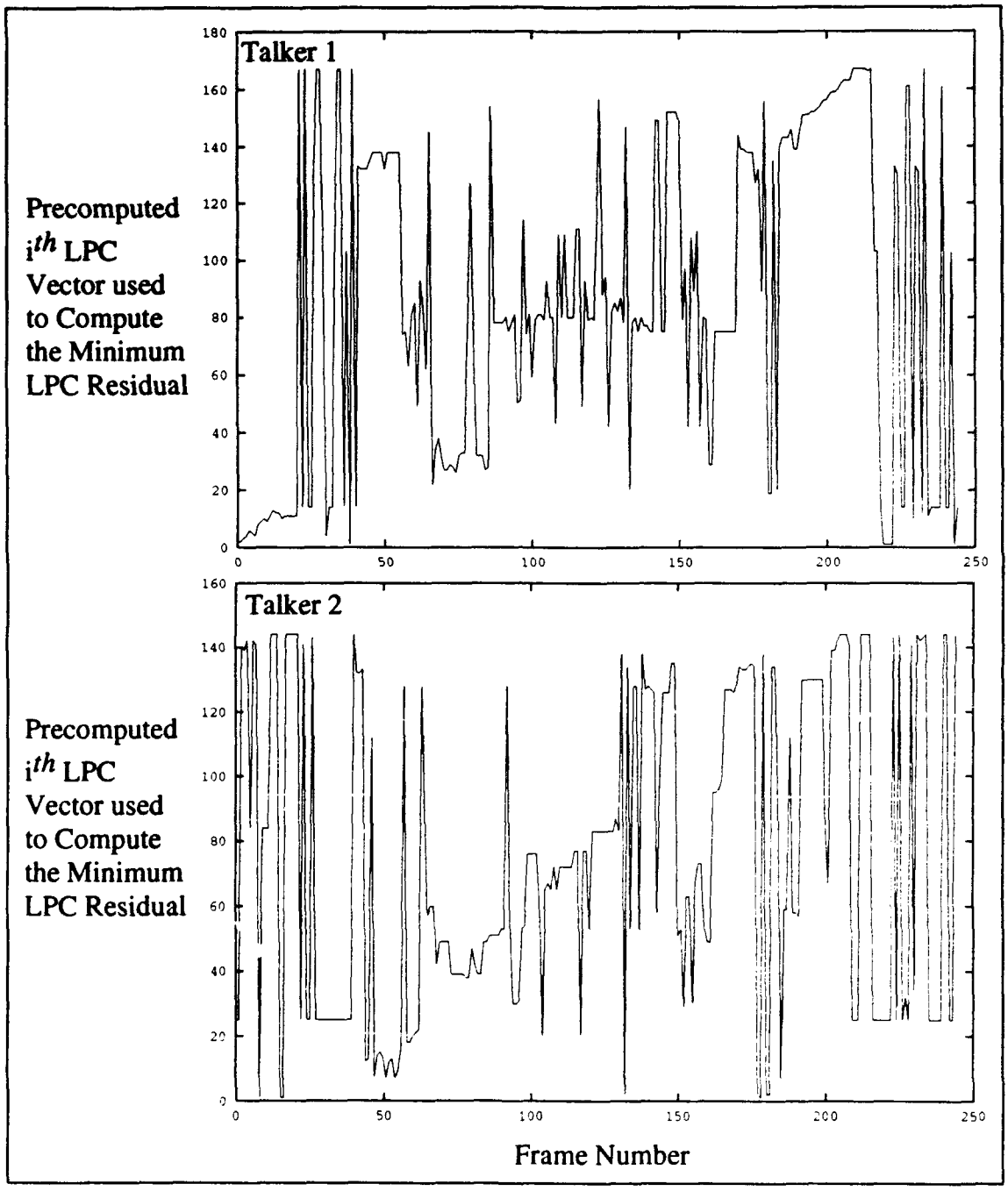


Figure 4.16. LPC Vectors Selected in Computing the Minimum-Prediction Residual, Recorded 0 dB SSR

Input Signals SSR Talker 1/Talker 2	Target Post Processed Signals:	
	Talker 1	Talker 2
+5 dB	Good	Marginal
0 dB	Good	Good
-5 dB	Marginal	Good

Table 4.6. Listening Test of Recorded Signals through the *A Priori* LPC Based Assignment Algorithm

The results of personal listening to the three post-processed recorded test co-channel speech signals are provided in Table 4.6. As indicated in Table 4.6, the talker 1 and talker 2 co-channel speech signals were separated nearly as well as the TIMIT test co-channel signals. Spectrogram plots are provided in Appendix B of the recorded recovered talkers listed in Table 4.6.

In examining this *proof-of-concept* test, the results have shown good success for the LPC based assignment methodology against the suite of test co-channel speech signals considered. Since the TIMIT speech signals enjoyed about a 15 dB SNR advantage over the recorded speech signals. as noted, these signals were separated better than the recorded signals, and the informal listening test confirmed the separated TIMIT co-channel signals sounded better. Because the *proof-of-concept* methodology was successful, the next test case is investigated where the model LPC vectors are computed from model speech signals. This "excluded sentences" LPC based spectral assignment methodology is discussed in the next section.

4.4.3 **LPC Based Distortion Method: Excluded Sentences.** In the "excluded sentences" LPC based distortion spectral assignment methodology, the speech signals that were used to pre-compute the model LPC vectors came from speech that was **not** part of the co-channel speech. Thus the minimum-prediction residual (equation 3.3) is computed between the "stronger" segment of separated co-channel speech and model LPC vectors that were calculated from model sentences that were not part of the co-channel signal, but from the same talkers in the co-channel signal.

Input Signals	Target Post Processed Signals:	
SSR Talker 1/Talker 2	Talker 1	Talker 2
+5 dB	Good	Poor
0 dB	Marginal	Marginal
-5 dB	Poor	Good

Table 4.7. Listening Test of Recorded Signals through the LPC Based Assignment Algorithm

Figure 4.17 shows the frame assignment decisions for the three test recorded co-channel signals. The plots shown in Figure 4.17 show the "excluded sentences" spectral assignment methodology made nearly the same decisions as the "pitch deviation" methodology for the plus 5 dB SSR case, and many of the same assignment decisions for the - 5dB and 0 dB SSR cases.

The results of personal listening to the post-processed recorded test co-channel speech signals are provided in Table 4.7.

Next, the three TIMIT test co-channel speech signals were processed by the "excluded sentences" LPC based spectral assignment methodology.

Figure 4.18 shows the frame assignment decisions for the three test TIMIT co-channel signals. The plots shown in Figure 4.18 show the "excluded sentences" spectral assignment methodology made nearly the same decisions as the "pitch deviation" methodology for the plus 5 dB SSR case, and many of the same assignment decisions for the - 5dB and 0 dB SSR cases.

The personal listening test results are provided in Table 4.8. These results show good recovery of both talkers at SSR levels above 0 dB. The recovered speech from the "excluded" entences" LPC based assignment algorithm was clearly understandable, and the background talkers voice was reduced to a low murmor.

Input Signals SSR Female/Male	Target Post Processed Signals:	
	Female	Male
+5 dB	Good	Poor
0 dB	Marginal	Marginal
-5 dB	Poor	Good

Table 4.8. Listening Test of TIMIT Signals through the LPC Based Assignment Algorithm

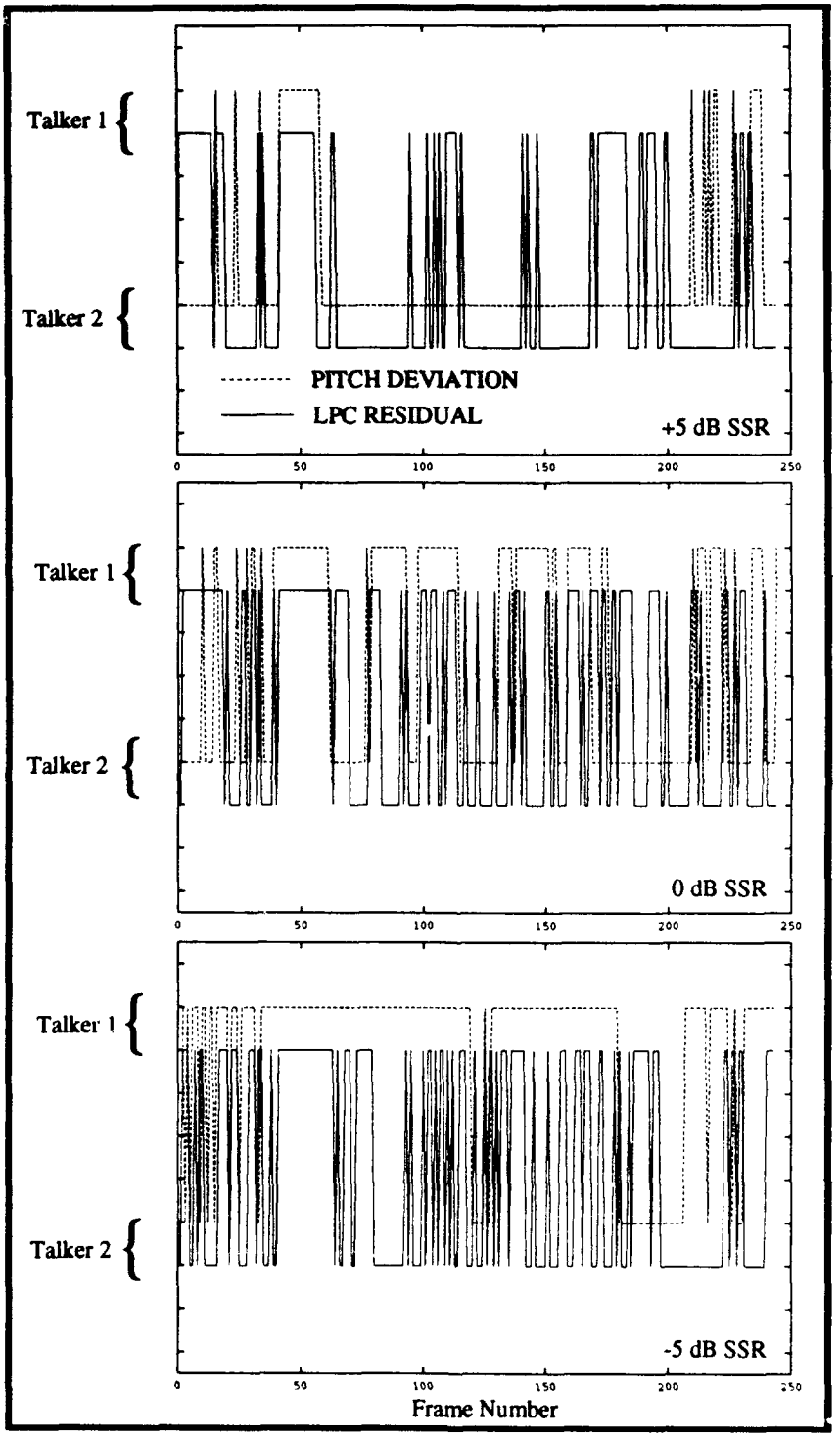


Figure 4.17. Frame Assignment for +5, 0, & -5 dB SSR (Talker 1/Talker 2) Recorded Co-Channel Signals, "Excluded Sentences" LPC Based Method

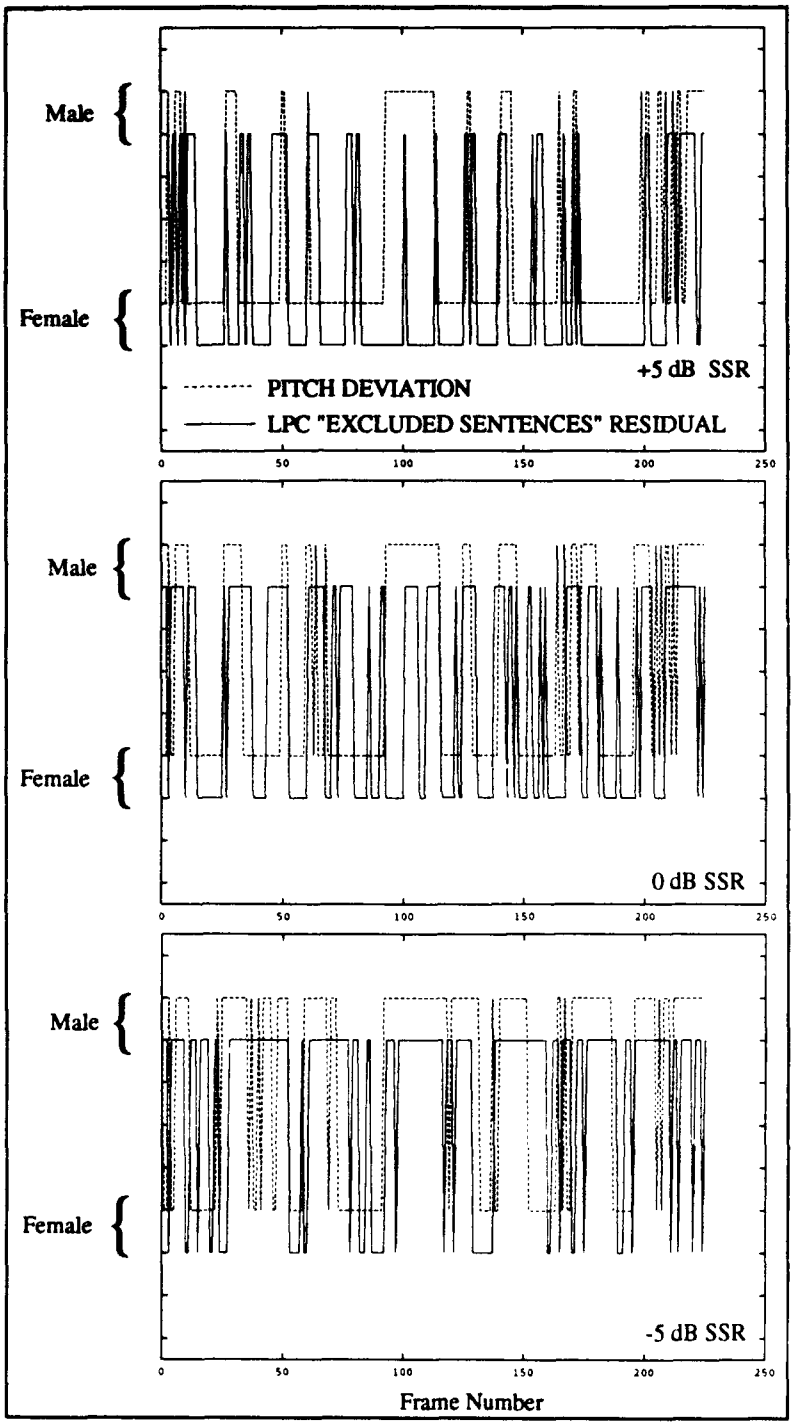


Figure 4.18. Frame Assignment for +5, 0, & -5 dB SSR (Talker 1/Talker 2) TIMIT Co-Channel Signals, "Excluded Sentences" LPC Based Method

Target Talker's SSR	*A Priori* "Pitch Deviation"	*A Priori* LPC	"Excluded Sentences" LPC
+5 dB	Good	Good	Good
0 dB	Good	Good	Marginal
-5 dB	Marginal	Marginal	Poor

Table 4.9. Summary of Listening Tests

4.5 Summary of Spectral Assignment Methodologies.

Three spectral assignment methodologies were investigated: the "pitch deviation", and the minimum-prediction residual (*a priori* sentences and "excluded sentences"). Each methodology separated and recovered the target talkers in the test cases of co-channel speech. The significance of the "excluded sentence" spectral assignment methodology was the fact that the sentences used to compute the model LPC vectors was **independent** of the speech in the co-channel signal. The importance of this technique is that if the operator of the co-channel separation processing algorithm can obtain "clean" speech from the target talkers, then this methodology can be used when the target talkers are interfered by another interfering speech signal. Thus no *a priori* information about the co-channel signal is needed to extract the target talker(s). The *a priori* pitch deviation assignment algorithm performed the best since the ML pitch detector is robust in noise, and this had the advantage of *a priori* information.

Table 4.9 provides a summary of the informal listening test against the target co-channel test signals processed by the three spectral assignment methodologies. The SSR column represents the dB value of the target talker in the co-channel speech signal.

V. Conclusions and Recommendations

This chapter provides a summary of this thesis effort, and suggests areas of further research into the co-channel speaker separation problem.

5.1 Conclusions.

The co-channel speaker separation problem remains a challenging task for signal processing applications. The inherent non-stationary properties of speech, the limited bandwidth occupation, the varying energy levels, and inherent large computational requirements, each contribute to the complexity of the co-channel speaker separation problem.

The LPC based distortion assignment algorithm developed in this thesis is an important processing technique that can help solve the co-channel speaker separation problem. The experiments in this thesis showed that the methodology, when placed in conjunction with the spectral energy separation algorithm provided by L. Lee and Morgan (11) provided a viable alternative processing technique to their *a priori* pitch based assignment algorithm. If the LPC based distortion assignment algorithm resulted in separating the co-channel signal perfectly, the processing burden is still severe.

The "excluded sentences" LPC based assignment algorithm performed nearly as well as the *a priori* pitch deviation algorithm. The "excluded sentences" algorithm could only accurately assign the separated co-channel speech segment at SSR levels greater than zero dB.

5.2 Recommendations.

The following items are suggested areas for future research and development in support of the co-channel speaker separation problem.

1. An area of further research is the system described by Naylor and Porter (17). Their co-channel separation technique was unique in that it did not require any *a priori* information.

2. Use clustering routines to further reduce the amount of pre-computed model LPC vectors.

3. Determine if there is an "optimum" amount of pre-computed model LPC vectors. Is there a finite amount of phonemes, for a given talker or language, that would make the pre-computed set of model LPC vectors complete?

4. Determine the usefulness of the co-channel separation processing to a speaker identification process. Given the co-channel separation processing algorithm discussed in this thesis, investigate whether current speaker identification techniques can accurately determine who the post-processed (separated) co-channel speakers are.

5. Investigation into other spectral distortion metrics. Is there a distortion metric other than the minimum-prediction residual that might be applied to the assignment of separated co-channel speech?

6. The LPC based distortion metric is inherently not robust to noise, and maybe some noise-reducing techniques could be applied to the co-channel signal prior to attempting co-channel separation.

7. Developing an adaptive window size and window stepping scheme for the co-channel processing algorithm. A larger window size will allow more time-domain data and hence better frequency domain analysis. A larger window step size would speed up the overall processing. The varying window size must not exceed the requirements for stationary of the frame considered.

8. Although Matlab is an excellent signal processing tool, the severe computational requirements of any co-channel speaker separation algorithm would require programming in C, or possible implementation in hardware to approach real-time processing. The approximate time to fully pre-compute the parameters and process a short sentence of co-channel speech requires several hours (2-3) processing in Matlab on a Sun Sparc2 workstation.

Appendix A. Data Conversion and Software Programs

This thesis effort requires the processing of information using a digital computer. Specifically, a Sun Sparc-2 workstation was used in conjunction with the following commercial/public domain software programs, self-generated software programs, and speech data files:

1. Matlab, by *The MathWorks, Inc.* is a high-performance interactive software package for scientific and engineering numeric computation, used in the co-channel signal processing algorithms.
2. Ariel S-32C Digital Signal Processor, and ProPort A/D and D/A converter and supporting record/playback utility software.
3. Entropic Signal Processing System (ESPS), version 4.1, by *Entropic Research Laboratory, Inc.* is a suite of programs for creating, manipulating, and analyzing digital signals.
4. *waves+*, version 2.0, by *Entropic Research Laboratory, Inc.* is an interactive graphics interface used to display the ESPS data signals and files.
5. TIMIT Data Files, a standard speech data set from the Defense Advanced Research Projects Agency (DARPA), sampled at 16 kHz, 16 bits/sample, linearly quantized.
6. SOund eXchange (SOX) A program called SOX (SOund eXchange) was used in converting the binary data files between the specific formats. SOX is a public domain software package.

A.1 Speech Data Files: Format and Description

The speech data files used in this research effort originate from two sources. The first source is the TIMIT speech files, and the other source is self-generated speech data files, generated directly from the Sparc-2 workstation using the Ariel S-32C DSP/ProPort A-D

converter. Fortunately for standardization and file conversion, both of these speech data file types are at a 16 kHz sampling rate and linear quantization.

The TIMIT speech data files are used to artificially create various co-channel speech signals analyzed in this thesis. The complete TIMIT acoustic-phonetic speech data base was designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. The data was prepared at the National Institute of Standards and Technology (NIST), with sponsorship from the Defense Advanced Research Projects Agency - Information Science and Technology Office (DARPA-ISTO). The individuals who spoke the sentences in the TIMIT data base come from varying ethnic backgrounds, education level, ages, and geographic location. The TIMIT data base includes 168 speakers, each speaker recording 10 sentences. Eight of these ten sentences are suggested to be used for test/training purposes, whiie two sentences from each talker are sample sets. A male talker and a female talker were selected arbitrarily to form the co-channel signal, and the training codebook.

The Ariel/ProPort A/D and D/A converter was used to record sample speech data files, and was used to playback the co-channel speech signals (before and after processing). The data format for the recording of speech signals was set to 16 kHz and 16 bits/sample linear quantization, in the standard Sun audio binary format.

A.2 Software Description

The two speech data file formats, and the Matlab co-channel processing code necessitated the data format conversion between these different platforms. A program called sox (SOund eXchange - universal sound sample translator) and some self generated C routines were used to convert between the speech data formats and MatLab data formats and vice-versa. Additional routines contained in the ESPS package enable the conversion to and from a Matlab binary file to an ESPS feature file or generation of spectrogram files.

A.3 **Data Conversion**

The TIMIT data files required the conversion between the CDROM (*.wav) binary data format to the (*.mat) binary format. Several conversion routines were included in the TIMIT CDROM for use in converting the binary speech files. A unix script file called "wav2mat" was implemented that converted the (*.wav) to *.mat). The program begins by reading the (*.wav) file header is read by a TIMIT utility program called h_read, and the number of samples in the speech file is obtained. Another TIMIT utility program called h_strip, strips off the binary (*.wav) file header, leaving a (*.raw) binary file. This (*.raw) binary file is then converted using the SOX utility program, where the (*.raw) data format is binary short words that have the most significant bit and least significant bit reversed. The SOX program is first used to reverse the data bits. Then C program is run that converts the (*.raw) binary file to a (*.mat) file.

To create a spectrogram plot, an encapsulated postscript .eps file from a Matlab .mat speech file, a Matlab/Unix script file called "creatfspec.m" was created. The following is file that contains the commands necessary to convert a Matlab speech file to a spectrogram plot. Basically, the program scales the speech file to the whole dynamic range of the D/A converter, and subtracts off the DC component. A file "test1.mat" is created, with the speech data saved under the variable test1. The Unix script file, "mat2fspec" is then executed that converts the test1.mat file to a test1.fspec spectrogram file. The script file then plots the spectrogram file to the screen, and the "xgrabsc" ESPS command is used to grab the spectrogram file and save the result in an encapsulated postscript file test1.eps.

```
%file: creatfspec.m
function factor=ampplay(sig)
amp=max(sig);
factor=(2^15-1)/amp;
test1=factor*sig;
test1=test1-mean(test1);
[row column]=size(test1);
if row == 1,
test1=test1';
end;
disp('You have achieved a standing vector!')
save test1.mat test1
!mat2fspec test1.mat
end;
```

Appendix B. **Spectrogram Plots**

This Appendix provides spectrogram plots of the processed speech signals. The dark bands within the spectrogram plots are the formant lines for the individual talker's recovered speech signal. These spectrogram plots are to be compared with the "clean" talker's spectrogram plots shown in Chapter IV

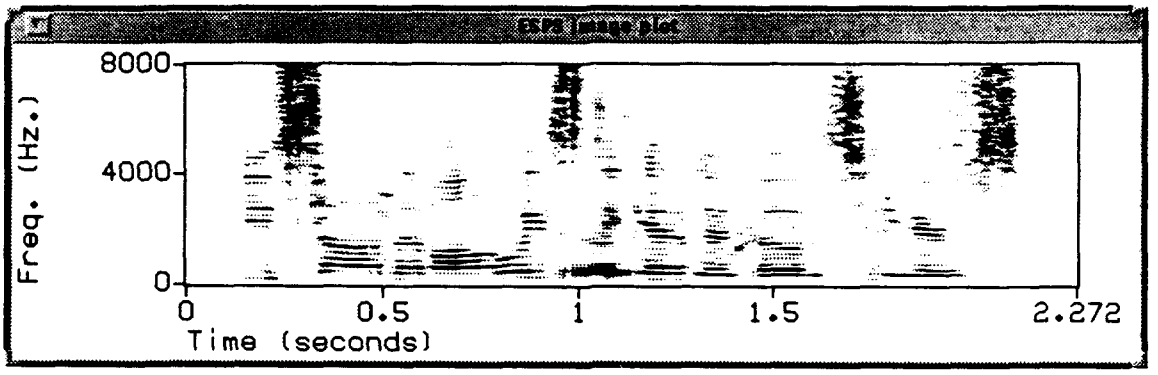


Figure B.1. Female, *A Priori* LPC Method, +5 dB, t111p5

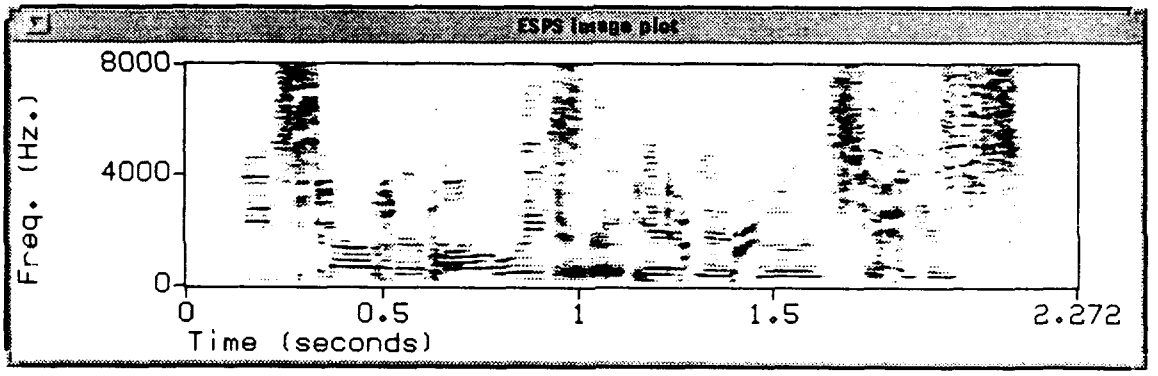


Figure B.2. Female, *A Priori* LPC Method, 0 dB, t111z

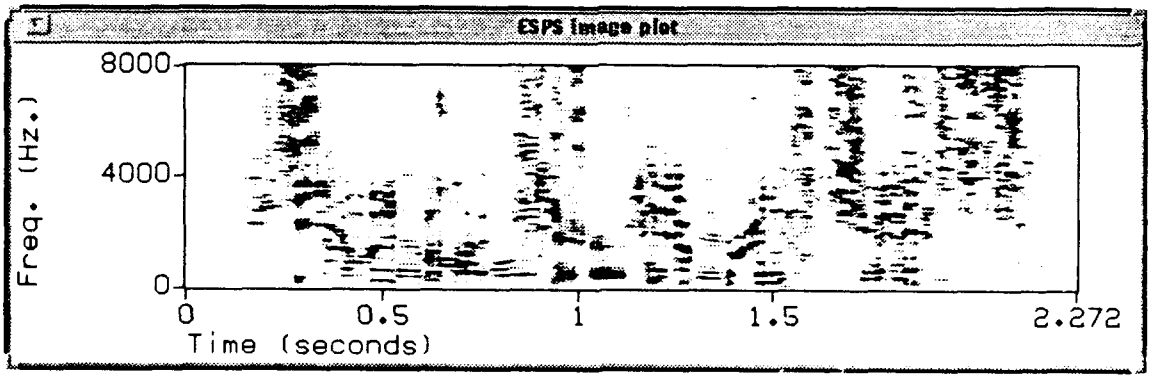


Figure B.3. Female, *A Priori* LPC Method, -5 dB, t111m5

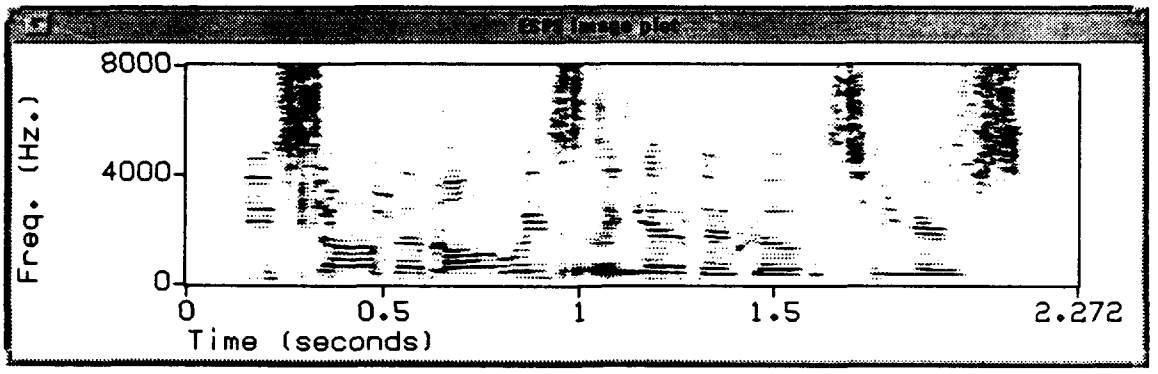


Figure B.4. Female, "Excluded Sentences", LPC Method, +5 dB, t117p5

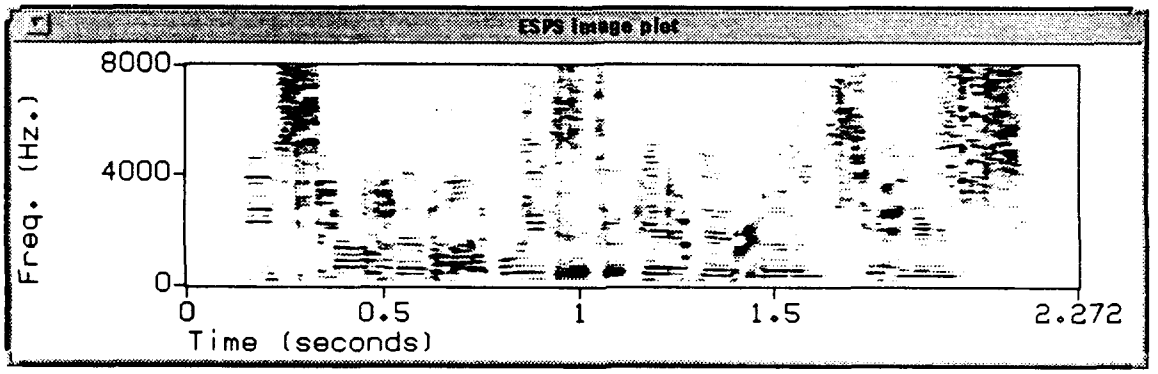


Figure B.5. Female, "Excluded Sentences", LPC Method, 0 dB, t117z

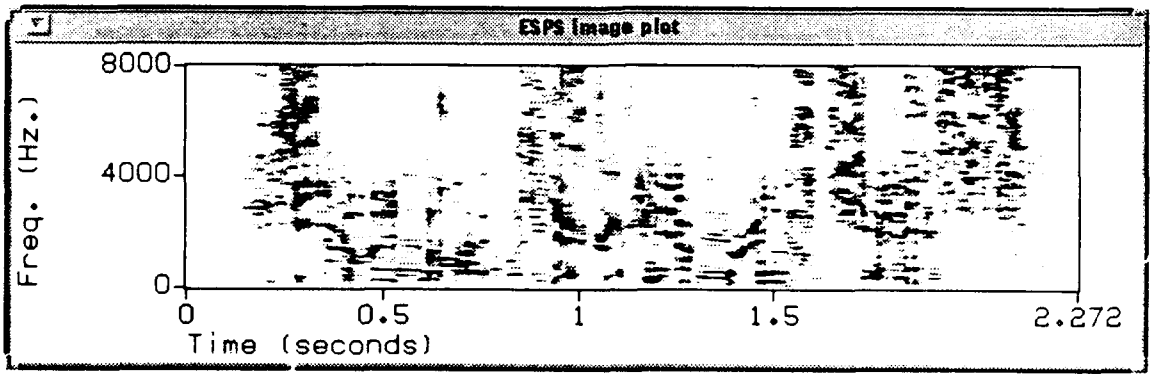


Figure B.6. Female, "Excluded Sentences", LPC Method, -5 dB, t117m5

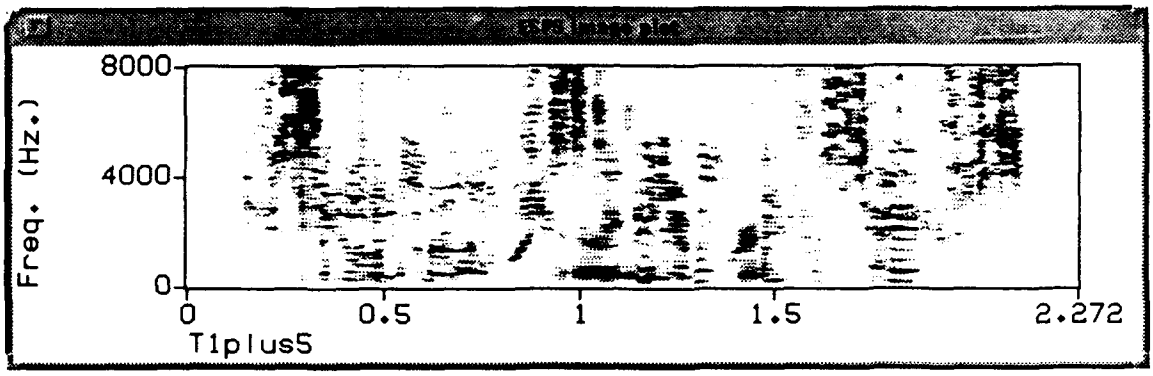


Figure B.7. Female, "Pitch Deviation" Method, +5 dB, t1plus5

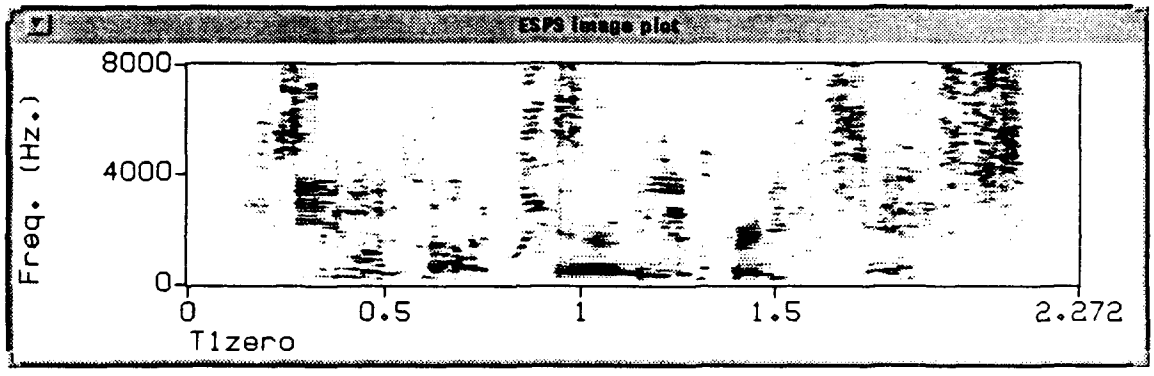


Figure B.8. Female, "Pitch Deviation" Method, 0 dB, t1pz

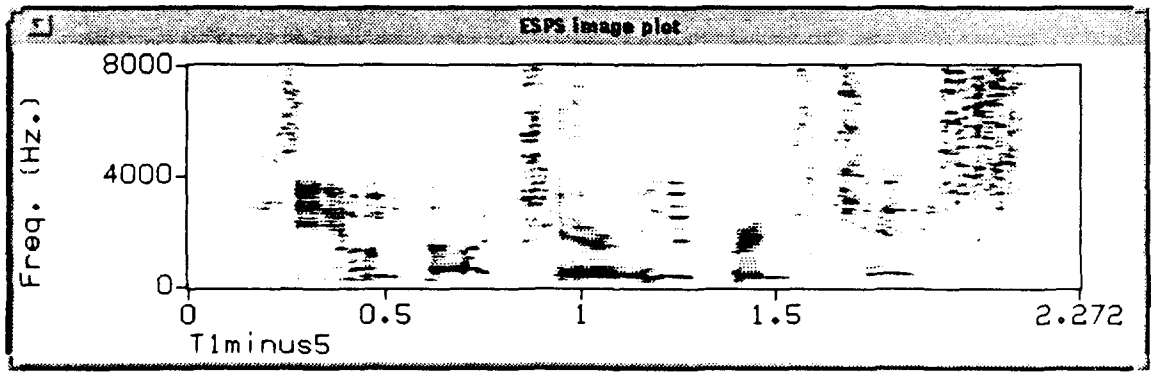


Figure B.9. Female, "Pitch Deviation" Method, -5 dB, t1minus5

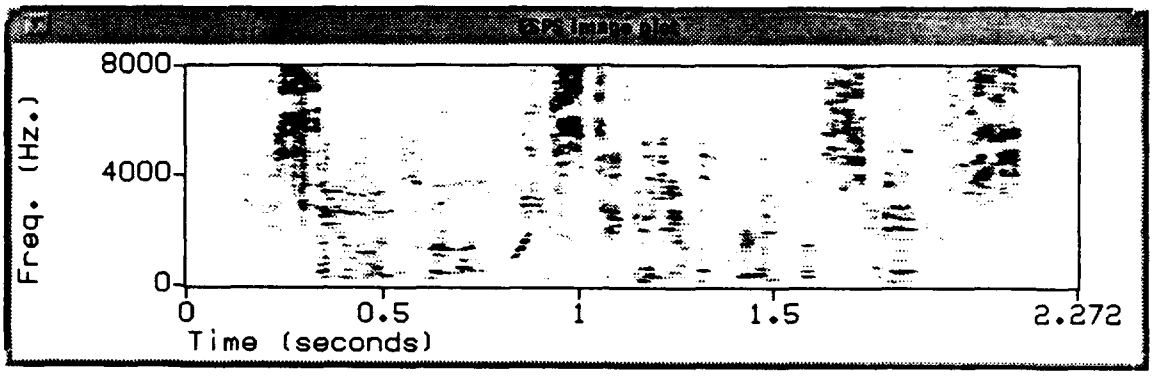


Figure B.10. Male, *A Priori* LPC Method, +5 dB, t211p5

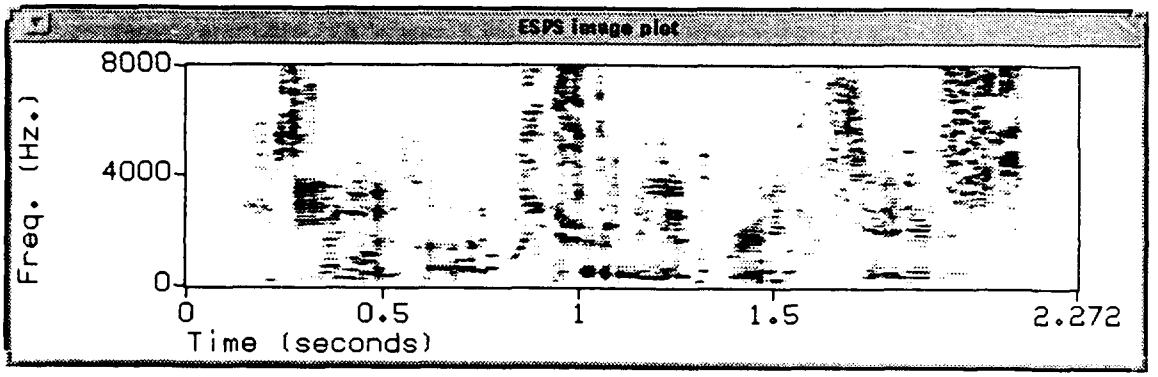


Figure B.11. Male, *A Priori* LPC Method, 0 dB, t211z

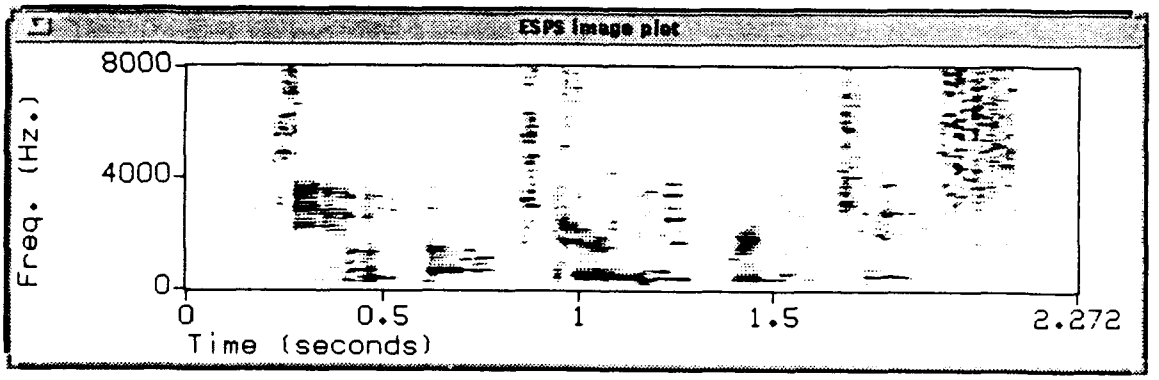


Figure B.12. Male, *A Priori* LPC Method, -5 dB, t211m5

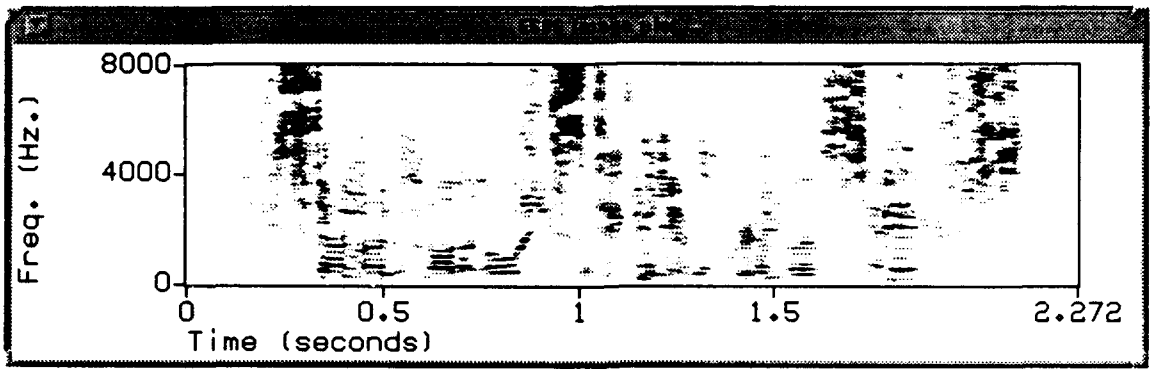


Figure B.13. Male, "Excluded Sentences", LPC Method, +5 dB, t217p5

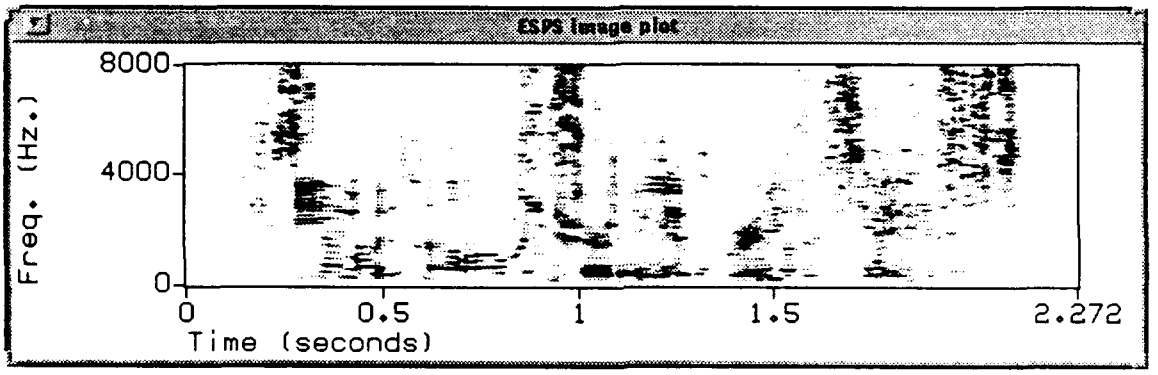


Figure B.14. Male, "Excluded Sentences", LPC Method, 0 dB, t217z

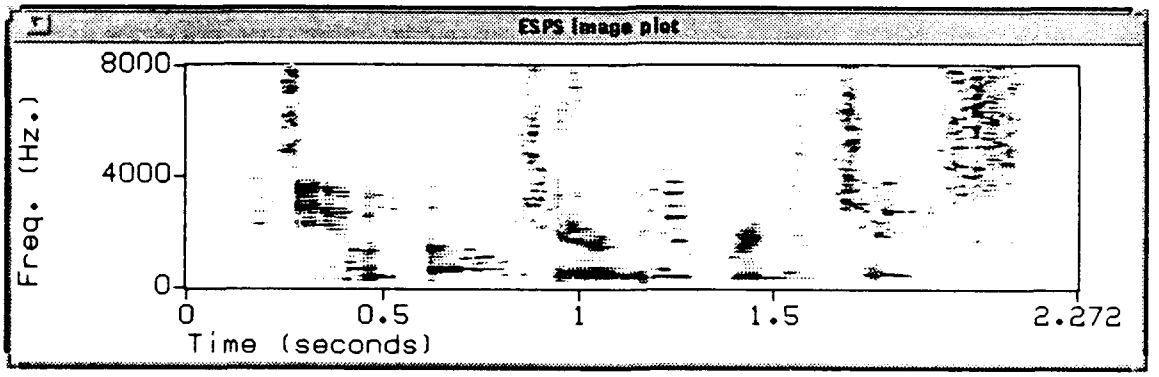


Figure B.15. Male, "Excluded Sentences", LPC Method, -5 dB, t217m5

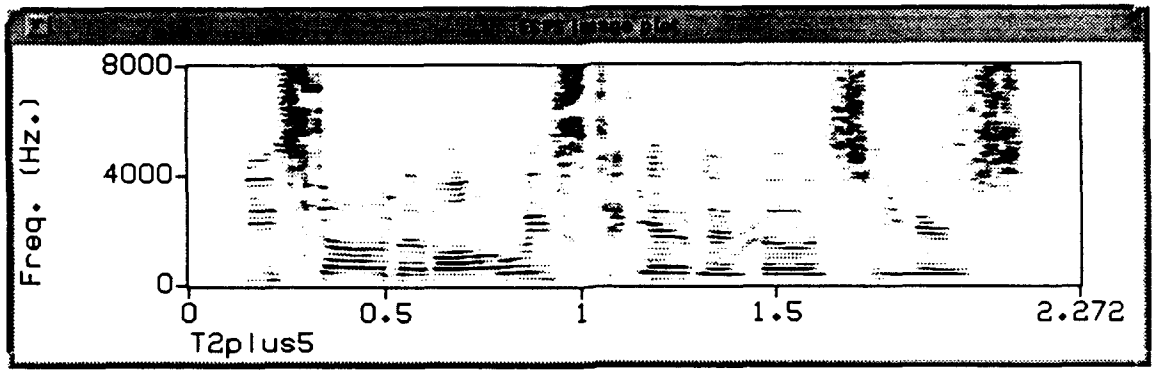


Figure B.16. Male, "Pitch Deviation" Method, +5 dB, t2plus5

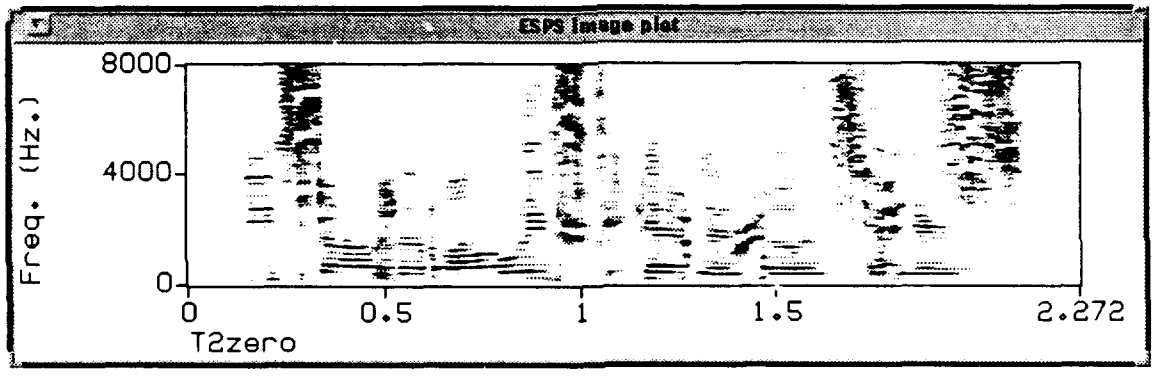


Figure B.17. Male, "Pitch Deviation" Method, 0 dB, t2pz

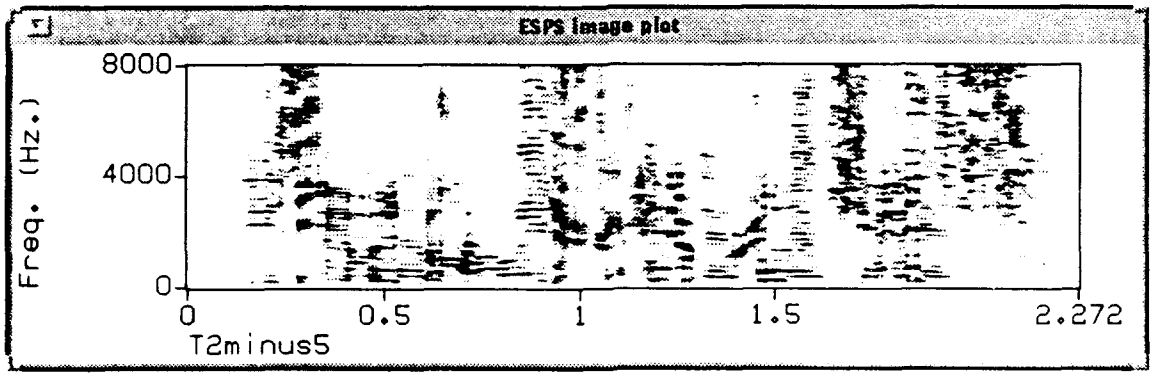


Figure B.18. Male, "Pitch Deviation" Method, -5 dB, t2minus5

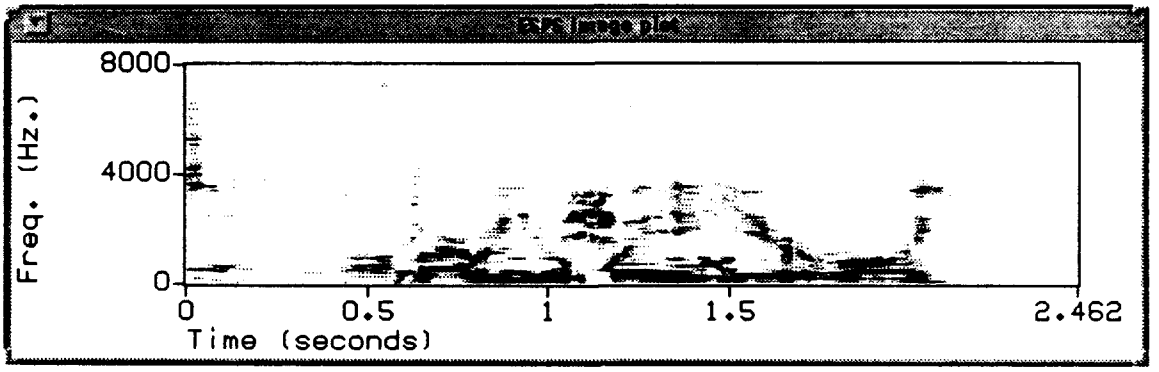


Figure B.19. Talker 1, *A Priori* LPC Method, +5 dB, t1tc11p5

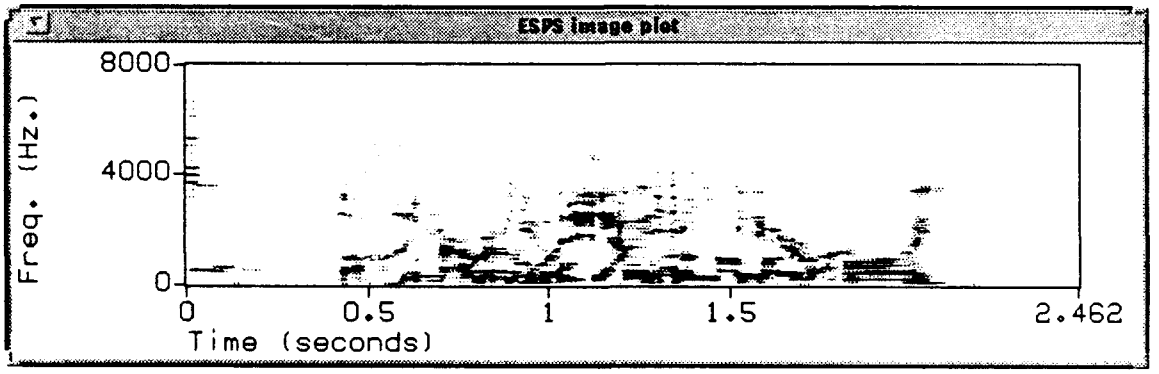


Figure B.20. Talker 1, *A Priori* LPC Method, 0 dB, t1tc11z

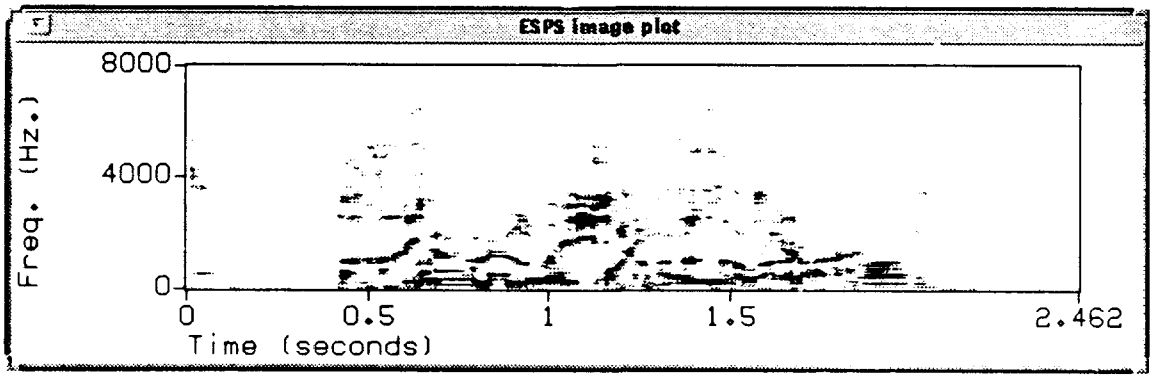


Figure B.21. Talker 1, *A Priori* LPC Method, -5 dB, t1tc11m5

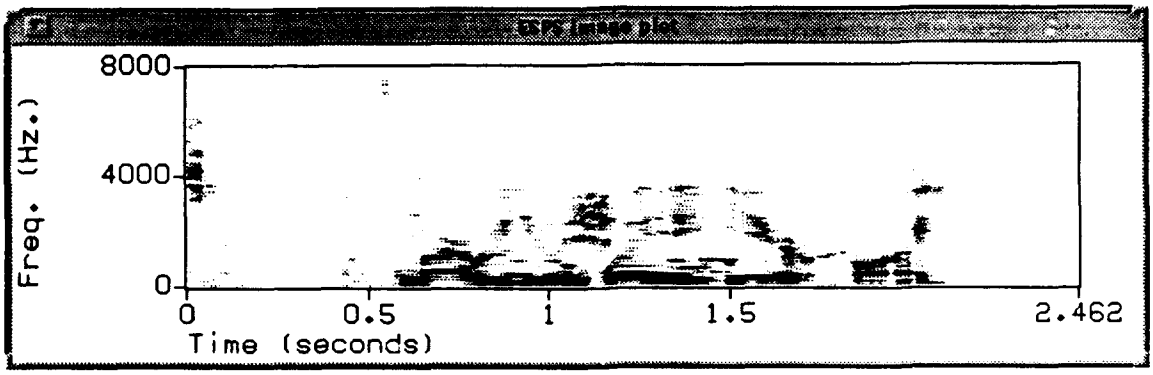


Figure B.22. Talker 1, "Excluded Sentences", LPC Method, +5 dB, t1tcl7p5

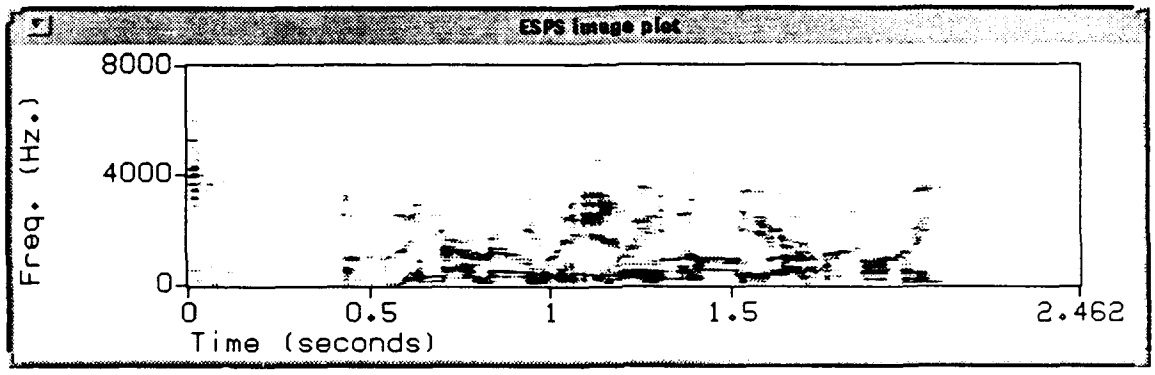


Figure B.23. Talker 1, "Excluded Sentences", LPC Method, 0 dB, t1tcl7z

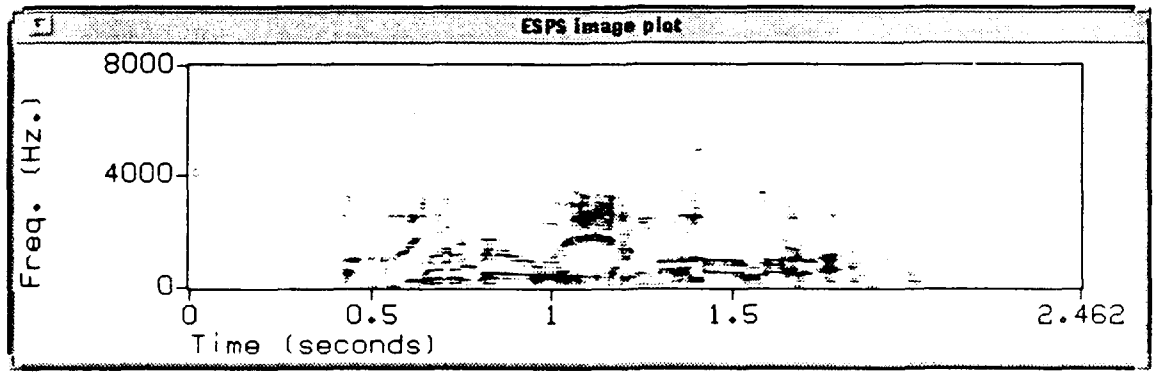


Figure B.24. Talker 1, "Excluded Sentences", LPC Method, -5 dB, t1tcl7m5

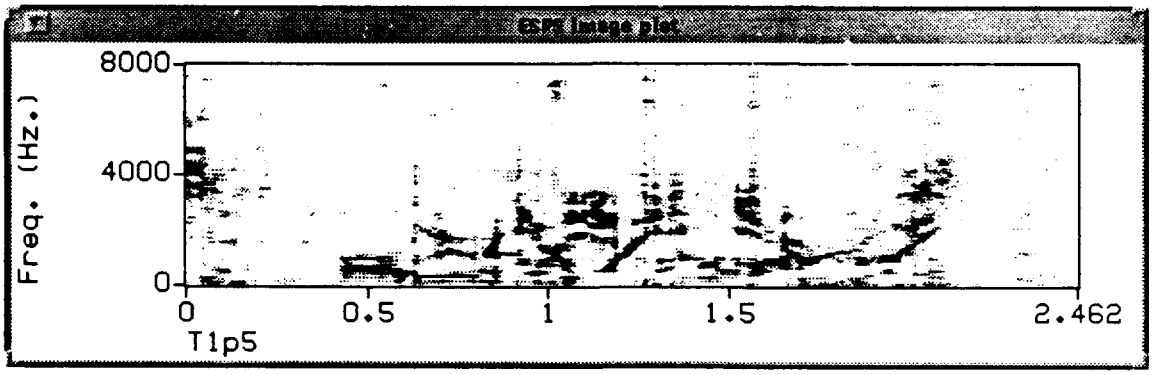


Figure B.25. Talker 1 "Pitch Deviation" Method, +5 dB, t1tcp5

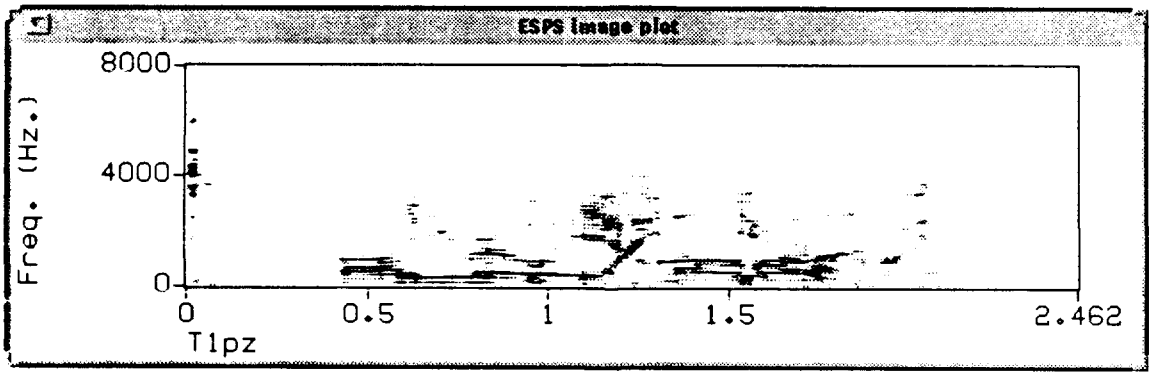


Figure B.26. Talker 1, "Pitch Deviation" Method, 0 dB, t1tcpz

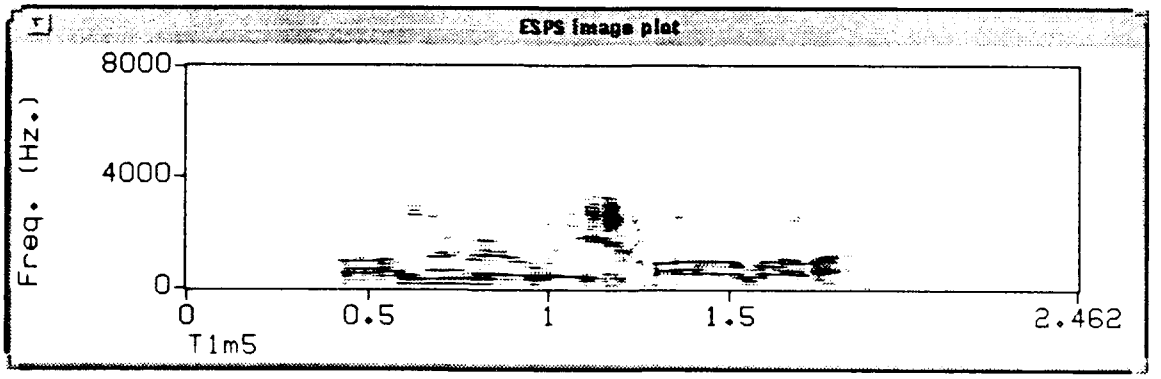


Figure B.27. Talker 1, "Pitch Deviation" Method, -5 dB, t1tcm5

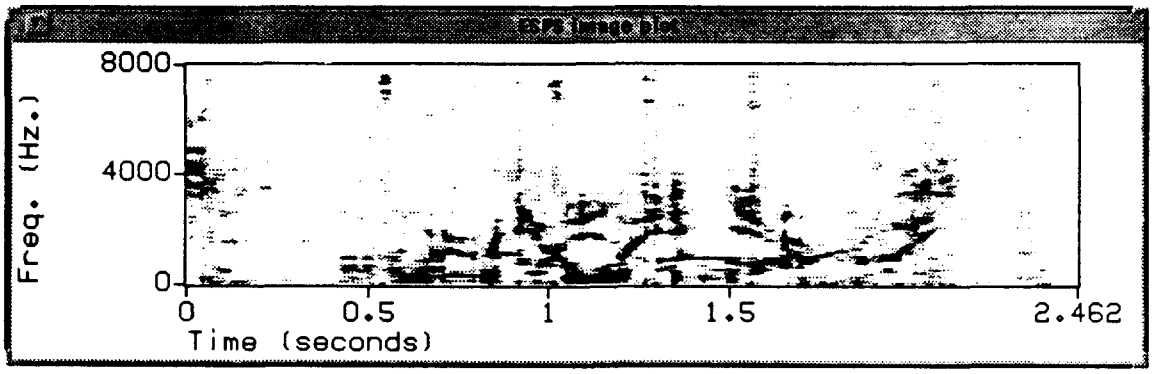


Figure B.28. Talker 2, *A Priori* LPC Method, +5 dB, t2tcl1p5

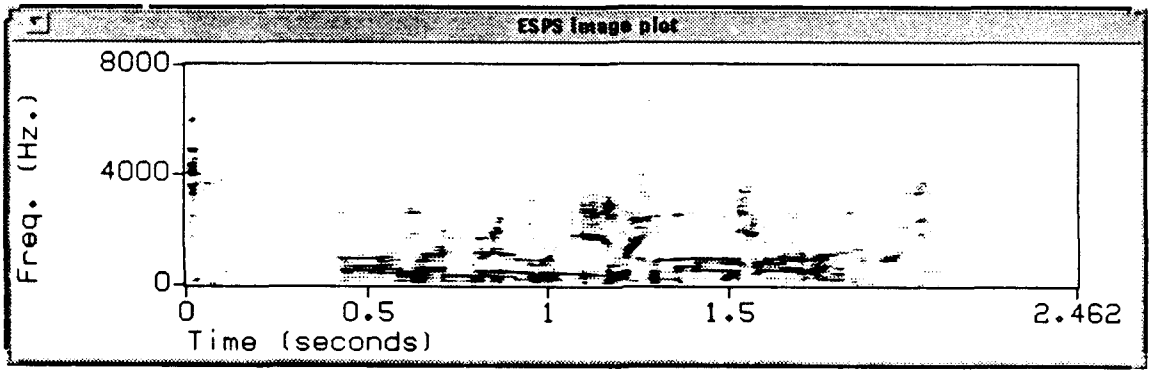


Figure B.29. Talker 2, *A Priori* LPC Method, 0 dB, t2tcl1z

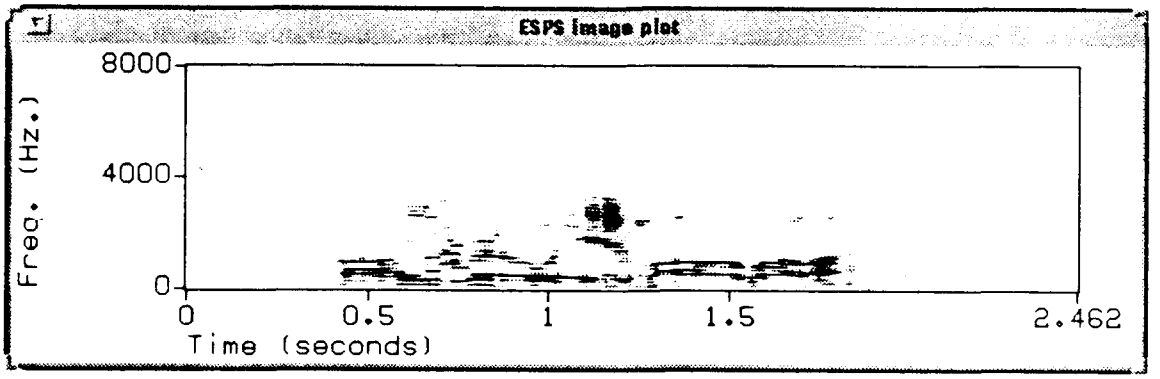


Figure B.30. Talker 2, *A Priori* LPC Method, -5 dB, t2tcl1m5

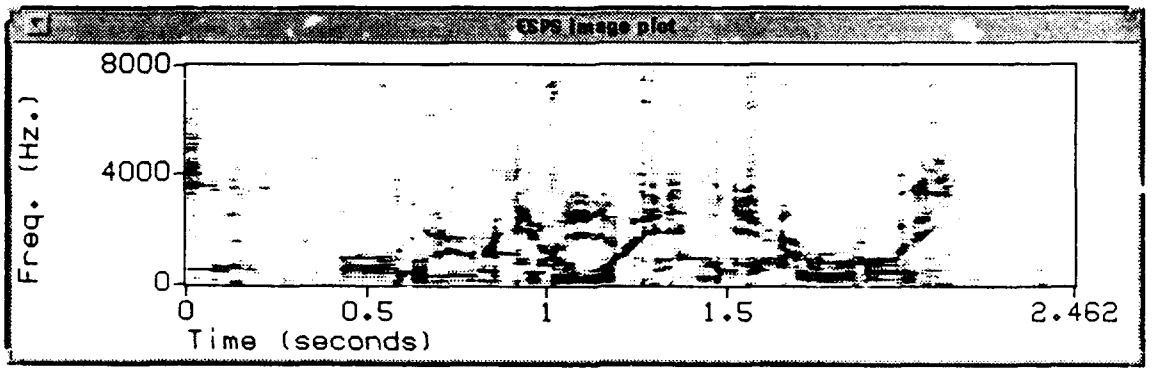


Figure B.31. Talker 2, "Excluded Sentences", LPC Method, +5 dB, t2tcl7p5

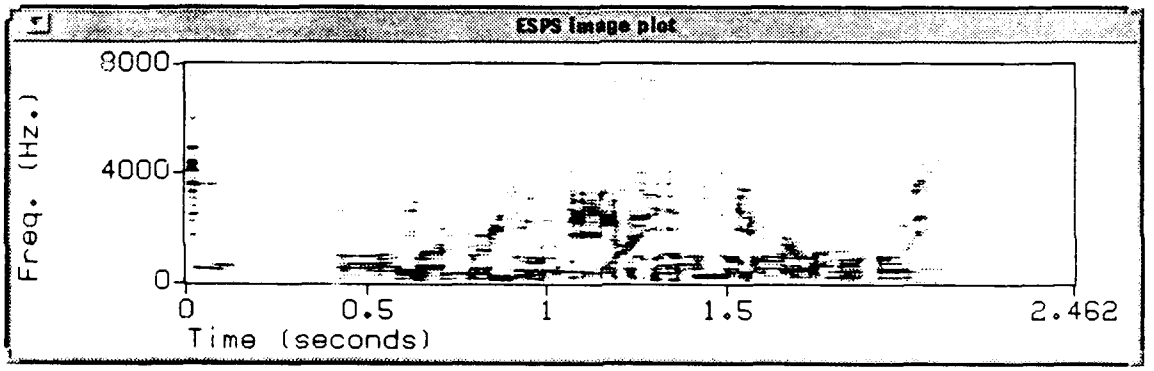


Figure B.32. Talker 2, "Excluded Sentences", LPC Method, 0 dB, t2tcl7z

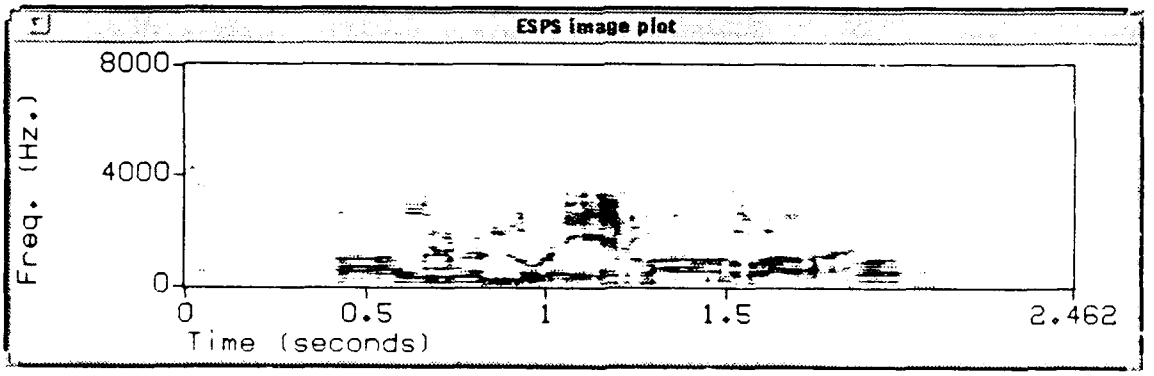


Figure B.33. Talker 2, "Excluded Sentences", LPC Method, -5 dB, t2tcl7m5

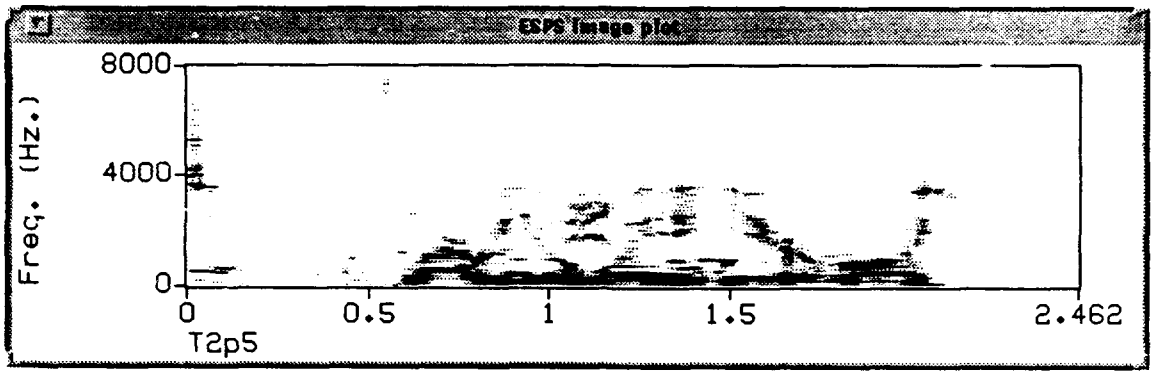


Figure B.34. Talker 2, "Pitch Deviation" Method, +5 dB, t2tcp5

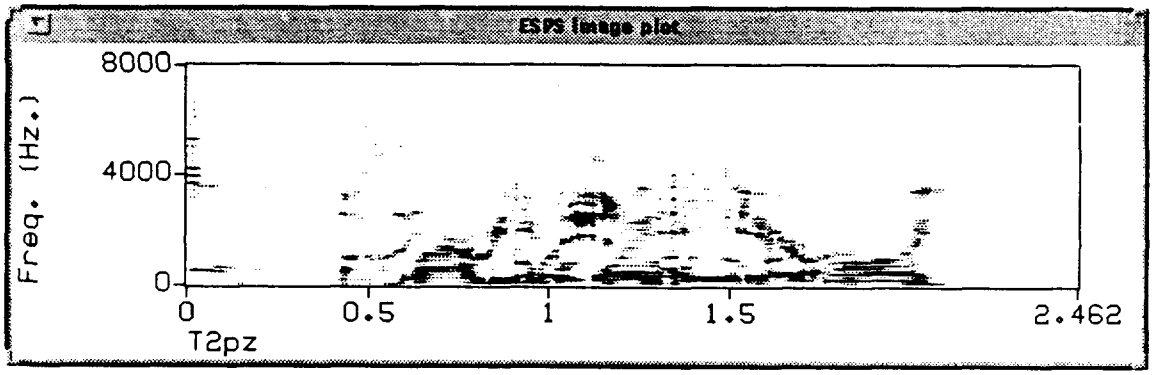


Figure B.35. Talker 2, "Pitch Deviation" Method, 0 dB, t2tcpz

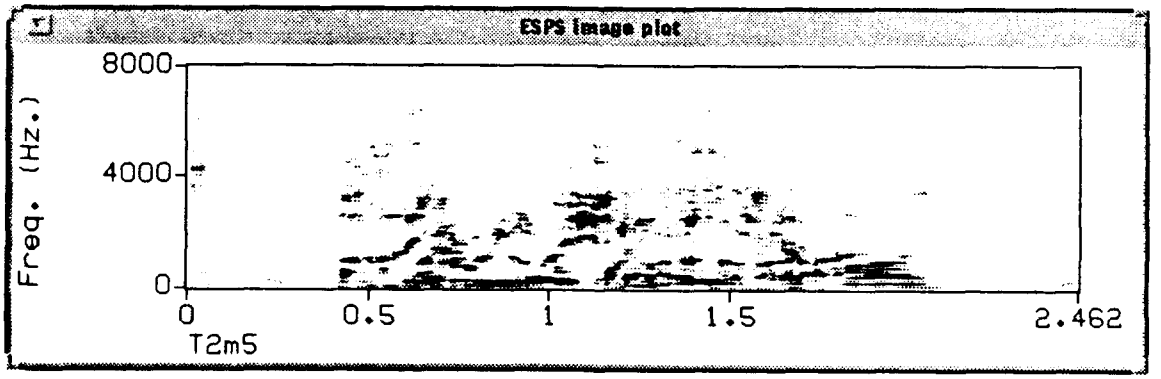


Figure B.36. Talker 2, "Pitch Deviation" Method, -5 dB, t2tcm5

Bibliography

1. C. et al, Rogers. "Neural Network Enhancement for a two Speaker Separation System," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 357–360 (1988). International Neural Network Society 1988.
2. DARPA. *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT)*, October 1990 1990.
3. Dick, Robert J. *Co-Channel Interference Separation*. Technical Report RADC-TR-80-365, RADC, Griffis AFB, NY 13441: Pattern Analysis and Recognition Corporation, December 1980 (ADA096059).
4. Frazier, R. et al. "Enhancement of Speech by Adaptive Filtering," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 418–423 (1976).
5. Hanson, Brian A. and David Y. Wong. *Processing Techniques for Intelligibility Improvement to Speech with Co-Channel Interference*. Contract Report, Rome Air Development Center, 1983 (ADA135702).
6. Hess, Wolfgang. *Pitch Determination of Speech Signals*. Springer-Verlag, 1983.
7. Itakura, Fumitada. "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Transactions on Acoustics, Speech, and signal Processing, ASSP-23*(1):67–72 (February 1975).
8. Kay, Steven M. *Modern Spectral Estimation, Theory and Application*. Prentice-Hall, Inc., 1987.
9. Kopec, Gary E. and Marcia A. Bush. "An LPC Based Similarity Measure for Speech Recognition in the Presence of Co-Channel Speech Interference," *IEEE International Conference on Acoustics, Speech and Signal Processing, 1*:270–273 (1989).
10. Lee, C. K. and D. G. Childers. "Co-channel Speech Separation," *Journal of the Acoustical Society of America, 83(1)*(1):274–280 (January 1988).
11. Lee, Leonard and David Morgan. *An Algorithm for Co-Channel Speaker Separation with Applications for Speech Enhancement and Suppression*. Technical Report, Nashua NH: Lockheed Sanders, May 1991.
12. Makhoul, John. "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, 561–580 (1975).
13. Marple Jr., Stanley Lawrence. *Digital Spectral Analysis with Applications*. Prentice-Hall, Inc., 1987.
14. McAulay, R. J. *Optimum Classification of Voiced Speech, Unvoiced Speech and Silence in the Presence of Noise and Interference*. Contract Report, Lexington MA: Massachusetts Institute of Technology Lincoln Laboratory, June 1976 (ADA028518).

15. McAulay, R. J. *Optimum Speech Classification and its Application to Adaptive Noise Cancellation*. Contract Report, Lexington MA: Massachusetts Institute of Technology Lincoln Laboratory, November 1976 (ADA036324).

16. Min, K. et al. "Automated two Speaker Separation System," *IEEE International Conference on Acoustics, Speech and Signal Processing, 1* (1988).

17. Naylor, J. and J. Porter. "An Effective Speech Separation System which Requires no *A Priori* Information," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:937–940 (May 1991).

18. Naylor, Joe A. *Interference Reduction Model*. Contract Report, Griffiss AFB NY: Rome Air Development Center, October 1987 (ADB118985).

19. Oppenheim, Alan V. and Ronald W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall, 1989.

20. Parsons, Thomas W. "Separation of Speech from Interfering Speech by Means of Harmonic Selection," *Journal of the Acoustical Society of America, 60*(4):911–918 (October 1976).

21. Parsons, Thomas W. *Voice and Speech Processing*. McGraw-Hill, Inc., 1987.

22. Paul, B. *Homomorphic Pitch Detection*. Contract Report, Massachusetts Institute of Technology Lincoln Laboratory, 1978 (unk).

23. Quatieri, Thomas F. and Ronald G. Daniesewicz. "An Approach to Co-Channel Talker Interference Suppression Using a Sinusoidal Model for Speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 56–69 (January 1990).

24. Rabiner, L. R. and M. R. Sambur. "An Algorithm for Determining the Endpoints of Isolated Utterances," *The Bell System Technical Journal, 54*(2):297–315 (February 1975).

25. Rabiner, L. R. and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

26. Smyth, Christopher C. *Computer Programs for Measuring Variations in Speech*. Technical Report, Aberdeen Proving Ground, MD: US Army Human Engineering Laboratory, July 1989 (ADA211126).

27. Stubbs, Richard J. and Quentin Summerfield. "Algorithms for Separating the Speech of Interfering Talkers: Evaluations with Voiced Sentences, and Normal-Hearing and Hearing-Impaired Listeners," *Journal of the Acoustical Society of America, 87*(1):359–372 (January 1990).

28. Wise, James D. et al. "Maximum Likelihood Pitch Estimation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 418–423 (October 1976).

29. Zissman, M. A. et al. "Speech-State-Adaptive Simulation of Co-Channel Talker Interference Suppression," *IEEE International Conference on Acoustics, Speech and Signal Processing, 1*:361–364 (May 1989).

Vita

Captain Thomas S. Andrews was born on November 17, 1961. Capt Andrews graduated from John Glenn High School in Westland Michigan, in June 1980. He enlisted in the Air Force in June 1981, and was assigned to the Tactical Air Warfare Center, Eglin Air Force Base, Florida, where he served for three and a half years as an administrative specialist. He was selected for the Airman's Education and Commissioning Program in 1984, and he transferred to the University of South Florida to pursue a Bachelor of Science degree in electrical engineering. Capt Andrews graduated from South Florida in July 1987 with a BSEE, and he was commissioned as a second lieutenant, from the Officer's Training School in October 1987. Capt Andrews served as a foreign communications systems analyst at the Foreign Technology Division, Wright-Patterson Air Force Base, Ohio, from November 1987 to May 1991. In June 1991, he entered the School of Engineering, Air Force Institute of Technology at Wright-Patterson Air Force Base, Ohio, to pursue a Master of Science in Electrical Engineering degree with an an emphasis in communications and radar signal technology. He is married to Annette (Magallanes) Andrews of South El Monte, California and they have one child, Joshua.

Permanent address: 6409 Hemingway Road
Huber Heights, Ohio 45424

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1992	Master's Thesis

4. TITLE AND SUBTITLE
CO-CHANNEL SPEAKER SEPARATION

5. FUNDING NUMBERS

6. AUTHOR(S)
Thomas S. Andrews

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology, WPAFB OH 45433-6583

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/GE/ENG/92S-07

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Capt Rick Ricart
RL/IRAA Griffiss AFB NY 13449

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Abstract

In the co-channel speaker separation problem, the goal is to recover two separate speech signals from a monaural channel which contains the sum of the two speech signals. A new methodology is developed that if given that a segment of co-channel speech is separated into a "stronger" and "weaker" segment, the correct assignment of these separated segments to the appropriate talker can be made using a Linear Predictive Coding (LPC) based minimum-prediction residual computation. The uniqueness of the developed technique is that no *a priori* information is required of the co-channel speech signal. The information needed to appropriately assign these separated segments from the co-channel speech signal are "clean" speech that is separate from the co-channel speech signal that are used to compute model LPC vectors. This "clean" speech is derived from the same channel that the co-channel speech signal is derived from. This technique has shown the ability to correctly assign the given "stronger" and "weaker" segments to the appropriate talker at signal-to-signal ratios down to equal power levels. The resulting separated speech is clearly understandable, and the interfering talker's speech signal is effectively eliminated.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Co-Channel, Speaker Separation, Speech Processing, LPC, Itakura Minimum Prediction Residual	100
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL