

AN APPROACH  
FOR THE SEPARATION OF VOICES IN  
COMPOSITE MUSICAL SIGNALS

BY

ROBERT CRAWFORD MAHER

B.S., Washington University, 1984  
M.S., University of Wisconsin--Madison, 1985

THESIS

Submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy in Electrical Engineering  
in the Graduate College of the  
University of Illinois at Urbana-Champaign, 1989

Urbana, Illinois

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN  
GRADUATE COLLEGE DEPARTMENTAL FORMAT APPROVAL

THIS IS TO CERTIFY THAT THE FORMAT AND QUALITY OF PRESENTATION OF THE THESIS  
SUBMITTED BY Robert Crawford Maher AS ONE OF THE  
REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy  
IS ACCEPTABLE TO THE Department of Electrical and Computer Engineering  
*Full Name of Department, Division or Unit*

April 24, 1989  
*Date of Approval*

[Signature]  
*Departmental Representative*

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

THE GRADUATE COLLEGE

APRIL 1989

WE HEREBY RECOMMEND THAT THE THESIS BY

ROBERT CRAWFORD MAHER

ENTITLED AN APPROACH FOR THE SEPARATION OF

VOICES IN COMPOSITE MUSICAL SIGNALS

BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

*Janet Beaudry*

Director of Thesis Research

*Timothy N. Ford*

Head of Department

Committee on Final Examination†

*Janet Beaudry*

Chairperson

*Dennis Cooper*

*Leon A. Triggall*

*Scott A. Wyatt*

† Required for doctor's degree but not for master's.

AN APPROACH  
FOR THE SEPARATION OF VOICES  
IN COMPOSITE MUSICAL SIGNALS

Robert Crawford Maher, Ph.D.  
Department of Electrical and Computer Engineering  
University of Illinois at Urbana-Champaign, 1989  
James Beauchamp, Advisor

The simultaneous presentation of several sound sources in a performance setting is fundamental to most music. Ensemble musical signals consist of superpositions of multiple distinct sonic events which may or may not be synchronized in time, frequency, and/or amplitude. Once the distinct events are combined and recorded in a storage medium, e.g., a digital recording, the composite signal is often unsatisfactory in some way: the recording might suffer from poor ensemble balance, performance errors, or corruption from undesired background audience noises (sneezing, talking, etc.). Although it often might be helpful to process the constituent signals independently, separating the composite signal into its parts is a nontrivial task. The research reported here considers particular aspects of the separation problem: analysis, identification, tracking, and resynthesis of a specified voice from a digital recording of a musical duet setting. Analysis is accomplished out of real-time using a quasi-harmonic, sinusoidal representation of the constituent signals, based on short-time Fourier transform (STFT) methods. The procedure is evaluated via resynthesis of a "desired" signal from the composite analysis and tracking data. Other applications include signal restoration, digital editing and splicing, musique concrete, noise reduction, and time-scale compression/expansion.

This material is based upon work supported, in part, under a National Science Foundation Graduate Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

## DEDICATION

This dissertation is dedicated to my parents, Louis and Jane Maher, and to the public school teachers of the Middleton-Cross Plains School District, Dane County, Wisconsin.

## ACKNOWLEDGEMENTS

I wish to express sincere appreciation to Dr. James Beauchamp, my doctoral advisor, who provided the facilities, instruction, direction, wisdom, humor, and friendship so vital to a successful dissertation project.

George Gaspar, engineer for the Computer Music Project, maintained the critical digital and analog hardware used for this research, and Electrical and Computer Engineering doctoral student Kurt Hebel of the CERL Music Group provided vital support for the Sound Conversion and Storage System (SCSS), which he designed, built, and maintained. George and Kurt repeatedly have responded far above and beyond the call of duty in answering my questions and fixing malfunctions. I would also like to acknowledge Music Department faculty members Scott Wyatt and John Melby for their concern, comments, and instruction.

The work for this dissertation was supported primarily by a National Science Foundation Graduate Fellowship, and grants from the University of Illinois Research Board, the Audio Engineering Society Education Foundation, and IBM Corporation. The generous support of these organizations allowed me the freedom to pursue my work without the worries and impediments that often accompany graduate studies.

A simple acknowledgement does not do justice to the role of my wife, Lynn, in maintaining a "real" life outside the laboratory. Her unfaltering love helped in uncountable ways during the work on this dissertation, and I owe a special, personal debt to her that I may never be able to repay.

## PREFACE

This dissertation involves contributions from the fields of electrical engineering, computer science, physics, music, speech and hearing science, and psychology, so some conflicts in terminology are inevitable. Thus, a few basic terms and their usage are defined at the outset.

- >> A voice indicates a single musical signal or musical line, such as the melodic notes played by a flute or sung by a soprano. For example, a solo has one voice, a duet has two voices, a trio has three, etc.
  
- >> A harmonic signal is a signal representable by a time-variant Fourier series. In other words, a harmonic signal can be expressed as some number of sinusoidal components whose frequencies are all integer multiples of a base frequency, called the fundamental frequency or simply the fundamental. The sinusoidal components are often referred to as harmonics. Some inharmonic or quasi-harmonic signals may be representable as a sum-of-sinusoids, and the general terms partials or overtones are used to describe the discrete spectral components in this case. Although musical sounds are not usually harmonic in a strict sense, the signals of interest in this investigation are at least nearly harmonic so that a meaningful fundamental frequency can be specified at all times.
  
- >> The term timbre has been defined by the American National Standards Institute as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar." This definition is a statement



of perceptual judgment concerning the tone quality or tone "color" of a sound. Timbre is at least related to the frequency spectrum of the stimulus. Many attempts have been made to relate other measurable physical parameters to the multidimensional realm of timbre perception, but without total success. The term timbre will be used in this dissertation to describe the general time-variant spectral properties of a musical signal.

>> Although the words fundamental frequency and pitch are often considered synonyms in the vernacular, the psychoacoustic definitions differ: the fundamental frequency is a physical quantity measured in terms of waveform repetitions per second (Hertz), while pitch is a perceptual phenomenon defined in terms of an empirically derived unit of measure (such as the mel). The pitch of an acoustic stimulus is often directly related to its frequency, but the pitch typically varies with other parameters of the stimulus, such as its amplitude level, spectrum, or duration. With this difference in mind, the use of the word pitch will be eschewed in favor of frequency whenever conflicting usage might cause confusion.

## TABLE OF CONTENTS

CHAPTER	PAGE
1 INTRODUCTION .....	1
1.1 Description of Problem .....	2
1.2 Motivation and Applications .....	3
1.3 Statement of Research Goals .....	4
1.4 Research Facility Overview .....	5
1.5 Outline of Dissertation .....	6
2 REVIEW OF RELEVANT LITERATURE .....	7
2.1 Short-time Fourier Transform Methods .....	7
2.2 Time-variant Spectral Analysis of Musical Sounds .....	9
2.3 Co-channel Speech Separation .....	10
2.4 Segmentation and Analysis of Musical Signals .....	14
3 RESEARCH APPROACH AND METHODS .....	19
3.1 Short-time Fourier Transform Analysis .....	19
3.2 Sampling the STFT: $X(n,k)$ .....	26
3.3 Short-time Fourier Transform Synthesis .....	28
3.4 STFT Synthesis with Modifications .....	34
3.5 A Variation of the STFT Concept: The MQ Analysis/Synthesis Procedure .....	34
3.6 Tracking the Fundamental Frequencies in a Duet Signal .....	46
3.6.1 Common methods for pitch detection .....	46
3.6.2 A new duet fundamental frequency tracking approach .....	49
3.7 The Use of the MQ Procedure in the Duet Separation Task .....	55
3.7.1 MQ frequency resolution and spectral collisions .....	56

3.7.2	Separation strategy I: linear equations .....	59
3.7.3	Separation strategy II: analysis of beating components ...	62
3.7.4	Separation strategy III: signal models, interpolation, or templates .....	68
3.7.5	Further considerations .....	73
4	RESULTS AND DISCUSSION .....	77
4.1	Testing and Evaluation Outline .....	78
4.2	Evaluation of Duet Fundamental Frequency Tracking .....	85
4.3	Evaluation of Voice Separation with Known Fundamental Frequencies .....	97
4.4	Evaluation of Voice Separation with Frequency Tracking .....	107
5	CONCLUSIONS .....	118
5.1	Summary of Findings .....	118
5.2	Future Directions .....	120
	APPENDIX A IMPLEMENTATION NOTES .....	122
	APPENDIX B DESCRIPTION OF SOFTWARE MODULES .....	125
	LIST OF REFERENCES .....	129
	VITA .....	135

## LIST OF FIGURES

Figure 3.1a:	Filter-bank Analysis Viewpoint of the STFT. ....	22
Figure 3.1b:	Fourier Transform Viewpoint of the STFT. ....	24
Figure 3.2:	Sum of Overlapped Windows for Various Spacings. ....	33
Figure 3.3:	The Basic MQ Analysis Procedure. ....	37
Figure 3.4:	Parabolic Interpolation of Spectral Peak Location. ....	39
Figure 3.5:	MQ Analysis Representation of a Time-varying Signal. ....	42
Figure 3.6:	The Two-way Mismatch Error Calculation. ....	51
Figure 3.7:	Overlap Response for Two Closely Spaced Partials (Real Part). ....	62
Figure 3.8:	Examples of $A_1 \cos(\omega_1 t) + A_2 \cos(\omega_2 t)$ ....	64
Figure 3.9:	Instantaneous Frequency Functions for Waveforms from Figure 3.8. ....	66
Figure 3.10:	Linear Interpolation of Spectrum to Resolve a Collision. ....	70
Figure 3.11:	Cubic Interpolation of Spectrum to Resolve a Collision. ....	71
Figure 3.12:	The Use of Spectral Templates to Resolve a Collision. ....	74
Figure 4.1:	Artificial Duet Test Signal #1: Synthesis Frequencies. .....	79
Figure 4.2:	Artificial Duet Test Signal #2: Synthesis Frequencies. .....	81

Figure 4.3:	Artificial Duet Test Signal #3: (a) Musical score (b) Frequency Specification .....	82
Figure 4.4:	Artificial Duet Test Signal #4: Synthesis Frequencies. .....	83
Figure 4.5:	Duet Test Signal #5: (a) Musical Score (b) Frequency Specification .....	84
Figure 4.6:	Duet Test #6. ....	86
Figure 4.7:	Duet Used for Tests #7 and #8. ....	87
Figure 4.8:	(a) MQ Analysis and (b) Two-way Mismatch (TWM) Frequency Tracking of Artificial Duet Test Example #1. ....	89
Figure 4.9:	(a) MQ Analysis and (b) Two-way Mismatch (TWM) Frequency Tracking of Artificial Duet Test Example #2. ....	90
Figure 4.10:	(a) MQ Analysis and (b) Two-way Mismatch (TWM) Frequency Tracking of Artificial Duet Test Example #3. ....	91
Figure 4.11:	(a) MQ Analysis and (b) Two-way Mismatch (TWM) Frequency Tracking of Artificial Duet Test Example #4. ....	94
Figure 4.12:	Frequency Tracking Data for the INDIVIDUAL Soprano Voices Used in Example #5. (a) Arpeggio With Vibrato (b) Constant Pitch Without Vibrato .....	95
Figure 4.13:	(a) MQ Analysis (excerpt) and (b) Two-way Mismatch (TWM) Frequency Tracking of Duet Test Example #5 (Note tracking problem during $2 < t < 6$ ). ....	96
Figure 4.14:	Separation Spectra of Example #1 Using a priori Frequency Data. (a) voice 1 (b) voice 2 .....	98
Figure 4.15:	Separation Spectra of Example #2 Using a priori Frequency Data. (a) voice 1 (b) voice 2 .....	101

Figure 4.16:	Collision Repair of Example #2.	
	(a) upper voice (b) lower voice .....	102
Figure 4.17:	Separation Spectra of Example #3 Using a priori	
	Frequency Data. (a) voice 1 (b) voice 2 .....	104
Figure 4.18:	Separation Spectra of Example #4 Using a priori	
	Frequency Data. (a) voice 1 (b) voice 2 .....	105
Figure 4.19:	Separation Spectrum of Example #5 Using a priori	
	Frequency Data (Excerpt from Upper Voice). .....	106
Figure 4.20:	Results for Example #1 Using the Complete	
	Separation Process. (a) voice 1 (b) voice 2 .....	108
Figure 4.21:	Results for Example #2 Using the Complete	
	Separation Process. (a) voice 1 (b) voice 2 .....	109
Figure 4.22:	Results for Example #3 Using the Complete	
	Separation Process. (a) voice 1 (b) voice 2 .....	110
Figure 4.23:	Results for Example #4 Using the Complete	
	Separation Process. (a) voice 1 (b) voice 2 .....	111
Figure 4.24:	Results for Example #5 Using the Complete	
	Separation Process (Excerpt from Upper Voice). .....	112
Figure 4.25:	Results for Example #6 Using the Complete	
	Separation Process. (a) voice 1 (b) voice 2 .....	114
Figure 4.26:	Results for Example #7 Using the Complete	
	Separation Process	
	(Fundamental Frequencies Manually Edited).	
	(a) voice 1 (b) voice 2 .....	116

Figure 4.27: Results for Example #8 Using the Complete Separation Process.

(a) voice 1 (b) voice 2 ..... 117

Figure (B.1): Module Flow Diagram. .... 125

## CHAPTER 1

## INTRODUCTION

Separation of signals superimposed in time is a problem of interest to several branches of electrical engineering. The problem often appears when a desired signal is corrupted by some undesired interference, as in the study of radar and sonar data, the removal of reverberation in recorded speech or music, the separation of simultaneous talkers in a communications channel, the suppression of additive noise, etc. From a practical standpoint, the separation task relies on prior knowledge of some aspect of the superimposed signals whereby a set of separation criteria may be identified. For example, if the superimposed signals occupy nonoverlapping frequency bands, the separation problem reduces to the specification and design of frequency selective filters. In other cases the competing signals may be described in a statistical sense, allowing separation via correlation or some nonlinear detection scheme. Unfortunately, many superimposed signals do not allow such simple decomposition methods, so other strategies applicable to signal separation must be discovered.

In the case of ensemble music, several signals generated by possibly different musical instruments are combined in an acoustic signal which may be captured on a recording medium via a transducer of some kind. The individual musical signals of the ensemble may or may not be synchronized in time, frequency, or amplitude, and a simple statistical description of each signal is not available for the separation task. Despite this complexity, a human listener can usually identify the instruments playing at a given point in time. Further, a listener with some training can often reliably transcribe



each voice in terms of standard musical pitch and rhythm, i.e., musical notation. Unfortunately, the methods and strategies employed by human observers are not introspectable, and thus cannot serve easily as models for emulation in the form of an automatic system. For this and related reasons, the automatic ensemble separation problem remains unsolved and intriguing.

### 1.1 Description of Problem

Considered in this dissertation is the problem of automatic decomposition of a musical recording into its constituent signal components (for example, the extraction of the "trumpet" signal from a recording of a trumpet and tuba duet). As noted above, the human model provides an operational example of many features of such a system, but we currently lack the knowledge of its methods and mechanisms. Instead, a digital signal processing approach has been used for this dissertation based on physical measurements rather than psychoacoustic models.

For the purposes of this investigation, several restrictions on the properties of the input signal have been employed to limit the complexity of the decomposition procedure. First, recordings containing only two separate voices (musical duets) are considered. Second, each voice of the duet is required to be nearly harmonic and to contain a sufficient number of partials (overtones) for unambiguous estimation of the voice's fundamental frequency. Third, the fundamental frequencies of the two voices are restricted to non-overlapping ranges, i.e., the lowest note of the upper voice must be higher than the highest note of the lower voice. Finally, reverberation, echos, and other correlated noise sources are discouraged since, in effect, they

represent additional "background voices" in the recording and violate the duet assumption.

Despite these restrictions, the remaining problems are formidable: the partials of one voice will often collide with the partials of the other voice; the duet voices may occur simultaneously or separately (and neither voice may occur during shared rests); level imbalances between voices or added noise may hinder the detection process, and so forth. Development of methods to alleviate these difficulties has required the most time and effort during the work for this dissertation.

The analysis/extraction/synthesis system has been implemented in the C programming language on a general-purpose digital computer. The processing time required for this implementation may be hundreds of times real-time (1 second of sound may require several minutes of computer processing), but for development purposes this disadvantage is more than balanced by the flexibility and testability available with a software simulation approach.

## 1.2 Motivation and Applications

Recordings of music may suffer from many degradations: performance errors during a live recording, imbalance between instrumental lines, complex post-recording signal processing requirements, etc. In many cases these problems might be reduced or eliminated if the basis signals comprising the recorded material could be extracted and processed separately. This research was intended to develop and evaluate a method to accomplish this task.

Additional applications of this work may be found in background noise reduction, where musical material may be corrupted by sneezing or coughing from an audience, wind noise, or other competing signals. The process could also be useful as an aid in musical composition and sound production. With a fully operational system, the analysis could be used to alter stereo imaging and reverberation quality or to drive an automatic music transcription system.

### 1.3 Statement of Research Goals

The major goal of this dissertation work has been to demonstrate the feasibility of composite signal decomposition using a time-frequency analysis procedure. The research effort can be stated in terms of two fundamental questions:

- (1) How may we automatically obtain accurate estimates of the time-variant fundamental frequency of each voice from a digital recording of a duet?
- (2) Given time-variant fundamental frequency estimates of each voice in a duet, how may we identify and separate the interfering partials (overtones) of each voice?

Question (1) treats the problem of estimating the time-variant frequencies of the spectral components contributed by each voice. Assuming nearly-harmonic input, specification of a fundamental frequency identifies the partial component frequencies of that voice. Conflicting (coincident) partial frequencies may be identified and marked for further processing by comparing the harmonic series of each voice of the duet.

Question (2) involves the fundamental constraints on simultaneous time and frequency resolution. The desire for high-resolution frequency domain information necessitates observation of the input signal over a long time span. However, long observation spans may result in an unacceptable loss of time resolution. Thus, we must deal with inherent uncertainty in determining the time-vs.-frequency representation of the input duet signal.

Note that it is possible to deal with question (2) without solving question (1), assuming the time-variant fundamental frequency pair for the duet can be identified and tabulated by some manual means. Thus, the two fundamental questions can be treated independently, if desired.

#### 1.4 Research Facility Overview

The research effort has been implemented in software, using the facilities of the University of Illinois Computer Music Project. The CMP comprises a 700 square-foot music composition and acoustics research facility in the School of Music. The CMP contains 16-bit A/D and D/A systems supporting sample rates to 50 kHz per channel stereo, IBM RT PC, LMC, IBM PC AT, and Macintosh computers, modems for dialup connection to the University computing facilities, and extensive audio recording and playback equipment. The general-purpose nature of the audio and computing equipment has provided a comprehensive and productive environment for this research project, at the expense of computation rates much slower than desired for production level

signal processing work. Although a final implementation of this project using special-purpose signal processing hardware would be desirable, the available computing environment has been nearly ideal for the development and evaluation work described herein.

### 1.5 Outline of Dissertation

The narrative portion of this dissertation begins in Chapter 2 with a brief review and summary of the relevant literature. Covered in Chapter 3 are the fundamental theoretical tenets and procedures employed in this research. A detailed examination of the short-time Fourier transform methods used for this work is also included. Discussed in Chapter 4 are the significant results, successes, and failures, followed by concluding statements and a research summary in Chapter 5. Described in Appendix A are several practical issues related to the specific software implementation used for this study, and contained in Appendix B is a description of the software modules.

## CHAPTER 2

### REVIEW OF RELEVANT LITERATURE

A review of the reported work in composite signal separation reveals several approaches to this problem. Most previous work has considered the case in which the desired and interfering signals are human speech, although some work with musical signals has been reported.

Discussed in this chapter is some of the published work related to the topic of this dissertation. In the interest of narrative simplicity and brevity, the mathematical details of each topic will not be repeated here. Any specific concepts of direct importance to this dissertation will be treated in Chapter 3.

#### 2.1 Short-time Fourier Transform Methods

The fundamental analysis approach employed in this investigation has been the discrete short-time Fourier transform, or simply the discrete STFT [Allen, 1977; Allen and Rabiner, 1977]. STFT methods have been well reported over the last 15 years, primarily for analysis and synthesis in speech [Oppenheim and Schaffer, 1975; Portnoff, 1976, 1980 and 1981; Crochiere, 1980; Griffin and Lim, 1984 and 1988; McAulay and Quatieri, 1986; Quatieri and McAulay, 1986; Dembo and Malah, 1988], and in music [Risset and Mathews, 1969; Beauchamp, 1969 and 1975; Moorer, 1975 and 1978; Dolson, 1983 and 1985; Smith and Serra, 1987; Strawn, 1987; Maher and Beauchamp, 1988].

As will be described in Chapter 3, the STFT process can be formulated as an identity analysis/synthesis system; i.e., the unaltered analysis data may

be used to synthesize a perfect copy of the original input signal. Within some restrictions, useful modifications may be made to the analysis data prior to resynthesis. For example, we can achieve time-scale compression or expansion of the signal while maintaining its original pitch [Portnoff, 1981; Quatieri and McAulay, 1986]. For this investigation the identity property of the STFT has provided the means to operate in either the time or frequency domain as appropriate for a given processing situation.

It is often necessary to identify how the spectral content of a signal varies with time. In particular, for signals representable as a finite sum-of-sinusoids, the STFT can be used to estimate the amplitude, frequency, and phase of each sinusoid as a function of time. Note, however, that the discrete Fourier transform (DFT) provides a sampled version of the short-time spectrum at points equally spaced in frequency. Thus, if the fundamental frequency of the input signal is an integral divisor of the sample rate, the fundamental frequency and each of its harmonics will coincide with frequency samples provided by the DFT [Oppenheim and Schaffer, 1975]. In general, of course, it may be inconvenient or impossible to ensure any particular relationship between the input signal frequency and the sample rate. In this case, the input signal can be digitally interpolated to adjust its sample rate prior to the STFT (time domain interpolation) [Crochiere and Rabiner, 1983], or the short-time spectrum may be interpolated to estimate values between the available DFT frequency points (frequency domain interpolation). Either technique can be used to provide an improved estimate of the parameters of the signal's harmonic components.

## 2.2 Time-variant Spectral Analysis of Musical Sounds

Digital signal processing (DSP) techniques were used in the analysis of musical sounds beginning in the 1960s [cf., for example, Luce, 1963; Freedman, 1965, 1967, 1968; Beauchamp, 1965, 1966, 1969; Risset and Mathews, 1969]. Most of the early work was intended to resolve the harmonic structure and time-varying characteristics of acoustic musical instruments using a sum-of-sinusoids model. Research was based on analysis/synthesis systems, in which the qualities of the analysis models were evaluated by synthesizing signals from analysis data for comparison with original sounds.

The early analysis/synthesis efforts using digital methods were all pitch-synchronous, meaning that the input signal was divided into equal length segments approximately one pitch-period in duration (pitch-period =  $1/\text{fundamental frequency}$ ). This allowed a standard Fourier series representation for each cycle of the input waveform. The Fourier series coefficients were obtained either by an explicit summation calculation or using a fast Fourier transform (FFT) algorithm [Luce, 1963; Freedman, 1965; Beauchamp, 1966; Moorer, 1978; Strawn, 1987]. The time evolution of the amplitude and phase of each partial could be ascertained by examining changes in the Fourier series coefficients from cycle to cycle.

These pitch-synchronous analysis methods required close to an integral number of digital samples per cycle of the input waveform. Since the period of the signal was not necessarily a multiple of the sample period, linear interpolation was often used to generate values between the given sample points [Luce, 1963; Beauchamp, 1969]. The continual adjustment necessary to



match the input period with the length of the analysis window for signals having substantial vibrato or other frequency variations was problematical.

More recently, the sinusoidal analysis procedure (for speech) of McAulay and Quatieri [1986] has been applied to the problem of time-variant analysis of musical signals [Serra, 1986; Smith and Serra, 1987; Maher and Beauchamp, 1988]. In this method, each "peak" in a high-resolution STFT is attributed to an underlying sinusoidal component with time-varying amplitude, frequency, and phase. While using a fixed window length, the analysis procedure makes no attempt to synchronize with the fundamental frequency of the input signal. However, interpolation of the short-time magnitude spectrum is used to improve the frequency resolution of the analysis and to avoid problems due to mismatch between the period of the input signal and the length of the analysis transform. A nearest-neighbor tracking and matching procedure is used to connect the features (peaks) of each frame of the STFT with corresponding features in adjacent frames, providing a list of sinusoidal component trajectories (tracks) in terms of amplitude and frequency vs. time. A version of the McAulay and Quatieri method (referred to as the 'MQ' method for the remainder of this dissertation) will be presented in detail in Chapter 3.

### 2.3 Co-channel Speech Separation

In the case of speech there have been several investigations of the performance of human listeners in monaural and binaural voice-separation tasks [Cherry, 1953; Sayers and Cherry, 1957; Mitchell et al., 1971; Brokx and Nootboom, 1982; Zwicker, 1984; Weintraub, 1985]. There is evidence that listeners use both monaural cues and binaural processing to invoke the so-

called cocktail-party effect, in which a listener isolates the speech of a desired conversation while ignoring the numerous competing talkers in the vicinity. Yanagida et al. [1985] considered separation of the speech of several talkers by the use of multiple microphones using multi-channel deconvolution, but most other research in this area has dealt with co-channel speech separation (separation of interfering speech from a monaural input signal). Note that the co-channel case cannot take advantage of the binaural cues normally available to a listener in the classic cocktail-party effect.

Shields [1970] and Frazier et al. [1976] attacked the speech separation and enhancement task using variable and adaptive filters. These researchers attempted to improve speech quality and separation by the design of frequency selective filters to pass and reject the desired and interfering spectral components, respectively. However, if the desired and interfering signals were of nearly the same strength, simply passing or rejecting certain spectral bands could not produce an adequate separation because of the inevitable overlap of the desired and interfering spectral components. In fact, intelligibility tests by Perlmutter et al. [1977] showed that the "desired" speech obtained by the Frazier method was actually less intelligible than the unprocessed co-channel input signal itself.

Everton [1975] approached the co-channel speech problem by estimating the center frequency and bandwidth of each formant (broad spectral resonance) for the two simultaneous talkers, as well as the fundamental frequency of each talker. The frequency and formant data were then supplied to a speech synthesizer system for artificial reconstruction of the desired speech. Thus, Everton's method may be considered an analysis/synthesis procedure:

parameters of a synthesis model were derived from an analysis of the input signal. Note that the use of an analysis/synthesis method implies that the signal of interest can be represented adequately by some finite number of synthesis parameters, and that the parameters can be obtained reliably from the input signal.

Parsons [1976] developed and evaluated the harmonic selection algorithm for separating co-channel speech signals. The harmonic selection approach assumed that each talker in the composite input signal was using voiced (periodic) speech, i.e., the magnitude spectrum of each speech signal consisted of a harmonic series of peaks corresponding to the fundamental frequency and its overtones. In this method, the short-time Fourier transform (STFT) of a two-talker speech signal was obtained. For each short-time analysis frame, the fundamental frequencies of the two talkers were estimated, and the spectral features corresponding to each fundamental frequency and its overtones were separated. The separated components were used to reconstruct the desired speech signal in an STFT synthesis procedure. Although Parsons did not conduct any formal listening tests, Stubbs and Summerfield [1988] recently evaluated Parsons' harmonic selection algorithm using simultaneous vowel sounds: the harmonic selection algorithm was found to improve intelligibility for both normal hearing and hearing-impaired listeners.

More recent work on the co-channel speech separation problem has been reported using the harmonic magnitude suppression (HMS) algorithm [Hanson and Wong, 1984; Naylor and Boll, 1987]. Hanson and Wong applied their HMS algorithm to co-channel speech samples in which the desired talker was the weaker of the two voices, reasoning that a signal with a negative signal-to-

noise ratio (SNR) would be the case most in need of improvement in intelligibility. Thus, rather than attempting to estimate the parameters of the relatively weak "desired" speech, Hanson and Wong estimated the magnitude spectrum of the interfering voice, then subtracted it from the magnitude spectrum of the co-channel speech signal. In other words, the approach was to suppress the interfering speech, presumably leaving the desired speech in a more intelligible form.

The study by Hanson and Wong did not consider unvoiced (noisy) interfering speech suppression, but Naylor and Boll [1987] extended the HMS algorithm to unvoiced speech, additive noise, and channel distortion. Naylor and Boll also modified the HMS algorithm to enhance the desired signal in cases where the interfering signal was the weaker of the two components.

Lee and Childers [1988] developed a co-channel speech separation procedure employing multisignal minimum-cross-entropy spectral analysis (multisignal MCESA). The MCESA procedure was used to refine an initial spectral estimate of the desired speech by the use of the autocorrelation of the co-channel signal. The separation process consisted of two steps. First, a preliminary estimate of the desired speech signal spectrum was obtained using either Parsons' harmonic selection algorithm [1976] or the HMS algorithm of Naylor and Boll [1987]. Next, this preliminary spectral estimate and the autocorrelation function of the co-channel input signal were processed using multisignal MCESA. The desired speech was reconstructed using either STFT

methods or a linear predictive coding (LPC) speech synthesis process. Lee and Childers found that while intelligibility was improved by the separation procedure, the resulting speech was often "mechanical" and less natural in quality.

Danisewicz and Quatieri [1988] considered a co-channel speech separation approach using a sinusoidal analysis/synthesis model of speech. In this model, speech was represented as a sum of sinusoids with time-varying amplitudes, frequencies, and phases [McAulay and Quatieri, 1986]. For the co-channel input signal, Danisewicz and Quatieri calculated least-squared error estimates of the sinusoidal model parameters for both talkers, and reconstructed the desired speech using the parameter estimates in an additive synthesis procedure based on the sinusoidal model. The research included a multi-frame interpolation strategy to help predict the behavior of the model parameters during analysis frames where both talkers were at nearly the same fundamental frequency, causing difficulties with the least-squared error separation criterion. Danisewicz and Quatieri found the results to be useful for a range of signal conditions, and suggested that further work on multi-frame interpolation and continuity might improve the separation procedure.

#### 2.4 Segmentation and Analysis of Musical Signals

The analysis of musical signals includes work in psychoacoustics, perception, and modeling of timbre [cf., for example, Helmholtz, 1885; Fletcher, 1934; Berger, 1964; Saldanha and Corso, 1964; Luce and Clark, 1967; Grey, 1975; Benade, 1976; Beauchamp, 1981; Slawson, 1982; Gordon, 1984; McAdams, 1984; Dolson, 1985; Wessel, 1985], and other work has treated digital

signal processing applications in music processing [Dodge and Jerse, 1985]. For example, Stockham [1971] applied several DSP methods in the restoration and enhancement of old mechanical recordings of Enrico Caruso.

Wold and Despain [1986] have reported on preliminary work in the separation of composite musical signals by parameter estimation in structurally accurate nonlinear physical models of each sound source. They report that up to 300 states must be estimated in order to separate two clarinet sounds--a formidable computational task. Moreover, such an approach involves re-estimation of many states whenever the derived model changes (at note boundaries, for example), and distortions of the model parameters may be necessary to account for room resonances, acoustic transmission effects, and so forth.

Some of the relevant work on the analysis of musical recordings has been intended for automatic transcription of musical pitch and timing information into standard music notation or some other form of tabulation [Moorer, 1975; Piszczalski and Galler, 1977; Chowning et al., 1984; Mont-Reynaud and Goldstein, 1985; Schloss, 1985]. Although the sound separation task described in this dissertation differs from the musical transcription task in several ways, the two topics do share some fundamental concerns.

Attempts to transcribe an arbitrary musical recording into notation have been fraught with difficulties due to the immense differences between various input signals. A wide range of variables due to choice of orchestration, musical style, and common performance practices must be confronted. The transcription task requires identification of the attack and release times of

each musical note, determination of musical pitch, and correlation of identified notes with standard musical forms and constructs. A fully operational automatic system would be required to specify measures and bar lines, clefs and key signatures, note types (such as whole, half, or sixteenth), musical lines and phrases, etc. In short, the automatic transcription system must be aware of the correct musical notation form implied by the only available observation of the performance, the recorded signal itself. Sophistication at such a high level rapidly enters the domain of artificial intelligence (AI) [Chowning et al., 1984], or at least knowledge-based recognition of musical form [Chafe et al., 1982].

Moorer's doctoral dissertation [1975] was a seminal effort in the area of automatic transcription of polyphonic music. In order to simplify the task, Moorer considered only musical duets and disallowed vibrato, glissandi, staccato notes (less than 100 milliseconds in duration), and consonant tunings (simultaneous notes in which the fundamental frequency of one note was the same frequency as one of the partials of the other note). The approach was to identify the basic periodicities in the input signal using the "optimum comb" (or absolute magnitude difference function) method [Moorer, 1974], then to extract the assumed harmonics using a series of bandpass filters centered at each harmonic frequency. In other words, Moorer identified the harmonic components present at a given point in time, and then attempted to deduce a basic harmony to account for the observed frequencies. Various heuristic rules were employed to segment the raw data into notes, rhythms, and musical lines.

Moorer reported that the system identified the pitches and starting times of notes with significant accuracy but frequently underestimated their duration. This result indicated a fundamental problem with performance-based automatic transcription: A performer uses his experience with a particular musical style and musical instrument in such a way that the notes actually played may differ substantially in a quantitative sense from the printed score. Thus, quantitative measurements may not be enough to produce a transcription suitable for music printing.

Piszcalski and Galler [1977] attempted to transcribe recordings of monophonic music using an approach similar to but different from that of Moorer. They obtained a time-variant spectral analysis using fast Fourier transforms (FFTs) of successive segments of the input signal. Next, spectral peaks in the analysis data were used to infer which musical pitches might best account for the observation on a frame-by-frame basis, note boundary decisions were made, and a printed score was produced. The attempts of Piszcalski and Galler did not consider polyphonic input, but did include effort toward a higher level of analysis: use of musical constructs to infer a performer's intent from a recording of an actual musical performance.

Since Moorer's thesis work at Stanford University (1975), investigation of automatic transcription has continued at Stanford's Center for Computer Research in Music and Acoustics (CCRMA, pronounced 'karma'), [Chafe et al., 1982; Foster et al., 1982; Chowning et al., 1984; Mont-Reynaud and Goldstein,



1985; Schloss, 1985]. The common approach has followed Moorer's pattern: a time-frequency analysis is followed by one or more stages of segmentation and refinement based on musical knowledge, pattern recognition, and various heuristic methods.

Chowning and Mont-Reynaud [1986] reported further work at Stanford in artificial intelligence applications for musical analysis and transcription. The recent work has focused primarily on the representation and automatic identification of simultaneous acoustic sources in a monaural signal, which is closely related to the work reported herein. However, while the increased level of abstraction evident in the work of Chowning and Mont-Reynaud provides more insight into the areas of AI and machine perception, it focuses less on the pragmatic issues related to signal processing system design and implementation, as considered in this dissertation. For the moment, it is sufficient to acknowledge that any complete solution to the problems of automatic transcription and composite signal separation will undoubtedly require both cognitive (i.e., AI) and procedural (i.e., DSP) processing.

CHAPTER 3  
RESEARCH APPROACH AND METHODS

Described in this chapter is the theoretical basis for the analysis/synthesis methods employed in this dissertation. The concepts associated with short-time spectral analysis and the McAulay-Quatieri approach are considered first. Next, the simultaneous frequency tracking algorithm developed for musical duet signals is explained. Finally, the various techniques for separating the composite signal components are discussed.

The fundamental approach for this research is time-frequency analysis using the short-time Fourier transform (STFT). As mentioned in Chapter 2, the STFT has been widely used in the analysis of time-varying signals, such as speech and music.

For the duet separation problem the approach is to identify and separate the spectral components belonging to each voice from a sequence of high-resolution, short-time spectra of the composite signal. The STFT provides such a representation, so it was chosen as the analysis front-end for this investigation.

### 3.1 Short-time Fourier Transform Analysis

The STFT may be expressed in discrete form as [Allen and Rabiner, 1977]

$$X(n, k) = \sum_{m=-\infty}^{\infty} w(n-m)x(m)e^{-j2\pi mk/L} , \quad (3.1)$$

with the definitions

$x(m)$  = a signal defined for any sample time  $m$   
 $w(m)$  = a lowpass impulse response (window) function defined for any  $m$   
 $L$  = number of equally spaced (in frequency) analysis  
         channels between 0 Hz and the sample rate  
 $X(n,k)$  = short-time Fourier transform of  $x(m)$  at every sample time  $n$ ,  
         at normalized radian frequency  $2\pi k/L$ , where  $2\pi$   
         corresponds to the sample rate

For the remainder of this thesis, the window function  $w(m)$  is assumed to be real with even symmetry about the origin (noncausal, zero phase) and nonzero only for a finite range of points centered about the origin (see Harris [1978] for a description of various window functions). Note that with  $w(m)$  nonzero only for a finite range of  $m$ , the summation in (3.1) becomes finite. The STFT analysis Equation (3.1) takes the one-dimensional signal  $x(m)$  and generates a two-dimensional representation,  $X(n,k)$ . The corresponding synthesis equation will be discussed later.

Equation (3.1) can be examined from two viewpoints: filter-bank analysis and overlapped Fourier transform analysis. The two viewpoints differ only in that they represent different interpretations or implementations of the transformation expressed in Equation (3.1).

The filter-bank approach treats the STFT as a bank of  $L$  identical bandpass analysis filters, or analysis channels, centered at equally spaced frequencies,  $2\pi k/L$ . The notation  $X_k(n)$  is used here to emphasize the  $k$  subscript, i.e., the  $k$ 'th analysis channel is observed as a function of time,  $n$ . The operation is repeated for each index  $k$  at each time  $n$ .

Equation (3.1) can be rearranged to express this viewpoint as

$$X(n,k) = X_k(n) = \sum_{\text{all } m} x(m) e^{-j2\pi nk/L} w(n-m) , \quad (3.2)$$

which is either the discrete convolution

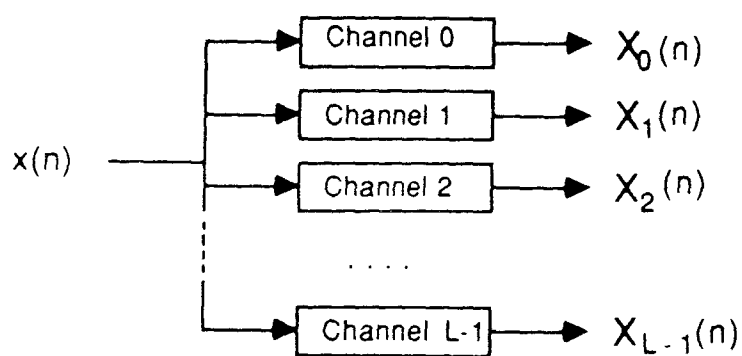
$$X_k(n) = [ x(n) e^{-j2\pi nk/L} ] * w(n) \quad (3.3a)$$

or

$$X_k(n) = e^{-j2\pi nk/L} \{ x(n) * [e^{+j2\pi nk/L} w(n)] \} . \quad (3.3b)$$

In (3.3a), the input signal  $x(n)$  is modulated by the complex exponential term, while the window remains a lowpass filter. The modulation translates the input signal's spectrum at normalized frequency  $2\pi k/L$  (corresponding to  $k \cdot (\text{sample\_rate})/L$  Hz) down to zero frequency, where it is filtered by the lowpass window response. The grouping of terms given in (3.3b) represents an equivalent interpretation. Equation (3.3b) can be derived from (3.2) using a change of variables, (e.g.,  $q = m-n$ ). The complex exponential factor  $\exp(-j2\pi nk/L)$  is a term which accounts for the difference between the time scale relative to a fixed time origin, as in (3.3a), and the sliding time scale of the window function, as in (3.3b). This linear phase shift does not change the magnitude of  $X_k(n)$ , only its phase reference [Crochiere, 1980]. The second term of the convolution in (3.3b) is a modulation of the window function  $w(n)$  by a complex exponential. The result is a frequency shift of the transform of  $w(n)$  by the amount  $2\pi k/L$  (or  $k \cdot (\text{sample\_rate})/L$  Hz),

translating its lowpass response into a bandpass response centered at frequency  $2\pi k/L$ , resulting in the filter-bank format. Either of these viewpoints is sometimes referred to as the heterodyne filter, because of the modulation of either the signal spectrum or the window function spectrum in heterodyne fashion. The filter-bank representation is shown in Figure 3.1a.



For each analysis channel 'k':

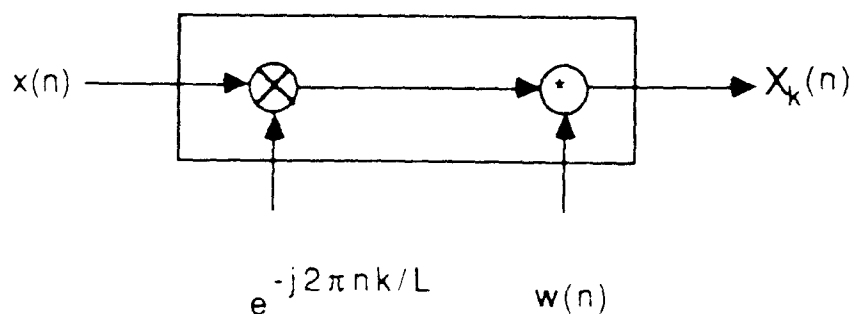


Figure 3.1a: Filter-bank Analysis Viewpoint of the STFT.

---

The second viewpoint of Equation (3.1) is as an overlapped Fourier transform. This viewpoint can be expressed as

$$X(n,k) = X_n(k) = \sum_{\text{all } m} y(m) e^{-j2\pi mk/L}, \quad (3.4)$$

where  $y(m) = w(n-m)x(m)$ , the windowed input signal.

In other words, an intermediate signal  $y(m)$  is computed by reversing and shifting the analysis window, then multiplying it by the input signal  $x(m)$ . In this case the notation  $X_n(k)$  can be used to emphasize that the Fourier transform computed at time  $n$  is observed as a function of the frequency index  $k$ . For each time  $n$ , in effect, "snapshots" of the spectrum of  $x(n)$  are computed. This representation is shown in Figure 3.1b.

Recognizing that the limits on the summation in Equation (3.4) must be [zero] and [L-1] in order to compute  $X_n(k)$  in normal DFT form, the substitution  $\{ q = m-n \}$  may be made in (3.4) to obtain

$$X_n(k) = e^{-j2\pi nk/L} \sum_{\text{all } q} w(-q) x(n+q) e^{-j2\pi qk/L}, \quad (3.5)$$

which may be rewritten as a summation segmented into blocks of length  $L$ , using  $\{ q = pL + r \}$ ,

$$X_n(k) = e^{-j2\pi nk/L} \sum_{\text{all } p} \sum_{r=0}^{L-1} w(-pL-r) x(n+pL+r) e^{-j2\pi(pL+r)k/L}. \quad (3.6)$$

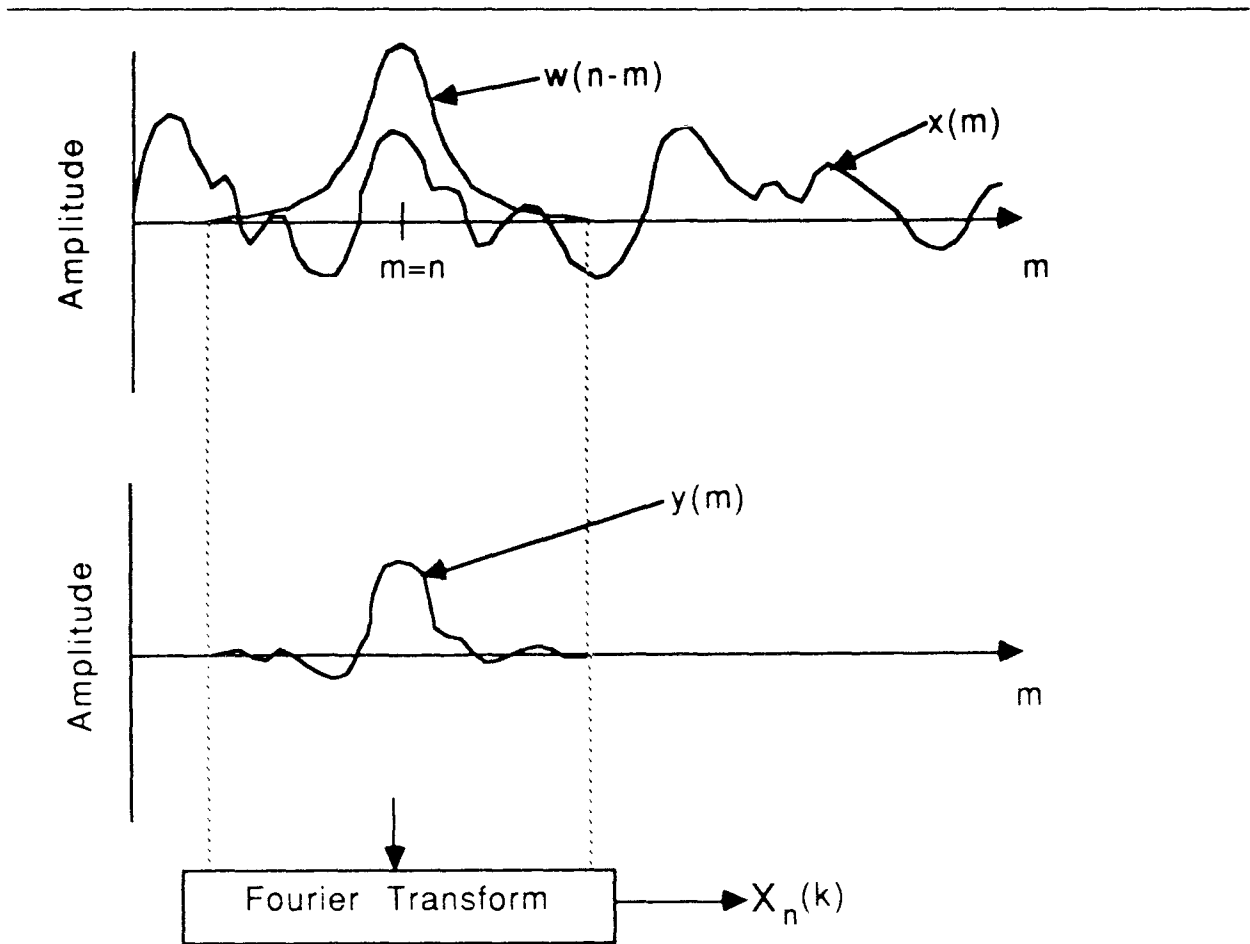


Figure 3.1b: Fourier Transform Viewpoint of the STFT.

Exchanging the order of summation,

$$X_n(k) = e^{-j2\pi nk/L} \sum_{r=0}^{L-1} \sum_{\text{all } p} w(-pL-r) x(n+pL+r) e^{-j2\pi(pL+r)k/L}, \quad (3.7)$$

and noting that with the total non-zero length of  $w(m) \leq L$ , the product  $w(-pL-r) x(n+pL+r)$  can be non-zero ONLY for  $p=0$  (since  $w(k) = 0$  for

$k \geq L/2$  or  $k < -L/2$ ),

$$X_n(k) = e^{-j2\pi nk/L} \sum_{r=0}^{L-1} w(-r) x(n+r) e^{-j2\pi rk/L} \quad (3.8)$$

The multiplication by  $\exp(-j2\pi nk/L)$  in (3.8) can be accomplished by a circular shift of the windowed data prior to the summation, and the summation itself can now be calculated using an FFT algorithm [Portnoff, 1976].

The Fourier transform viewpoint requires a series of overlapping short-time Fourier transforms of the input signal. The overlap may seem unnecessary, considering that the original signal can be reconstructed exactly from the inverse transforms of concatenated nonoverlapping segments. This observation would be useful and reasonable if the only interest was in obtaining an identity analysis/synthesis procedure. However, for the duet separation problem (and for other tasks) it may be useful to interpret and modify the frequency domain representation of the signal, which requires knowledge of the signal for every frequency index  $k$  at every time  $n$ .

Since Equation (3.1) can be expressed and interpreted in either the filter-bank or Fourier transform viewpoints, the form which is most useful or convenient for a particular task can be chosen.

At this point in the description  $X(n,k)$  consists of  $L$  complex numbers for every time  $n$ . Thus, it would appear that the storage requirements of the STFT are  $2L$  times the storage needed for a real signal  $x(n)$ ! However, these data explosions can be reduced by careful examination of the STFT representation.



### 3.2 Sampling the STFT: $X(n,k)$

So far, each input sample has been represented by  $L$  complex numbers in the STFT,  $X(n,k)$ . One question is whether the function  $X(n,k)$ , considered as a time sequence in  $n$ , can be sampled at a lower rate while still obeying the Nyquist theorem. Another question is how big  $L$  must be to obtain an adequate frequency resolution in the STFT representation of the signal. In short, we wish to find the appropriate time and frequency sample rates.

The appropriate time sampling rate for  $X(n,k)$  can be selected by examining the filter-bank representation of the STFT. Since each of the channel sequences  $X_k(n)$  was shown to be the output of a lowpass filter with impulse response  $w(n)$ , each of the sequences must have a bandwidth less than or equal to that of the window. So if the window transform is bandlimited to  $B$  Hz, the output sequences  $X_k(n)$  need only be sampled at  $2B$  Hz or greater to satisfy the sampling theorem. Thus, if the input signal has a sample rate of 20 kHz and the data window is bandlimited to 250 Hz, the STFT  $X(n,k)$  must be evaluated at a rate of at least 500 Hz to avoid aliasing. A 500 Hz frame rate for a signal sampled at 20 kHz means the analysis need only be done every 40 samples of the input signal ( $20000/500$ ), rather than at every sample. This spacing between frames can be called the analysis hop,  $R$ .

The frequency sampling density (size of  $L$ ) can be realized by noting that the inverse Fourier transform of  $X(n,k)$  is time-limited due to the finite length of the window function. Thus, we may apply the frequency domain equivalent of the sampling theorem for bandlimited signals. The frequency sampling theorem for time-limited signals states that a signal of total

duration  $N$  must be sampled at  $L \geq N$  equally spaced frequencies between zero frequency and the time-sampling frequency  $f_s$  to fully represent  $x(n)$  without time-domain aliasing. Note that this condition is trivially satisfied by making  $L$  greater than or equal to the length  $N$  of the analysis window function.\* If  $L$  is chosen to be greater than  $N$ , the samples outside the window interval are deliberately set to zero. This zero padding in the time domain is equivalent to bandlimited interpolation in the frequency domain and represents an increased sampling density of the short-time spectrum.

By properly choosing the window length, sample rate, bandwidth, etc., for a particular class of input signals, the STFT storage requirements can be reduced considerably. An additional savings is due to the fact that the Fourier transform of a REAL-only time sequence is conjugate symmetric in  $k$ , that is,

$$X(n,k)=X^*(n,(L-k)\text{mod } L) \quad (3.9)$$

where  $L$  is the transform length, and  $(.)\text{mod } L$  is the  $L$ -modulus of  $(.)$ , and  $X^*(.)$  is the complex conjugate of  $X(.)$ .

The STFT sampling issues discussed above apply in the case where a time domain waveform is to be resynthesized from the unmodified STFT analysis data. If modifications are made to the analysis data (e.g., to implement time scale compression or expansion), the appropriate time and frequency domain sampling

---

\*  $L$  is typically chosen to be an integral power-of-two, because many FFT algorithms impose this requirement.

rates may need to be greater than in the unmodified case in order to avoid aliasing. Synthesis from modified data is considered in a later section of this chapter.

### 3.3 Short-time Fourier Transform Synthesis

Like the ordinary Fourier transform, the STFT can be inverted to resynthesize an exact copy of the original signal. The synthesis operation can be interpreted using the same filter-bank and Fourier transform viewpoints used in the analysis step.

For filter-bank synthesis the frequency channel sequences  $X_k(n)$  are modulated back to their original frequency bands and added together at each time  $n$ . This gives a synthesis equation

$$\hat{x}(n) = \sum_{k=0}^{L-1} X_k(n) e^{-j2\pi nk/L} \quad . \quad (3.10)$$

This equation can be verified by replacing  $X_k(n)$  with the expression of Equation (3.2), giving

$$\hat{x}(n) = \sum_{k=0}^{L-1} \left( \sum_{\text{all } m} x(m) e^{-j2\pi mk/L} w(n-m) \right) e^{+j2\pi nk/L} \quad . \quad (3.11)$$

Reordering and regrouping this expression,

$$\hat{x}(n) = \sum_{\text{all } m} x(m) w(n-m) \sum_{k=0}^{L-1} e^{+j2\pi(n-m)k/L} \quad . \quad (3.12)$$

Noting that the summation over  $k$  of the complex exponential is zero EXCEPT when the quantity  $(n-m)$  is an integer multiple of  $L$ , i.e.,  $0, +/-L, +/-2L, \dots$ , for which the sum is  $L$ . Thus, the summation over  $k$  can be replaced by a series of  $L$ -weight unit sample functions (delta functions,  $\delta(n)$ ) that "fire" only at multiples of  $L$ ,

$$\hat{x}(n) = \sum_{\text{all } m} x(m) w(n-m) \sum_{\text{all } r} L \cdot \delta(n-m-rL) \quad . \quad (3.13)$$

The summation over  $m$  is only nonzero when the delta function is nonzero, that is, only when  $\{ m = n-rL \}$ . Making this substitution,

$$x(n) = (L) \sum_{\text{all } r} x(n-rL) w(rL) \quad . \quad (3.14)$$

Choosing  $w(n)$  of duration  $N$  samples, and  $L \geq N$ , then  $w(rL)$  is zero EXCEPT when  $r=0$ . Finally,

$$x(n) = (L) w(0) x(n) \quad (3.15a)$$

or

$$x(n) = A \cdot x(n) \quad (3.15b)$$

where  $A$  is a constant scale factor, which can be included implicitly by appropriate scaling of the window function,  $w(n)$ , prior to analysis. Thus, the filter-bank synthesis procedure EXACTLY inverts the STFT.

Note that the filter-bank procedure requires  $X(n,k)$  at the original input sample rate, so if the analysis were performed with a hop other than one, the missing values of  $X(n,k)$  must be produced by interpolation of the  $X(sR,k)$  sequences. Methods for performing this task are discussed in [Portnoff, 1976].

The STFT synthesis procedure can be formulated as an overlap-add (OLA) operation. The STFT analysis frames are inverse transformed, then shifted, overlapped, and added in such a way as to resynthesize the input signal.

The desired synthesis equation is [Allen and Rabiner, 1977]:

$$\hat{x}(n) = \sum_{\text{all } m} \sum_{k=0}^{L-1} X_m(k) e^{+j2\pi nk/L} \quad (3.16)$$

Assuming that the STFT was obtained using a properly chosen analysis hop size,  $R$ , the desired synthesis equation becomes

$$\hat{x}(n) = \sum_{\text{all } s} \sum_{k=0}^{L-1} X(sR,k) e^{+j2\pi nk/L} \quad (3.17a)$$

The summation over  $k$  in (3.17a) is almost the inverse DFT of  $X(sR,k)$ , namely

$$\sum_{k=0}^{L-1} X(sR,k) e^{+j2\pi nk/L} = (L) x(n) w(sR-n) . \quad (3.17b)$$

Combining (3.17a) and (3.17b),

$$x(n) = \sum_{\text{all } s} (L) x(n) w(sR-n) \quad (3.18)$$

or simply,

$$x(n) = (L) x(n) \sum_{\text{all } s} w(sR-n) . \quad (3.19)$$

The summation in (3.19) is the sum at time  $n$  of copies of the window function  $w(n)$  shifted by multiples of the hop size,  $R$ . So if the sum in (3.19) is a constant for all  $n$  and  $s$ , the OLA process can EXACTLY invert the STFT. Fortunately, it can be shown that any function  $w(n)$ , which is bandlimited to frequency  $B=1/(2R)$  and has discrete-time Fourier transform  $W(w)$ , can be expressed using the Poisson summation formula [Allen, 1977]

$$\sum_{\text{all } s} w(sR -n) = (1/R) \sum_{\text{all } m} W(m/R) e^{-j2\pi mk/R} . \quad (3.20)$$

Since  $W(w)$  is bandlimited to  $(1/2R)$ , all terms of the summation over  $m$  on the

right hand side are essentially zero EXCEPT the  $m=0$  term.\* Thus, the Poisson formula reduces to

$$\sum_{\text{all } s} w(sR - n) = (1/R) W(0) = \text{CONSTANT for all } s, n \quad (3.21)$$

The result is that any lowpass window function can be used in the OLA procedure if the analysis frames are overlapped with a spacing of at most  $R = 1/(2B)$ . This may seem counterintuitive at first, but it may be more clear by considering what happens to the sum of overlapping window functions as the size of the hop is made smaller and smaller: The ripple in the sum of the overlapped windows can be reduced to some arbitrarily small amount, as shown in Figure 3.2. The Poisson formula shows how big the hop may be for a given window specification and performance criterion. It is important to realize that the summation term in Equation (3.19) may be a constant for certain window functions (e.g., rectangular) when the spacing of the overlapped windows is greater than the maximum predicted by the Poisson formula. In these cases the original signal can be resynthesized exactly from the unmodified short-time transform using the overlap-add procedure. However, the frequency domain representation of the signal is an undersampled version of the STFT, which may result in unacceptable time-domain aliasing if the STFT data were modified prior to resynthesis. Therefore, we may use an undersampled STFT to reduce computation only if no modification of the STFT data is required prior to resynthesis.

---

\* However, any time-limited window function cannot be strictly bandlimited, so Equation (3.21) must be taken as an approximation that can be made accurate to any arbitrary degree by the choice of window function parameters.

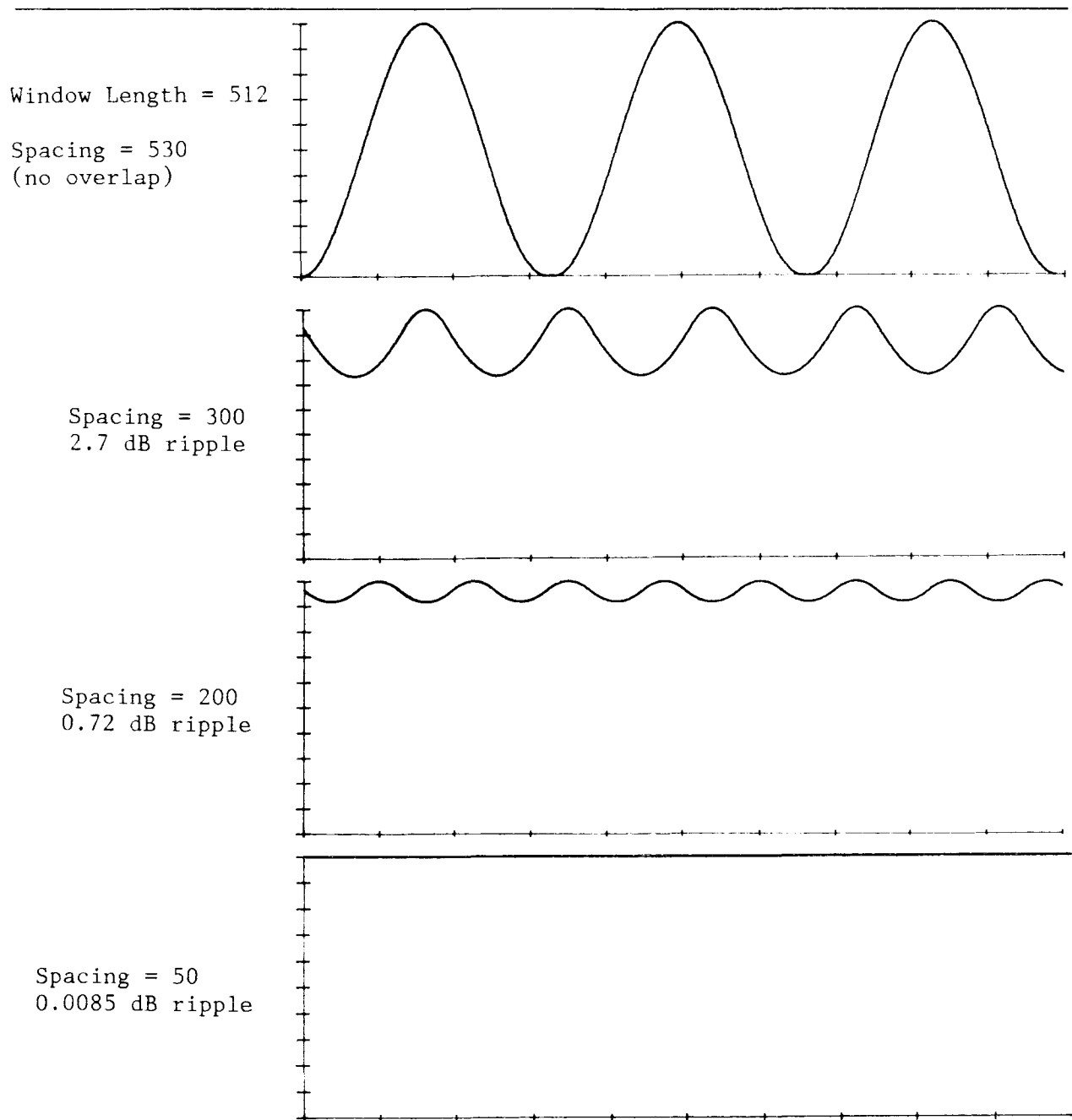


Figure 3.2: Sum of Overlapped Windows for Various Spacings.



### 3.4 STFT Synthesis with Modifications

In many applications, including the voice separation problem, it is desirable to obtain the STFT, perform some linear (or nonlinear) modifications, and then resynthesize the processed version. Examples of modifications are data quantization in a vocoder, time-varying filtering, or nonlinear time-scale and/or frequency-scale changes. The main concern for synthesis from modified STFT data is to ensure that the modifications do not violate the frequency-sampling (L) and time-sampling (R) choices made during the analysis process. For example, applying a multiplicative operation to the STFT data is equivalent to a convolution operation of the inverse transforms in the time domain. Convolution generally "smears" the time extent and detail of a signal, so any original assumptions concerning the time characteristics of the input signal must be adjusted to include the effects of the smearing.

Several of the voice separation methods described in Chapter 2 contain modifications of the STFT data prior to resynthesis, so the analysis parameters must be chosen with care. Depending on the nature of the input signal, it may be necessary to choose a different value of L and/or R to account for the change in impulse response length due to the convolution to prevent time domain aliasing (audible reverberation) during synthesis.

### 3.5 A Variation of the STFT Concept: The MQ Analysis/Synthesis Procedure

McAulay and Quatieri (1986) proposed an analysis/synthesis procedure for speech based on a sinusoidal representation. Their approach was to model speech waveforms as a sum of possibly inharmonic, time-varying, sinusoidal

components. Described in this section is the basic MQ analysis/synthesis model, and the modifications made for the duet decomposition problem considered in this dissertation are indicated.

The basic McAulay and Quatieri (MQ) signal model assumes a priori that each segment of the input,  $x(n)$ , consists of a finite number of sinusoidal components,  $J$ . Each component may have arbitrary amplitude ( $\alpha_k$ ), angular frequency ( $\omega_k$ ), and phase ( $\phi_k$ ). Thus,

$$x(n) \approx \sum_{k=1}^J \alpha_k \cos(\omega_k \cdot n + \phi_k) . \quad (3.22)$$

Note that both harmonic and inharmonic signals can be accommodated in the MQ model, unlike the equally spaced filter bank of the standard STFT.

Like the STFT, the MQ process assumes that the parameters of the signal may be time-variant, so the amplitude, frequency, and phase parameters must be updated frequently to remain a valid representation of a time-varying input signal. The computational challenge for this model resides primarily, of course, in the decomposition of the input signal into sinusoidal components with meaningful amplitude, frequency, and phase parameters.

Synthesis using the model can be accomplished via a simple additive procedure, where each of the  $J$  components is regenerated by a sinusoidal oscillator with amplitude, frequency, and phase modulation applied according

to the analysis model parameters. However, extra care is required to "unwrap" the phase parameter, due to the inherent range ambiguity of the principal value.

According to the original MQ analysis algorithm, the input signal is segmented into blocks of length  $N$  (possibly overlapping), each block is windowed with an appropriate lowpass window (as in the STFT analyzer), and its discrete Fourier transform is computed via an FFT algorithm. For each DFT the magnitude spectrum is calculated, and all peaks in the spectrum are identified simply by searching for groups of three adjacent spectral samples where the magnitude's slope changes from positive to negative. McAulay and Quatieri assume that each peak may be attributed to the presence of an underlying sinusoidal component during the current segment of the input signal. Once all the peaks are determined, the complex (real,imaginary) spectrum is used to identify the phase information for each peak, and the amplitude, frequency, and phase values for each peak are stored in a data structure. The number of peaks chosen in each data frame can be limited by 1) selecting only the tallest  $K$  peaks or by 2) imposing some amplitude threshold. The MQ analysis procedure is depicted in Figure 3.3.

One difficulty in the peak identification procedure is due to the limited density of frequency points resulting from the discrete Fourier transform (DFT). Indeed, the actual frequency of an underlying sinusoidal component may lie BETWEEN the frequency samples of the DFT. This limitation can be reduced by 1) increasing the density of the DFT frequency samples by means of a longer zero-padded DFT and by 2) using an interpolation method on the magnitude spectrum itself [Smith and Serra, 1987].

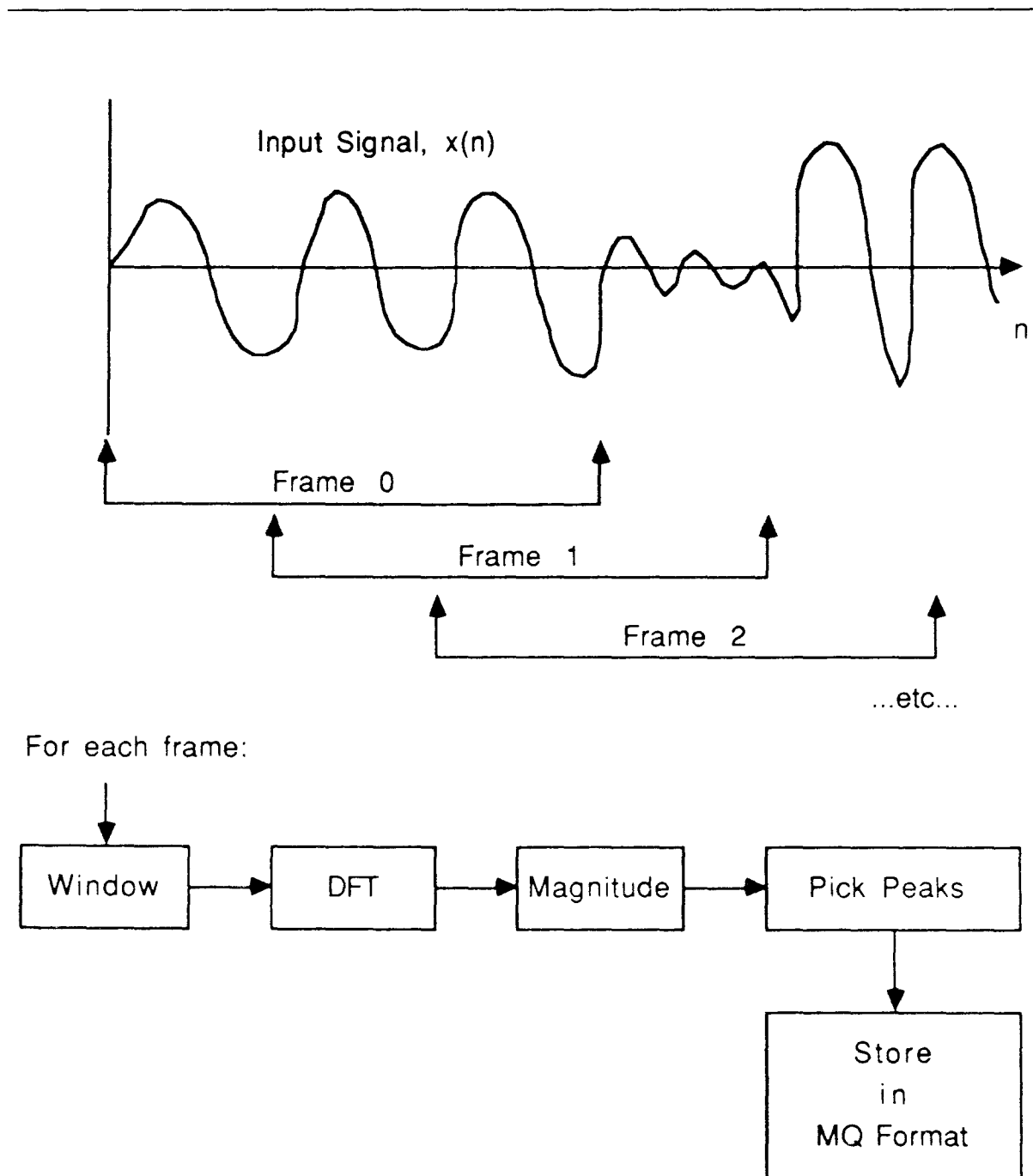


Figure 3.3: The Basic MQ Analysis Procedure.

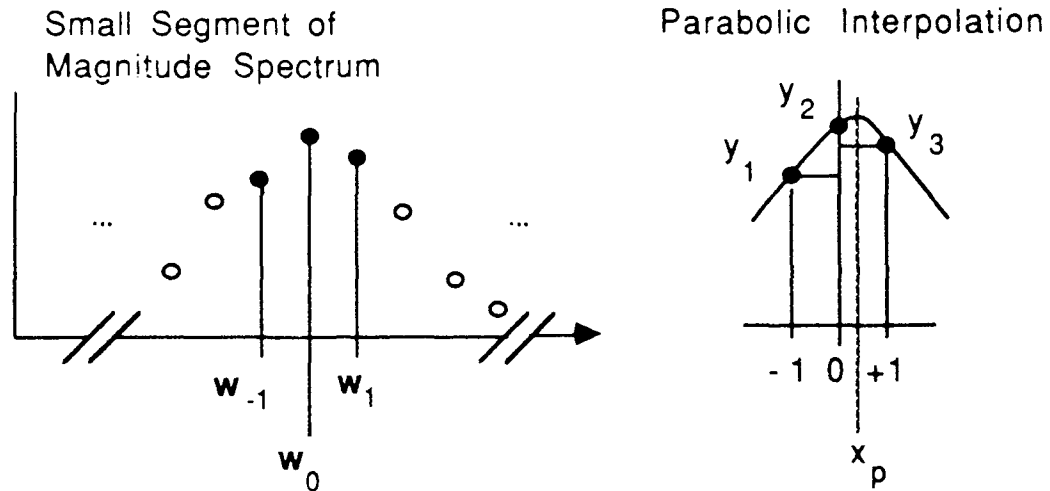
---

The frequency density of the DFT is increased by zero-padding the input data frame prior to the FFT procedure. Typically, padding has been used to increase the FFT frame length to be the next higher power of 2. As mentioned previously, the zero-padding of the time-domain input frame corresponds to band-limited interpolation in the spectral domain. However, zero-padding adds to the length of the input frame, increasing the computation required to calculate the DFT. Thus, the computation load can be reduced by choosing a padding factor only as large as is necessary for an efficient secondary interpolation method to achieve the desired accuracy.

A simple second-order (3-point) interpolation scheme is used to refine the estimated peak location between the DFT points [Smith and Serra, 1987]. The use of second-order interpolation implies that the spectral magnitude shape in the vicinity of the peak is nearly parabolic, which is reasonable if the spectrum has been sufficiently oversampled by zero-padding.

The three points comprising the identified peak define a unique parabola passing through them. Constructing a coordinate system in which the arbitrary abscissa units  $-1, 0, +1$ , are aligned with the equally spaced DFT points surrounding the peak, a simple system of equations yields the parabolic peak location, as shown in Figure 3.4. This identifies both the amplitude and frequency of the underlying sinusoid attributed to the peak.

After the frequency and amplitude estimates are obtained from the magnitude spectrum, the complex coordinates (real,imaginary) of the peak are estimated by performing independent second-order interpolations for the real and imaginary parts of the three points surrounding the peak. These functions



coordinates:  $(-1, y_1), (0, y_2), (+1, y_3)$

frequencies:  $\omega_{-1}, \omega_0, \omega_1 \rightarrow$  DFT point frequencies

parabola:  $y(x) = A \cdot x^2 + B \cdot x + C$

equations: I.  $y(-1) = y_1 = A - B + C$

II.  $y(0) = y_2 = C$

III.  $y(+1) = y_3 = A + B + C$

results:  $A = (y_1 - 2y_2 + y_3)/2; B = (y_3 - y_1)/2; C = y_2$

x-coordinate of the parabola peak:

$$y'(x_p) = 0 = 2A \cdot x_p + B$$

$$x_p = -B/2A$$

$$= 0.5(y_1 - y_3)/(y_1 - 2y_2 + y_3)$$

frequency estimate for peak:

$$\omega = \omega_0 \cdot (1 - x_p) + \omega_1 \cdot x_p$$

magnitude estimate for peak:

$$\alpha = y(x_p)$$

Figure 3.4: Parabolic Interpolation of Spectral Peak Location.

are evaluated at the frequency obtained from the magnitude parabola maximum, and the resulting complex coordinate provides an estimate of the phase of the underlying sinusoid, i.e.,  $\phi_k = -\text{atan}(\text{imag}/\text{real})$ .

The analysis and peak identification process is repeated for successive frames of the input signal. The spacing,  $R$ , between the frames (the hop size) is chosen as a tradeoff between the computation expense associated with a small hop, and the loss of time resolution due to a large hop. The hop size can be selected according to the frequency-domain sampling criteria described previously for the STFT analyzer.

In most practical cases the sound spectrum presented to the MQ analyzer varies considerably with time, so that the number of detected components and their frequencies will, in fact, change from frame to frame. For this reason a matching procedure is performed to connect components from frame  $[i]$  with corresponding ones from frame  $[i+1]$  and thus track the time-varying sinusoidal components. The following steps are used in the tracking procedure:

- 1) The frequencies of components (i.e., the peaks) identified in each frame are sorted from lowest to highest.
- 2) Each peak in frame  $[i]$  is compared to the peaks in frame  $[i+1]$ . If the frequency of a peak in frame  $[i+1]$  lies within an arbitrary capture range of a peak in frame  $[i]$ , and no other match is better, then the two matching peaks are linked in the analysis data base. The capture range specifies how much the frequency of a component may change between analysis frames and still be considered a valid match.

- 3) Any peak in frame [i] that cannot be matched to any of the peaks in frame [i+1] is considered "dead," and a null peak is inserted into frame [i+1]. The null peak is assigned zero amplitude, the same frequency as its progenitor, and a phase value calculated from the previous phase value, the frequency, and the known time interval (R/sample rate) between frames (using  $\text{phase} = \int \text{frequency } dt$ ).
- 4) Similarly, any peak in frame [i+1] that was not matched by a peak in frame [i] is "born," and a corresponding null peak is inserted into frame [i].

Once the peak-matching process on the current frame is complete, the procedure is repeated for the subsequent data frames. The final output database consists of chains of peaks, or "tracks," which trace the behavior of the underlying sinusoidal components comprising the signal, as shown in Figure 3.5.

Synthesis with the MQ method is performed with an additive synthesis procedure based on the MQ analysis data. Each frequency track is used to control a sinusoidal oscillator, whose amplitude, frequency, and phase are modulated in such a way that they exactly match the measured values at the frame boundary times and change smoothly between frames. Linear interpolation has been found to be adequate for the amplitude values, but the frequency and phase values require more careful treatment [McAulay and Quatieri, 1986; Smith and Serra, 1987].

Since the phase values are obtained relative to the sliding analysis frame reference, i.e., modulo  $2\pi$ , some means must be incorporated to choose



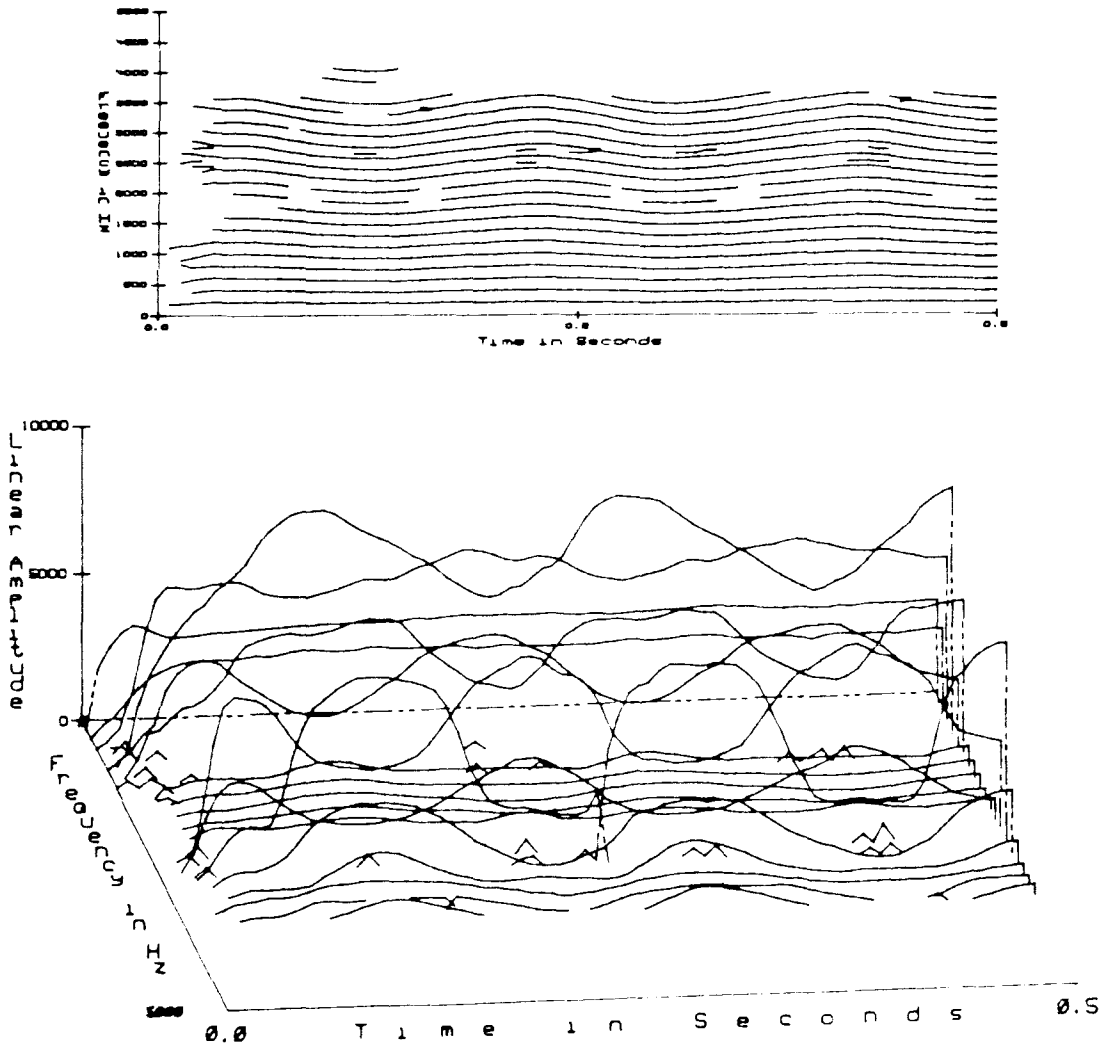


Figure 3.5: MQ Analysis Representation of a Time-varying Signal.

the correct phase angle from the measured principal angle. The process of selecting the actual phase value from the modulo  $2\pi$  principal value is called

phase unwrapping.<sup>\*</sup> Since the continuous-time phase and frequency functions are related by the time derivative, they are incapable of fully independent variation. The problem becomes one of choosing a phase interpolation function between each pair of linked peaks whose slope (instantaneous frequency) matches the measured frequency at the frame boundaries, and whose phase corresponds to the measured phase unwrapped to provide a maximally smooth frequency function. The phase and frequency continuity constraints (phase and its time derivative specified at frame boundaries) can be met with at least a cubic function, and the smoothness constraint can be quantified by minimizing the integral of the function's squared second derivative. In other words, the unwrapped phase which makes the frame-to-frame phase function closest to linear is chosen. Using a continuous cubic phase function [McAulay and Quatieri, 1986],

$$\Theta(t) = A \cdot t^3 + B \cdot t^2 + C \cdot t + D \quad (3.23a)$$

with time derivative (frequency),

$$\Theta'(t) = 3A \cdot t^2 + 2B \cdot t + C \quad , \quad (3.23b)$$

where  $t$  is a continuous time variable equal to zero at frame boundary  $[i]$ , and equal to  $T$  at frame boundary  $[i+1]$  ( $T = R \cdot \text{sample\_rate}$ ). Four boundary conditions on the phase function  $\Theta$  can be identified:

---

\* Note that if  $R$  were sufficiently small this would not be a problem.

- I:  $\Theta(0) = D =$  phase measured at frame [i]  
 II:  $\Theta'(0) = C =$  frequency measured at frame [i]  
 III:  $\Theta(T) = A \cdot T^3 + B \cdot T^2 + \Theta'(0) \cdot T + \Theta(0) + 2\pi M$   
        $=$  phase measured at frame [i+1]  
 IV:  $\Theta'(T) = 3A \cdot T^2 + 2B \cdot T + \Theta'(0)$   
        $=$  frequency measured at frame [i+1]

Thus, the parameters A, B, C, and D are readily calculated as a function of the "unwrapping" parameter, M. Using McAulay and Quatieri's smoothness criterion, M is chosen to minimize the expression [McAulay and Quatieri, 1986]

$$f(M) = \int_0^T [\Theta''(t;M)]^2 dt \quad (3.24)$$

giving M as the integer closest to

$$q = \{ [\Theta(0) + \Theta'(0) \cdot T - \Theta(T)] + [\Theta'(T) - \Theta'(0)] \cdot (T/2) \} / (2\pi) . \quad (3.25)$$

This provides the parameters A(M), B(M), C, and D for the chosen cubic function. The synthesis procedure is implemented on a block-by-block basis as the sum of the sinusoidal oscillators with linear amplitude interpolation and cubic phase interpolation, viz.

$$x(m) = \sum_{k=\text{track } 1}^J \text{AMP}(m,k) \cdot \cos(\Theta(m,k)) \quad (3.26)$$

where  $m=0, \dots, R-1$  is the sample time index between frames [i] and [i+1], J is

the number of frequency tracks at frame [i], and where AMP and  $\Theta$  are given by

AMP(m,k)= linear amplitude interpolation between matching peaks in  
frames [i] and [i+1], over R samples, at the m<sup>th</sup> sample  
for the k<sup>th</sup> frequency track

$$= \alpha(0,k) + m[\alpha(T,k) - \alpha(0,k)]/R$$

$\Theta(m,k)$ = cubic phase interpolation between matching peaks in  
the frames, over R samples, at the m<sup>th</sup> sample  
for the k<sup>th</sup> frequency track

$$= A(M) \cdot m^3 + B(M) \cdot m^2 + C \cdot m + D$$

The MQ process can be extended to allow a wide range of signal modifications. For example, the peak-matching and smooth-phase interpolation methods are useful for splicing and editing sound segments without clicks or pops. Also, the frequency tracks obtained from the analysis step can be scaled for shifting pitch without changing the evolution of the sound with respect to time. Similarly, time compression or expansion without pitch change can easily be accomplished within the MQ model.

Analysis and synthesis results obtained using the modified MQ procedure described in this section have been surprisingly good for a variety of sound sources, considering the simplicity of the model and the arbitrary design of the peak identification procedure and the cubic phase interpolation process. In many informal experiments accompanying this research, the MQ process was applied to isolated musical tones, speech, singing, and polyphonic music. With careful listening, the synthesis output was found to be perceptually distinguishable from the original input signal. In particular, the character of noise-like components of the input signal was often noticeably altered in

the synthetic sound, presumably due to an inadequate characterization of the noisy material by the sum-of-sinusoids MQ model. On the other hand, because of this attribute, the MQ procedure shows some potential for noise reduction of recordings, particularly if a careful choice of thresholds is made for the peak-picking step of the analysis. For the most part the synthesis was not found to be "better" or "worse" than the original, only "different." This informal result is encouraging, because it indicates that the MQ model retains the essence of the original recording, and therefore, we are led to believe that the information necessary for separating duets is present in the MQ analysis data.

### 3.6 Tracking the Fundamental Frequencies in a Duet Signal

The voice separation strategies considered in the next section require fundamental frequency estimates for each voice. This information could come from several possible sources, such as an accurate musical score, an automatic frequency tracking system, or a manual means of tabulation. Automatic methods are of primary interest in this dissertation.

#### 3.6.1 Common methods for pitch detection

In the literature, fundamental frequency tracking is often called pitch detection or extraction. Dozens of papers and reports describing algorithms for pitch detection have been published, particularly for applications in speech analysis and processing. The list of methods includes the cepstrum [Noll, 1966], autocorrelation [Sondhi, 1968], the period histogram and other harmonic-based methods [Schroeder, 1968; Piszczalski and Galler, 1979], the

"optimum comb" and average magnitude difference function (AMDF) [Moorer, 1974; Ross et al., 1974], and methods based on linear prediction [Markel and Gray, 1976; Rabiner and Schafer, 1978]. These methods were developed primarily for estimating the time-varying fundamental frequency of a single sound source. The original plan for this investigation was to modify an existing single-source pitch detection algorithm for use in the duet analysis problem. However, evaluation of several potential algorithms revealed a set of inherent problems.

For methods such as autocorrelation and the optimum comb, periodicities in the input signal are identified by searching for a delay lag,  $T_0$ , which maximizes the integrated product (autocorrelation) or minimizes the integrated absolute value of the difference (optimum comb and AMDF). The fundamental frequency estimate is then  $f_0 = 1/T_0$ . The search for the extremum is problematic, because the autocorrelation function and AMDF are not unimodal: many subextrema are present. These methods are particularly sensitive to octave errors and other problems. In the case of speech, octave errors can often be avoided by restricting the search span to a range less than one octave, but musical signals generally span a larger frequency range. Moreover, when two sources are present in the input signal, interactions between the numerous pairs of partials cause additional difficulties. In short, these methods were found to be impractical for the duet case.

The cepstrum is defined as the power spectrum of the logarithm of the power spectrum [Noll, 1966]. Using a signal model in which a single periodic source signal (excitation) is convolved with a system response function (filter), the spectrum consists of the product of the source transform and the

system transform. For pitch detection the cepstrum is used to isolate the spectral ripple due to the partials of the source signal from the spectral variations due to the system response function. Assuming that the source signal appears as a "high-frequency" ripple in the power spectrum, while the system transform contributes a "low-frequency" factor, the periodicity of the input signal can be evaluated by separating the power spectrum according to frequency. Logarithms of products can be expressed as sums of logarithms, so the logarithm of the power spectrum is the sum of the logarithm of the source signal power spectrum and the logarithm of the system function power spectrum. Taking the power spectrum again (obtaining the cepstrum) reveals a low-frequency component due to the system function, and a high-frequency peak corresponding to the period of the source function. The source fundamental frequency is the inverse of this period. Unfortunately, the convenient separation of source and system fails when two signals are present, as in the duet case. Here the power spectrum will contain the cross products between the source components and the two system functions, making it impossible to perform a clean separation and identification of the individual fundamentals.

Harmonic-based methods, such as the Schroeder histogram approach (1968), provide a somewhat better platform for pitch detection of polyphonic signals. The Schroeder method takes a list of harmonic frequencies and computes a series of submultiples for each entry in the list. For example, a frequency component at 100 Hz could be the fundamental of a 100 Hz signal, the second

partial of a 50 Hz signal, the third partial of a 33.333 Hz signal, etc. The list of submultiples is partitioned into a histogram, which counts the "votes" for each possible fundamental frequency. The histogram bin containing the most votes denotes the most probable fundamental.

The histogram method can suffer from the octave error problem, because the histogram bin one octave below a given bin will be a valid fundamental for all the components matching the higher frequency bin. As with the AMDF, octave errors can be avoided if some specific information is known about the frequency characteristics of the input signal, or if the search range can be limited to less than one octave.

For the sum of two harmonic spectra the histogram approach can still be effective. In fact, it was used in co-channel speech separation work [Parsons, 1976; Stubbs and Summerfield, 1988]. The basic procedure is divided into two passes: In the first pass, the largest histogram bin is identified as the fundamental frequency of one of the voices. All harmonics matched by the fundamental frequency from the first pass are removed from the list of partials; then the histogram process is repeated for the remaining partials. Note that this method assumes the partial frequencies can be determined somehow from the composite signal without problems due to frequency measurement inaccuracy and collisions between the partials of the two voices.

### 3.6.2 A new duet fundamental frequency tracking approach

The duet frequency tracking method developed for this dissertation combines some of the concepts described in the last section. The approach is



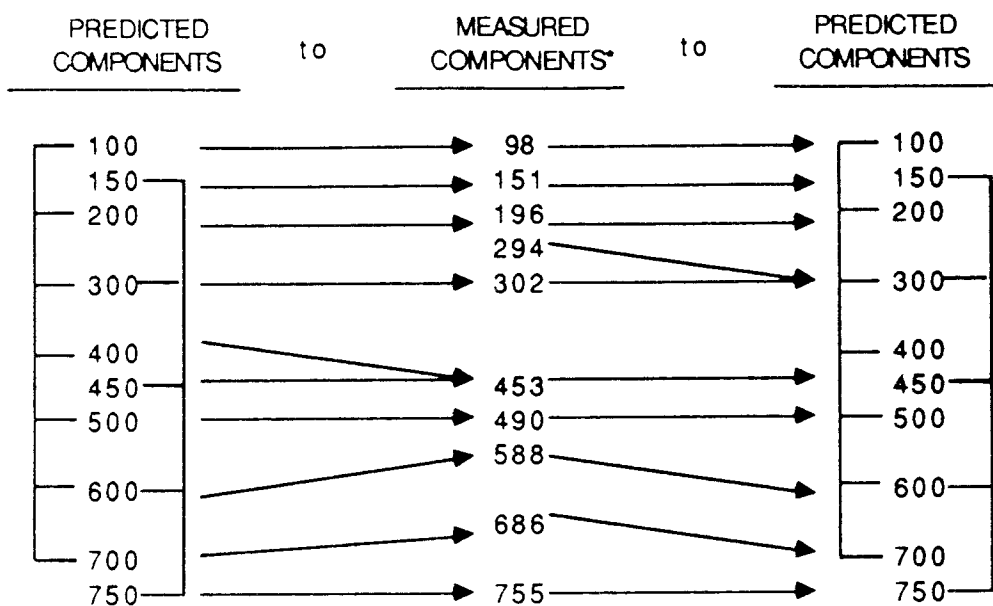
to choose a pair of fundamental frequencies which together minimize the mismatch between the predicted partial frequencies (harmonics of the two fundamentals) and the list of observed frequencies from the MQ analyzer. The mismatch error is calculated as the sum of squared normalized differences between each predicted partial frequency and the nearest measured partial frequency, and between each measured partial frequency and the nearest predicted partial frequency. This "two-way mismatch" calculation has two advantages: 1) It favors frequency choices which correctly predict the measured components, and 2) it does not predict components that are not found in the set of measured partials. An example of the two-way mismatch calculation for a particular pair of estimated frequencies is given in Figure 3.6.

The two-way mismatch calculation was devised to adjust the frequency tracker to best represent the measured set of frequency tracks. If the analysis data were noise free, we might only attempt to maximize the number of correct predictions (the coverage) of the measured frequencies, then simply choose the highest octave for a given level of coverage. However, any spurious frequencies due to noise or interference between the duet voices could cause the estimate to be very sensitive to small variations. The two-way mismatch approach helps prevent such errors by including the "bad" predictions as an unfavorable parameter in the error calculations.

In the two-way tracker implementation, the mismatch errors are weighted according to some simple rules. Specifically, if a measured frequency track has a relatively large amplitude, it is weighted more in the error calculation than a small amplitude track. That is, the error penalty for missing a large

---

Predicted Pair: 100 Hz, 150 Hz; Actual Pair: 98 Hz, 151 Hz.



\*To simulate real data, not all components are given.

Figure 3.6: The Two-way Mismatch Error Calculation.

---

amplitude track is greater than for missing a track with small amplitude. A ranking of errors from WORST to BEST (largest error penalty to smallest error penalty) is given by

- WORST (1) Missing a LARGE amplitude track by a LARGE frequency difference.  
 (2) Missing a SMALL amplitude track by a LARGE frequency difference.  
 (3) Missing a SMALL amplitude track by a SMALL frequency difference.  
 BEST (4) Missing a LARGE amplitude track by a SMALL frequency difference.

Defining  $f_e$  as the absolute difference in frequency between a predicted frequency,  $f_p$ , and the nearest measured track, frequency  $f_i$  and magnitude  $\alpha_i$ , and defining maxmag as the magnitude of the largest peak in the analysis frame, the error penalty is calculated using the following formulas:

Error from measured to predicted:

$$E_1 = (1 + 3f_e/f_i) - 0.2\alpha_i \cdot (1 - 7f_e/f_i)/\text{maxmag} \quad (3.27a)$$

Error from predicted to measured:

$$E_2 = (1 + 3f_e/f_p) - 0.2\alpha_i \cdot (1 - 7f_e/f_p)/\text{maxmag} \quad (3.27b)$$

The weights in (3.27) were determined empirically using a set of sample input signals and the error penalty criteria given previously. The total error for a given fundamental frequency prediction is given by the sum of all the errors  $E_{1,i}$  and  $E_{2,p}$  calculated for the analysis frame. Note that the global minimum of the error function will not be, in general, the only local minimum. Thus, a global search is necessary to locate the best frequency pair.

The frequency tracking procedure is provided with two nonoverlapping frequency ranges in which to concentrate its search. This information is supplied by the user from prior knowledge of the expected input signal. The search procedure first calculates the error value for a frequency pair with the upper voice frequency fixed, and the lower voice frequency increasing from the minimum of the low-frequency range to the maximum in semitone increments. When a local minimum is detected using the semitone increments, the region of the minimum is processed iteratively with a decreasing frequency increment to refine the true estimate of the minimum location.

Once the entire low-frequency range is processed and the overall minimum is obtained, the search procedure continues for a frequency pair with the lower voice frequency set to the "best" frequency obtained in the first step, and the upper voice frequency increasing in semitone steps across the high-frequency range. The global minimum pair for the entire frame is saved.

The fundamental frequency pair estimation algorithm can be summarized as follows:

- (1) The nonintersecting fundamental ranges of the lower and upper duet voices are specified,  $[f_{\min l}, f_{\max l}]$  and  $[f_{\min h}, f_{\max h}]$ .
- (2) The initial frequency pair estimate is set to  $\{f_{\min l}, (f_{\min h} + f_{\max h})/2\}$ .
- (3) The total mismatch error (3.27) for the frequency pair is calculated, and the low-frequency member of the pair is incremented by a semitone.
- (4) The error evaluation process is repeated until the low-frequency member of the pair reaches the range boundary. If a local minimum is detected,

the error evaluation process repeats the search in the vicinity with a reduced frequency increment to refine the estimate. The frequency pair with minimum error identifies the preliminary estimate of the low-frequency member.

- (5) The evaluation process of (4) is repeated for the high-frequency range. The resulting frequency pair is the initial estimate of the two fundamental frequencies in the duet.

The entire algorithm is repeated by using the initial estimate of the frequency pair just obtained as a starting point. This repetition helps insure that the true global minimum has been identified without resorting to testing every possible combination of frequency pairs.

The mismatch error calculation is performed as a global search only several times per second of the input signal. On the frames between the global search frames, the search is restricted to a semitone range ( $\pm 2.5\%$  of each fundamental frequency). If the global search turns up a better pair of frequencies outside the semitone range, the tracking process checks the intervening frames to isolate the frame at which the change occurred. By choosing the period between global searches to be less than an arbitrary minimum note duration, no frequency transitions will be missed. This approach reduces the amount of calculation necessary to track the pair of frequencies, under the assumption that the frequency pair will often remain roughly constant for many analysis frames during each musical note.

A nonlinear smoothing process is applied to each of the two outputs of the duet frequency tracker. The smoothing was found to be useful for reducing occasional single-frame errors due to signal noise. A five-point median smoother was found to be adequate for this purpose.\*

The two-way tracking algorithm requires substantial computation but provides good results. The performance tends to degrade as one voice of the duet becomes lower in amplitude than the other. In this case the tracker has little difficulty following the louder voice but has increasing trouble with the softer voice. A variety of related issues will be discussed in the next chapter.

### 3.7 The Use of the MQ Procedure in the Duet Separation Task

At first glance, the MQ model appears to be an ideal representation for the duet separation problem because the sinusoidal components of each voice should appear as independent tracks in the MQ analysis data. According to this reasoning, once the two fundamentals are determined by simply choosing the proper subset of tracks, each voice could be resynthesized separately. Unfortunately, this naive approach cannot be applied directly in most cases for reasons discussed in this section.

---

\* The median operation sorts a set of  $Q$  (usually odd) data points from smallest to largest, then returns the middle point:  
 $\{2,10,1\} \rightarrow \text{median}=2$ ;  $\{1,11,9,2,4\} \rightarrow \text{median}=4$ .

### 3.7.1 MQ frequency resolution and spectral collisions

The spectral resolution of the MQ procedure defines a limit on the number of frequency tracks present in the MQ analysis data. In this context, "resolution" measures how small the frequency difference between two equal-amplitude sinusoidal components may be before they appear to be a single peak in the magnitude spectrum. The resolution is determined primarily by the length,  $N$ , of each segment used in the DFT and the type of window function which is applied to the data. The fundamental time-bandwidth product, based on the uncertainty principle, places a constraint on the simultaneous observation of features in the time and frequency domains. Short observation intervals (small  $N$ ) provide more time resolution than long intervals (large  $N$ ), at the expense of less frequency resolution, and vice versa. In other words, time resolution and frequency resolution are inversely related. An expression for this relationship is

$$T_0 \cdot \Omega_0 \geq \frac{1}{2} \tag{3.28}$$

where  $T_0$  is the duration of the input data segment, and  $\Omega_0$  is the spectral bandwidth of the segment's Fourier transform.\* The inequality of (3.28) indicates that a windowed signal segment time-limited to  $T_0$  seconds always has a bandwidth of at least  $1/2T_0$ . The resolution problem is particularly significant when two components of unequal magnitude are closely spaced in

---

\* The numerical relationship of (3.28) may vary depending on how bandwidth is defined in a given situation.

frequency: The larger component may obscure the presence of the smaller. The ability of the analyzer to resolve closely spaced spectral components is an important consideration for the MQ process, since only the peaks in the magnitude spectrum are retained after the initial processing.

Assuming that we know the fundamental frequencies of both voices of a musical duet, the harmonic frequencies identify the spectral location of each partial. Consider the frequency pair {100, 150} Hz. The harmonics of the lower voice are (100, 200, 300, 400, 500,...), while the upper voice has partial frequencies (150, 300, 450, 600, 750,...). Note that the two voices share harmonics at (300, 600, 900,...). Moreover, because of the time-bandwidth issues just mentioned, overlaps may occur even if the spectral components are not exactly coincidental. The occurrence of harmonic sharing, or collision of spectral components, prevents a simple segregation of the frequency tracks into those belonging to one voice and those belonging to the other. More precisely, if the frequency separation of two components is smaller than the resolution limit of the analysis window transform used in the MQ process, the two components combine with each other and become a single track. Therefore, because of partial frequency collisions, most duets cannot be separated by simple segregation of the frequency tracks into groups belonging to one voice or the other.

In order to resolve spectral collisions between the partials of different voices we must ascertain the most probable contribution of each voice to the composite information observed in the short-time spectrum. Given that the two fundamental frequencies are known, it is possible to identify conflicting partials by comparing the predicted harmonic series of the two voices.



Assuming we have good estimates of the two fundamental frequencies, three approaches to this task may be considered:

- (1) For predicted partials closer than the spectral bandwidth of the analysis window a set of linear equations can be specified and solved for the contribution of each windowed sinusoid to the observed complex spectrum in the vicinity of the conflict.
- (2) For two closely spaced partials the resulting amplitude modulation (beats) and frequency modulation functions may be used to calculate the amplitudes of the colliding partials--assuming the partials' amplitudes and frequencies remain relatively constant for a period of time sufficient to estimate the various parameters involved.
- (3) With an accurate signal model for each voice, collisions of partials can be handled by synthesizing artificial amplitude/frequency tracks to replace the corrupted partials. For brief collisions, or when none of the other approaches is applicable, it may be necessary to resort to interpolation of the missing data from uncollided partial data in the same frame or in adjacent frames.

For this investigation the approach is: Compare the predicted harmonic series of the two voices to identify potential collisions, segregate all uncollided frequency tracks, and then apply one of the three methods listed above to reconstruct the colliding components. The criteria for choosing the appropriate strategy will be considered later in this chapter.

### 3.7.2 Separation strategy I: linear equations

The linear equations method for collision repair relies on several concepts and assumptions. First, because the time-domain window function,  $w(m)$ , is real, has even symmetry, and is noncausal, its DFT is real-only. Second, if the windowed input signal contains only components with constant frequency during the window interval, its spectrum will contain only copies of the window transform centered at the component frequencies. This is because multiplication of the constant frequency component by the window function is equivalent to convolving the two Fourier transforms. Third, if the window function is chosen so that its spectral bandwidth is less than both of the two fundamental frequencies, the window transform passband will cover at most two partials of the composite signal (no more than one from each of the duet voices). The significance of these assumptions will be apparent shortly.

A duet input signal,  $x(n)$ , can be expressed as the sum of the individual voice signals,  $x_1(n)$  and  $x_2(n)$ , where

$$x(n) = x_1(n) + x_2(n) \quad (3.29a)$$

and using a sinusoidal model,

$$x_1(n) = \sum_{k=1}^{J_1} \alpha_{1,k}(n) \cdot \cos(\omega_{1,k}(n) \cdot n + \varphi_{1,k}(n))$$

$$x_2(n) = \sum_{k=1}^{J_2} \alpha_{2,k}(n) \cdot \cos(\omega_{2,k}(n) \cdot n + \varphi_{2,k}(n)) \quad (3.29b)$$

where  $\alpha_{1,k}$ ,  $\omega_{1,k}$ , and  $\varphi_{1,k}$  are the time-varying amplitude, frequency (possibly inharmonic), and phase of each component, and  $J_1$  and  $J_2$  are the number of components in signals  $x_1$  and  $x_2$ , respectively. The discrete-time Fourier transform of the windowed input signal  $x(n-sR)w(n)$  can be expressed as

$$\begin{aligned} \Gamma(\omega) = & \sum_{k=1}^{J_1} \left( \frac{1}{2} \alpha_{1,k} \cdot W(\omega - \omega_{1,k}) e^{+j\varphi_{1,k}} \right. \\ & \left. + \frac{1}{2} \alpha_{1,k} \cdot W(\omega + \omega_{1,k}) e^{-j\varphi_{1,k}} \right) \\ & + \sum_{k=1}^{J_2} \left( \frac{1}{2} \alpha_{2,k} \cdot W(\omega - \omega_{2,k}) e^{+j\varphi_{2,k}} \right. \\ & \left. + \frac{1}{2} \alpha_{2,k} \cdot W(\omega + \omega_{2,k}) e^{-j\varphi_{2,k}} \right) , \end{aligned} \quad (3.30)$$

with phase assumed relative to the sliding time window, and the Fourier transform of  $w(n)$  denoted by  $W(\omega)$ . Equation (3.30) shows that the Fourier transform of the composite signal  $x_1(n) + x_2(n)$  contains shifted versions of the window transform,  $W(\omega)$ , scaled by complex factors. Denoting the discrete-time Fourier transforms  $x_1 \Leftrightarrow \Gamma_1(\omega)$  and  $x_2 \Leftrightarrow \Gamma_2(\omega)$ , the composite signal spectrum is  $\Gamma(\omega) = \Gamma_1(\omega) + \Gamma_2(\omega)$ . We know only the composite spectrum,  $\Gamma$ , and we need to estimate  $\Gamma_1$  and  $\Gamma_2$  in order to separate the two voices.

At a given frequency, the Fourier transform  $\Gamma(\omega)$  contains contributions from the  $2(J_1+J_2)$  shifted and scaled window transforms. If we assume that the only significant contributions to  $\Gamma(\omega)$  at a given frequency are due to the nearest two shifted window transforms, a simple set of linear equations may be used to solve for the overlapped parameters. Consider two partials having frequencies  $\omega_1$  and  $\omega_2$  separated by a small frequency difference  $\omega_1 - \omega_2$ . With

the spectrum of the combined signals denoted by  $\Gamma(\omega)$  and given a normalized window transform  $W(\omega)$  such that  $W(0)=1$ , we have

$$\begin{aligned}\Gamma(\omega_1) &= \Gamma_1(\omega_1) + W(\omega_1 - \omega_2) \cdot \Gamma(\omega_2) , \\ \Gamma(\omega_2) &= \Gamma_2(\omega_2) + W(\omega_1 - \omega_2) \cdot \Gamma(\omega_1) .\end{aligned}\tag{3.31}$$

From these two equations we can solve for the unknown complex quantities  $\Gamma_1(\omega_1)$  and  $\Gamma_2(\omega_2)$ . The equations in (3.31) are complex, but the real and imaginary parts may be separately computed. Thus, we can obtain estimates of the amplitude and phase of any pair of partials with frequency spacing less than the resolution bandwidth of the window transform. A schematic representation of the separation process is depicted in Figure 3.7.

It should be noted that a similar linear equation method (for co-channel speech) was proposed independently by Danisewicz and Quatieri (1988). Their method includes the effects of all shifted window transforms, not just the nearest two, as in this investigation. Danisewicz and Quatieri also show an interpretation of the frequency-domain linear equation solution in terms of an equivalent time-domain least-squares viewpoint.

Unfortunately, when the frequency difference between two partials approaches zero (near-perfect coincidence), the set of equations in (3.31) becomes singular, and no unique solution can be found. One of the alternate separation strategies must be applied when this occurs.

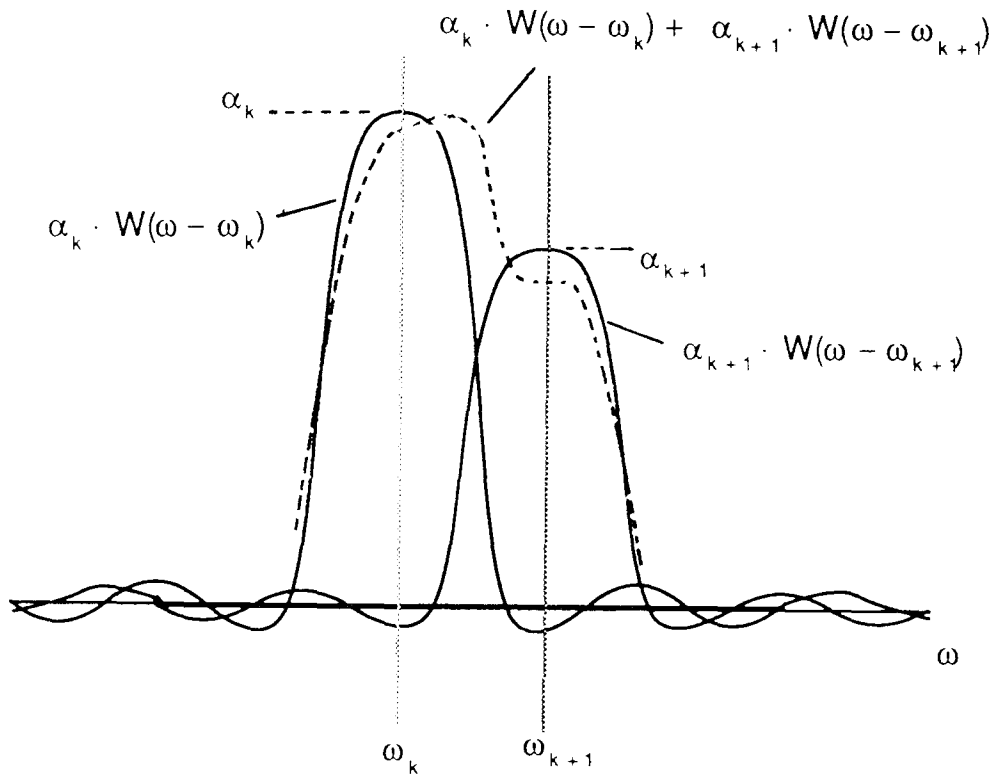


Figure 3.7: Overlap Response for Two Closely Spaced Partials (Real Part).

### 3.7.3 Separation strategy II: analysis of beating components

When a spectral collision occurs, several characteristic effects may be observed in the MQ analysis output. Consider the case in which two cosinusoids are present:

$$A_1 \cdot \cos(\omega_1 t) + A_2 \cdot \cos(\omega_2 t) = \operatorname{Re} \left\{ A_1 e^{-j\omega_1 t} + A_2 e^{-j\omega_2 t} \right\}. \quad (3.32)$$

Defining  $\omega_d = (\omega_1 - \omega_2)/2$  and  $\omega_a = (\omega_1 + \omega_2)/2$ , (3.32) becomes

$$\begin{aligned} &= \operatorname{Re} \left\{ e^{-j\omega_a t} \cdot [A_1 \cdot e^{-j\omega_d t} + A_2 \cdot e^{+j\omega_d t}] \right\} \\ &= \operatorname{Re} \left\{ e^{-j\omega_a t} \cdot [(A_1 + A_2) \cdot \cos(\omega_d t) - j(A_1 - A_2) \cdot \sin(\omega_d t)] \right\}, \end{aligned} \quad (3.33)$$

which may be expressed in polar form as

$$\begin{aligned} &= \operatorname{Re} \left\{ e^{-j\omega_a t} \cdot \left[ \sqrt{(A_1^2 + A_2^2 + 2A_1 A_2 \cos(2\omega_d t))} \right. \right. \\ &\quad \left. \left. \cdot e^{-j \arctan(\tan(\omega_d t) \cdot (A_1 - A_2) / (A_1 + A_2))} \right] \right\} \\ &= \sqrt{(A_1^2 + A_2^2 + 2A_1 A_2 \cos(2\omega_d t))} \\ &\quad \cdot \cos(\omega_a t + \arctan[\tan(\omega_d t) \cdot (A_1 - A_2) / (A_1 + A_2)]). \end{aligned} \quad (3.34)$$

Equation (3.34) shows that the sum of sinusoids in Equation (3.32) may be expressed as an amplitude and phase modulated cosine. For example, in the special case where  $A_1 = A_2$  and  $\omega_d \ll \omega_a$ , the composite signal takes the form of a balanced AM signal, ~100% modulated, with carrier frequency  $\omega_a$ . This effect is commonly described as beating, where the beat rate is equal to the difference between the component frequencies,  $|\omega_1 - \omega_2|$ . Several examples are shown in Figure 3.8.

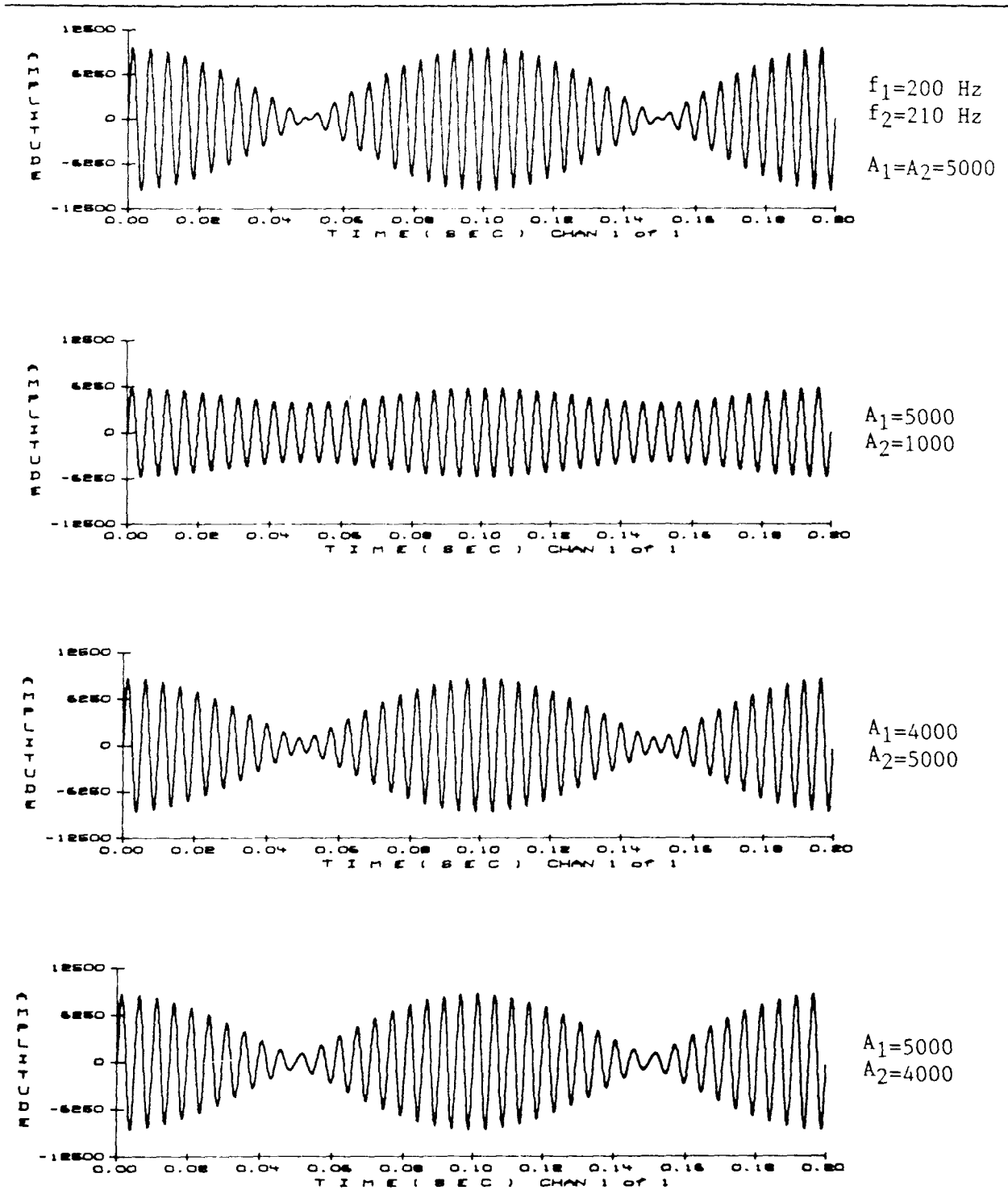


Figure 3.8: Examples of  $A_1 \cos(\omega_1 t) + A_2 \cos(\omega_2 t)$

The instantaneous frequency of the signal in (3.34) can be defined as the time derivative of the cosine phase,

$$\hat{\omega}(t) = \frac{d}{dt} \left( \omega_a t + \arctan \left[ \tan(\omega_d t) \cdot \frac{(A_1 - A_2)}{(A_1 + A_2)} \right] \right) \quad (3.35)$$

which can be expressed in the form\*

$$\hat{\omega}(t) = \omega_a + \frac{(A_1^2 - A_2^2) \cdot \omega_d}{A_1^2 + A_2^2 + 2A_1 A_2 \cdot \cos(2\omega_d t)} \quad (3.36)$$

For  $A_1 = A_2$  the second term becomes zero. For  $A_1 \approx A_2$  the second term remains near zero, except when the cosine in the denominator is close to -1; then the denominator becomes very small, giving a large positive pulse if  $A_1 > A_2$  or a large negative pulse if  $A_1 < A_2$ . The frequency pulse occurs when the amplitude function is nearly zero, so the perceptual importance of this effect is minimal. For  $A_1 \gg A_2$  the second term oscillates about the value  $\omega_d$ , with the oscillations due to the cosine term in the denominator becoming less significant as the ratio  $A_1/A_2$  grows. For  $A_1 \ll A_2$  the behavior is similar, but the second term approaches  $-\omega_d$  as the ratio  $A_2/A_1$  increases. Several examples corresponding to the waveforms of Figure 3.8 are shown in Figure 3.9.

---

\*Recall the formulas

$$(d/dt) \arctan u = [ 1/(1+u^2) ] du/dt$$

and

$$(d/dt) \tan u = [ \sec^2 u ] du/dt$$



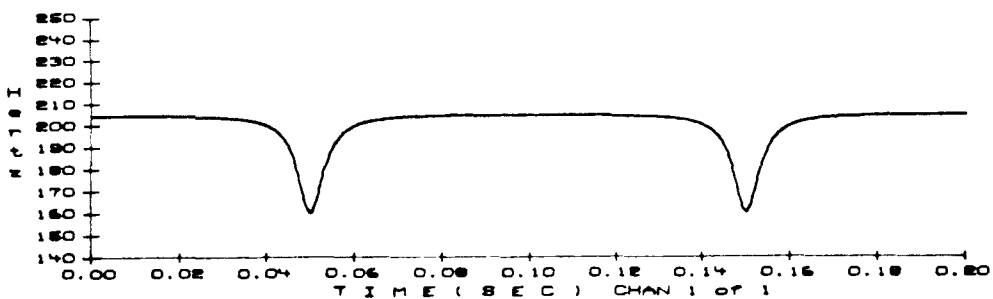
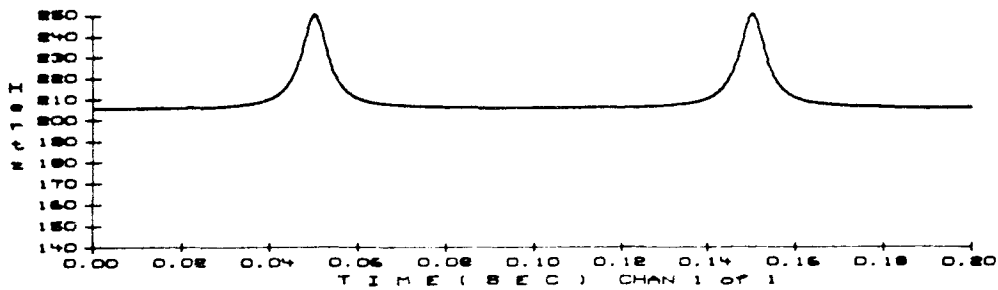
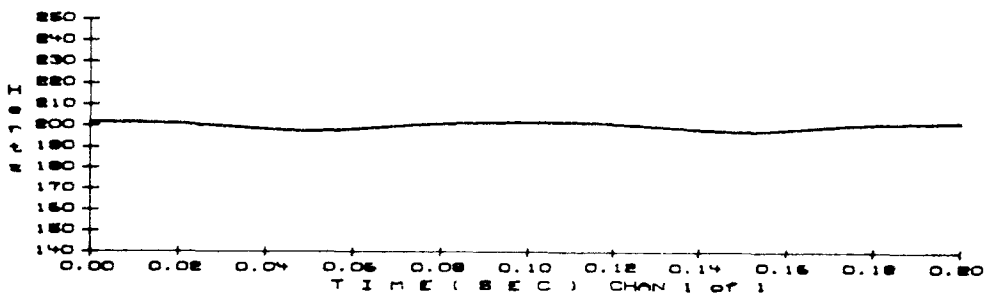
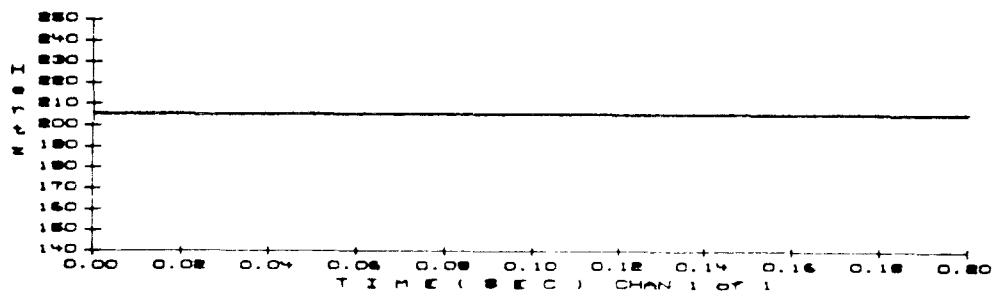


Figure 3.9: Instantaneous Frequency Functions for Waveforms from Figure 3.8.

As discussed above, the presence of two partials with frequency spacing less than the resolution bandwidth of the analysis window transform results in the composite signal of (3.34). Assuming that the colliding partials remain essentially constant for several cycles of the  $|\omega_1 - \omega_2|$  beat frequency, the amplitude values  $A_1$  and  $A_2$  can be determined using the amplitude and frequency beating functions: The maximum of the amplitude beat is  $A_1 + A_2$ , the minimum is  $|A_1 - A_2|$ , and if the amplitude minimum occurs when the frequency is a minimum, the lower frequency partial's amplitude is  $\max(A_1, A_2)$ , while if the amplitude maximum occurs when the frequency is a minimum, the lower frequency partial's amplitude is  $\min(A_1, A_2)$ .

For example, consider the pair of partials with known frequencies  $\{ 1000, 1005 \}$  Hz and unknown amplitudes. This gives  $\omega_d = |\omega_1 - \omega_2|/2 = 2.5$  Hz, and  $\omega_a = (\omega_1 + \omega_2)/2 = 1002.5$  Hz, for use in Equations (3.34) and (3.36). If we measure a maximum amplitude of 250 and a minimum amplitude of 50 due to beating, we can solve for the two amplitudes: namely,  $(250 + 50)/2 = 150$ , and  $(250 - 50)/2 = 100$ . However, we can not tell which amplitude goes with which frequency from the amplitude beating alone. If the minimum of the amplitude beat coincides with the minimum of the instantaneous frequency, we know that the lower partial (1000 Hz) has the larger amplitude (150), and the higher partial (1005 Hz) has the smaller amplitude (100).

Determination of the phase of the colliding partials using beating analysis is not possible, because only the relative phase between the two components is known, not the absolute phase relative to the time origin. Fortunately, the phase value for a partial on a given frame can often be estimated by adding the phase from the previously processed frame to its

frequency value multiplied by the spacing between frames. In the current implementation this process of phase accumulation seldom causes audible degradation.

Unfortunately, successful use of component beating functions to solve the partial collision problem requires that the duet voices contain no significant amplitude and frequency fluctuations of their own. Some instrumental and vocal timbres obey this restriction, but most, in general, do not. Thus, some means for resolving partial collisions is necessary when methods I and II fail.

#### 3.7.4 Separation strategy III: signal models, interpolation, or templates

When partial collisions cannot be resolved by either of the separation strategies described above, the interfering partials must be reconstructed by some other means. Specifically, if a flexible model can be determined for each voice of the duet, the missing partials can be generated artificially.

Models of musical sounds can range from differential equation specifications for a musical instrument based on physical principles to empirically-derived parameters based on time-variant analysis. A primary difficulty with modeling musical signals lies in matching the tonal quality--the timbre--of a particular instrument under conditions of deliberate and accidental performance variations typical of most music. For example, a simple model (such as an FM model) able to produce a believable synthetic trumpet sound may not be amenable for predicting appropriate behavior of a trumpet for every possible combination of pitch, embouchure, vibrato, etc.,

nor may it be able to choose the "correct" performance mode for a given musical context. Certainly, a single, static spectrum is insufficient to capture the essence of a complex instrument. On the other hand, an empirical model would require a large data base of salient parameters, including rules for their appropriate application, in order to repair corrupted partials in a duet recording.

However, in the case where only a small number of partials of each voice collide, we may assume that the parameters of the interfering partials can be estimated (interpolated) from the remaining uncorrupted partials. As an example, consider voices with fundamental frequencies 100 Hz and 175 Hz:

voice 1 partials:	100	200	300	400	500	600	700	800	900	...	
voice 2 partials:		175		350		525		700		875	...

The seventh partial of voice 1 and the fourth partial of voice 2 coincide at 700 Hz, while most of the other partials are spaced by at least 25 Hz. On a given analysis frame the amplitude of the 700 Hz partial could be estimated from the remaining uncorrupted partials in the short-time spectrum. Several possible situations, leading to different solutions, must be considered.

If the harmonic spacing (i.e., the fundamental frequency) of a voice is known to be small compared with the bandwidth of any features in the instrument's spectral envelope, a curve connecting the spectral peaks corresponding to that voice provides an estimate of the spectral envelope. For a low-frequency voice with smooth spectral resonances a simple linear interpolation of adjacent partials (of that voice) can yield a reasonable

estimate for a corrupted partial, as shown in Figure 3.10 for the spectrum of a male singing voice. Alternatively, a priori knowledge of the gross spectral character of the voice, e.g., a "smooth" spectral envelope, could be used for

---

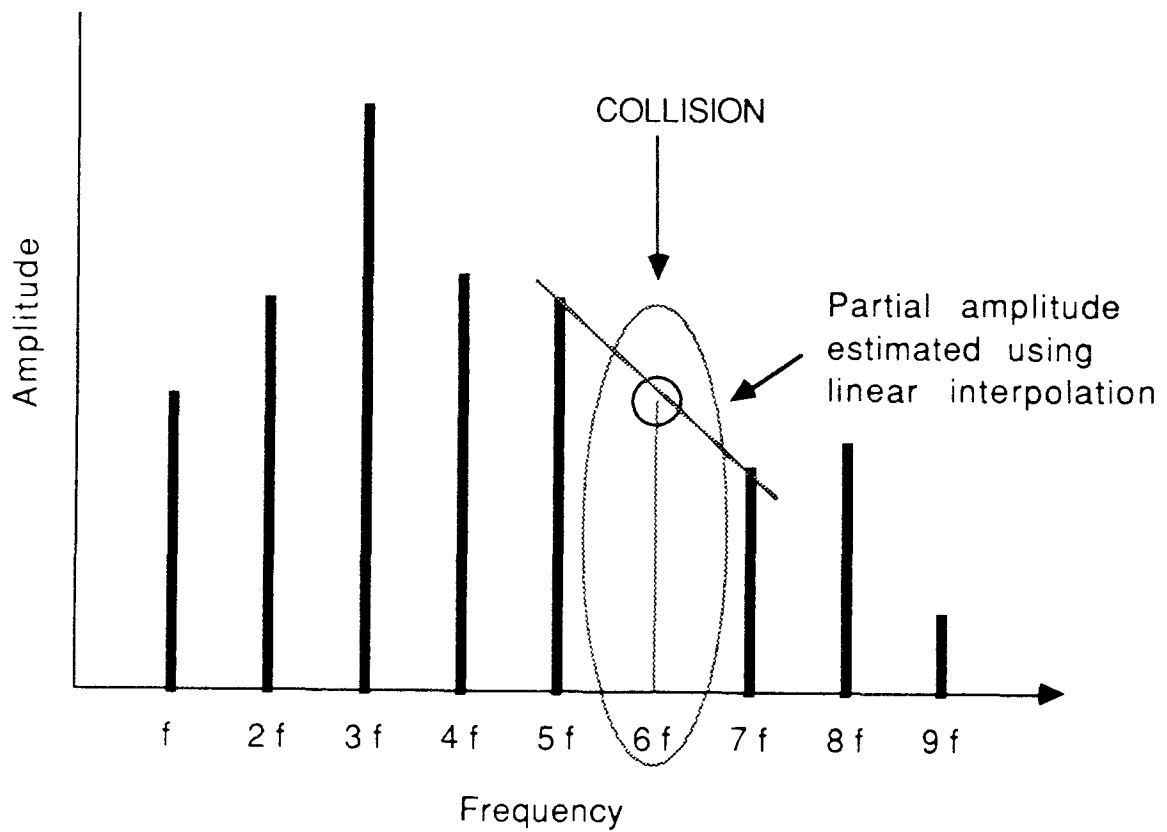


Figure 3.10: Linear Interpolation of Spectrum to Resolve a Collision.

---

reconstruction. A solution of this type may be extended to a higher-order, curve-fitting interpolation procedure based on several surrounding partials, as shown in Figure 3.11.

---

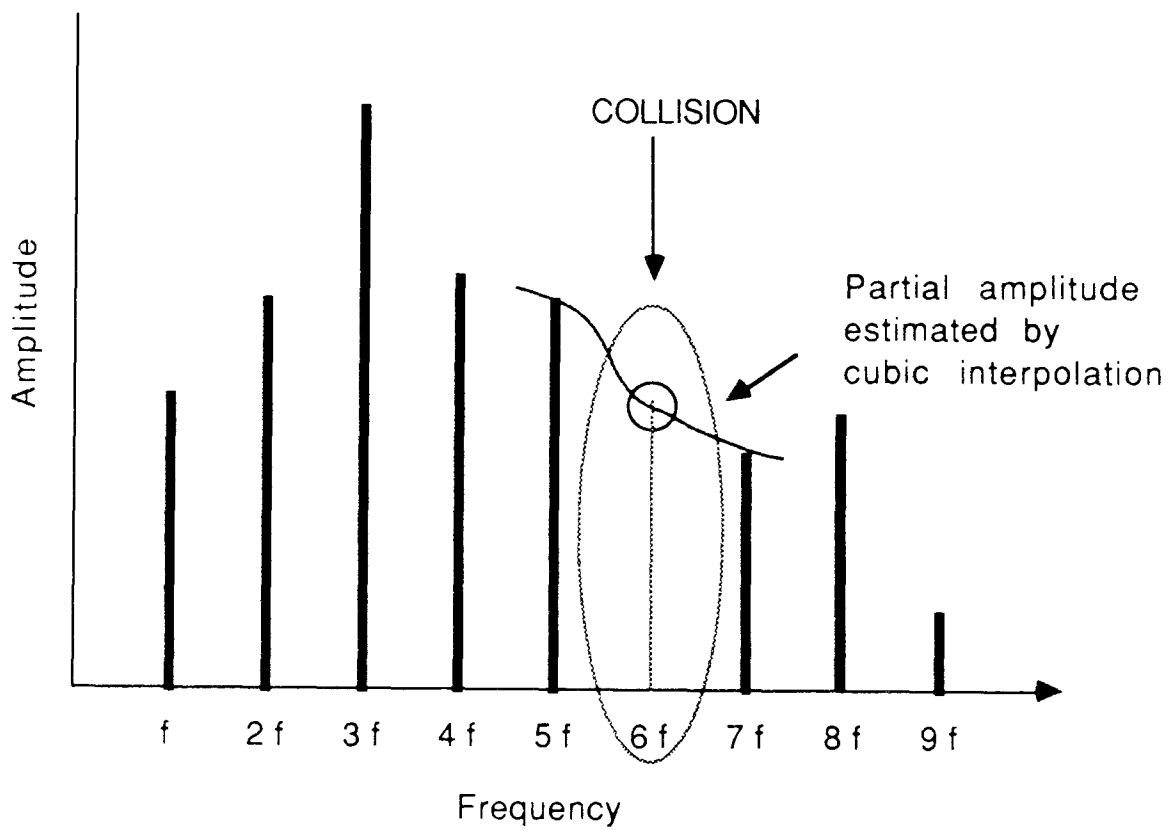


Figure 3.11: Cubic Interpolation of Spectrum to Resolve a Collision.

---

Simple interpolation of the spectrum may not be reasonable if the spacing of partials is large compared to the features of an assumed spectral envelope. In this situation the spectral envelope is undersampled by the spacing of the partials. This is primarily a problem for voices with high fundamental frequencies and/or sounds with many narrow, overlapping resonances, such as sounds produced by the violin. In this case weighted spectral templates may be used to assist in estimating the amplitude of a collided partial. The spectral templates can be precalculated from spectral envelope data for particular instruments for a wide range of amplitudes, fundamental frequencies, durations, etc. Note that the remainder of this section is somewhat speculative because, at present, only simple templates for soprano singers have been developed for this project.

A spectral template is a table giving relative amplitude as a function of frequency. For a particular fundamental frequency the template can be used to look up the estimated relative amplitude weight,  $T_i$ , for each partial number,  $i$ , where partial number  $i=k$  is corrupted by a collision. The measured amplitudes,  $Q_i$ , of the uncorrupted partials are obtained from the short-time spectrum. To minimize the total squared error between the  $Q_i$ 's and the corresponding scaled template values, the  $G \cdot T_i$ 's, the total squared error is computed,

$$E_{\text{total}} = \sum_{\substack{i=1 \\ i \neq k}}^J (Q_i - G \cdot T_i)^2 \quad . \quad (3.37)$$

Taking the derivative with respect to the amplitude scaling parameter,  $G$ , and

setting equal to zero,

$$0 = \sum_{\substack{i=1 \\ i \neq k}}^J 2T_i \cdot (G \cdot T_i - Q_i)$$

or

$$G = \left\{ \sum_{\substack{i=1 \\ i \neq k}}^J Q_i T_i \right\} / \left\{ \sum_{\substack{i=1 \\ i \neq k}}^J T_i^2 \right\} . \quad (3.38)$$

The estimated amplitude of the corrupted partial,  $k$ , is then given by

$$Q_k = G \cdot T_k . \quad (3.39)$$

An example of the template process is shown in Figure 3.12.

The application of templates to replace a partial damaged by a collision follows a few basic steps. For a given analysis frame, a spectral template is chosen from a precalculated set according to the local behavior of the signal. The scaling of the selected template is matched to the uncorrupted partials in a least-squares sense. Then the scale is used to estimate the corrupted partials, as described above.

### 3.7.5 Further considerations

In the current implementation the choice of the appropriate separation strategy for a pair of colliding partials is as follows:



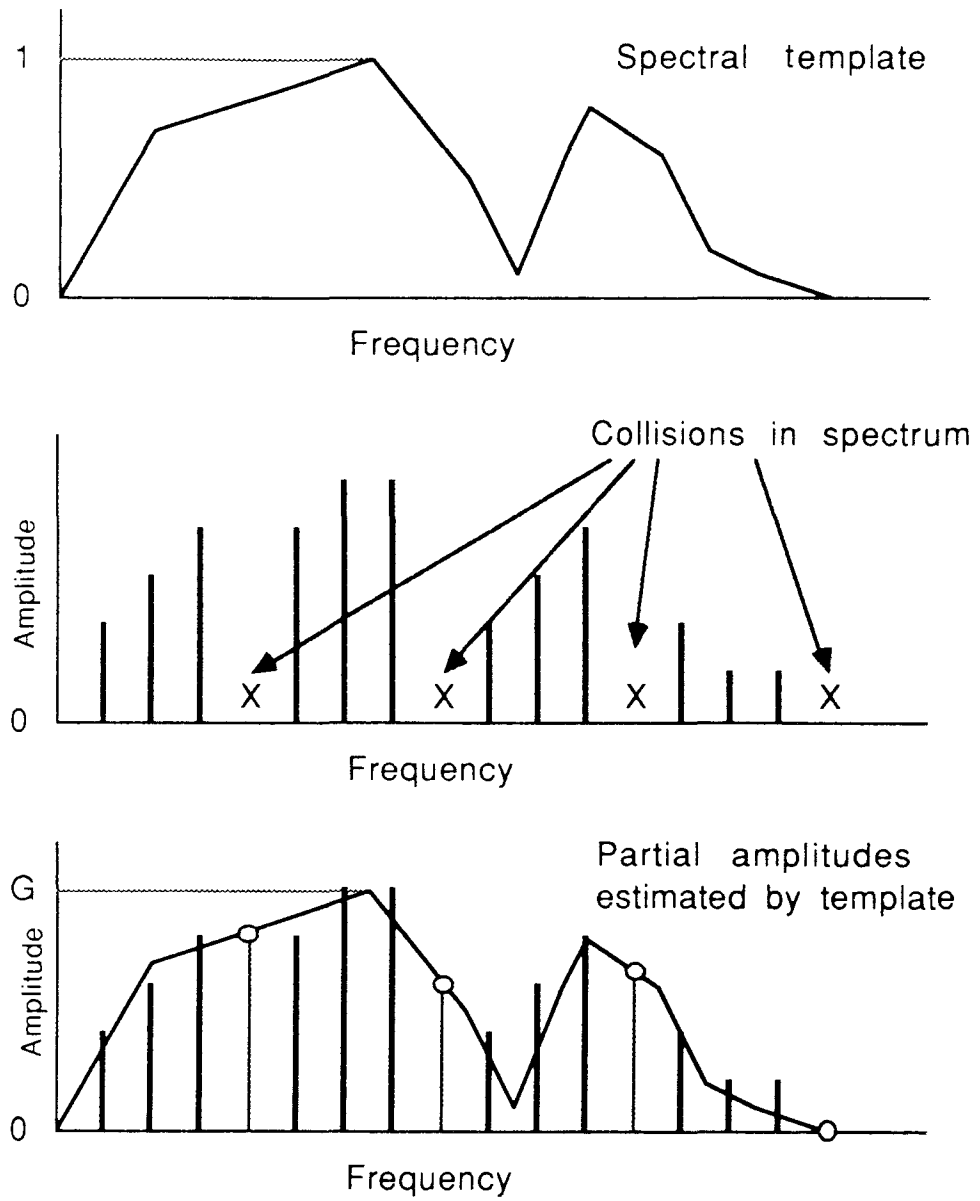


Figure 3.12: The Use of Spectral Templates to Resolve a Collision.

---

- 1) The two fundamental frequencies of the duet (obtained using the two-way mismatch procedure) are used to generate the harmonic series of the two voices. The minimum spacing between adjacent partials is calculated.
- 2) If a partial is at least 50 Hz away from every other partial, the component is considered "clean" and no collision repair occurs.\*
- 3) If two partials are separated by less than 50 Hz but more than 25 Hz, the linear equations solution (strategy I) is applied.
- 4) If two partials are separated by less than 25 Hz, the beating analysis solution (strategy II) is attempted. However, if the collision is less than two or three beat periods ( $< 3/|\omega_1 - \omega_2|$ ), estimates of the beating parameters are not reliable. In this case the spectral interpolation or template solutions (strategy III) are applied.

In the case where the fundamental frequency of one of the voices is an integer multiple (e.g., an octave) of the other, all partials of the higher voice coincide with partials of the lower voice. Extraction of the upper voice may become very difficult when this occurs. The lower voice, however, will have at least some of its partials uncorrupted because the partials of the upper voice will be spaced by at least twice the fundamental frequency of the lower voice. In this situation, an attempt can be made to reconstruct the lower voice spectrum using the separation strategies considered above. The upper voice may be extractable if the spectral envelope of the lower voice's

---

\* This is for a Kaiser window with 6 dB bandwidth of 40 Hz. The criterion would be changed appropriately if the window size is changed.

partials happens to be confined to a frequency range below the partials of the upper voice so that no (or few) collisions occur.

A final consideration is how the separation strategies are applied. Because the duet voices are independent, a note from one voice may start or stop while a note from the other voice is sustained. Thus, the spectral collision situation may change suddenly as the voices enter and exit. If we simply switch the appropriate separation strategy from one method to another as the spectral collisions vary, audible discontinuities may be heard in the output signal due to estimation variations between the separation methods. This problem is solved by additional continuity comparisons between the results of different separation strategies, particularly when a change occurs in the collision status.

CHAPTER 4  
RESULTS AND DISCUSSION

Unlike research work involving speech processing, no standard method is available to measure the "intelligibility" of musical signals. In the duet separation problem, the subjective "quality" of the output depends upon the intended application. For example, an attempt to restore an old 78 rpm record for re-release as a compact disc would probably demand a higher standard of sound quality than an attempt to isolate a particular voice simply to aid manual transcription into standard musical notation. Collection of meaningful psychoacoustic data on aspects of musical fidelity is quite difficult because of the multi-dimensional perceptual issues involved.\* Despite these issues, some method is necessary to evaluate the performance of music processing systems such as the one described in this dissertation.

We also must attempt to balance several external influences: At one extreme, a desire for "perfect" results may tempt us into solving problems with a specific input example in mind, resorting to quick-fixes and special-case program code. The result is often a program that works perfectly for a certain example case, but fails for many ostensibly similar inputs. The opposite extreme occurs when development proceeds at a highly theoretical level, leaving the implementation details as an afterthought. In this case, the theories may be disrupted by the unforeseen vagaries of real input signals, such as noise, level variations, etc. Thus, we must choose a

---

\*One need only consider the colorful arguments among users of consumer audio gear regarding personal opinions of sound quality and fidelity.

sufficient quantity and variety of examples in order to develop useful performance specifications for the separation procedure.

#### 4.1 Testing and Evaluation Outline

The evaluation approach for this project involves both acoustically-generated ("real") signals obtained from musical recordings and artificial signals generated by software. The real signals provide examples of practical problems and solutions, while the synthetic signals define extraction performance limits using known signal parameters.

Several examples are presented in this section to demonstrate the performance of the fundamental frequency tracking and duet separation system:

##### ARTIFICIAL TEST SIGNALS

- 1) The test duet (Figure 4.1) contains one voice with a constant fundamental frequency of 800 Hz for a duration of one second and another voice with a linear fundamental frequency ramp from 1200 Hz to 880 Hz over a duration of one second. The constant frequency voice contains six partials with equal amplitudes, while the changing frequency voice contains six partials with amplitude weightings {1, 0.5, 0.33, 0.25, 0.2 and 0.266}. Both voices have equal peak waveform amplitudes. The signal was chosen to evaluate the collision correction ability of the separation process and the behavior of the frequency tracking procedure for piecewise constant and rapidly varying fundamental frequency pairs.

---

Voice 1: Fundamental frequency: 800 Hz  
6 equal amplitude partials

Voice 2: Fundamental frequency: 1200 to 880 Hz  
6 partials with amplitude weighting:  
1, 0.5, 0.33, 0.25, 0.2, 0.266

Voice 1 and voice 2 have the same peak amplitudes.

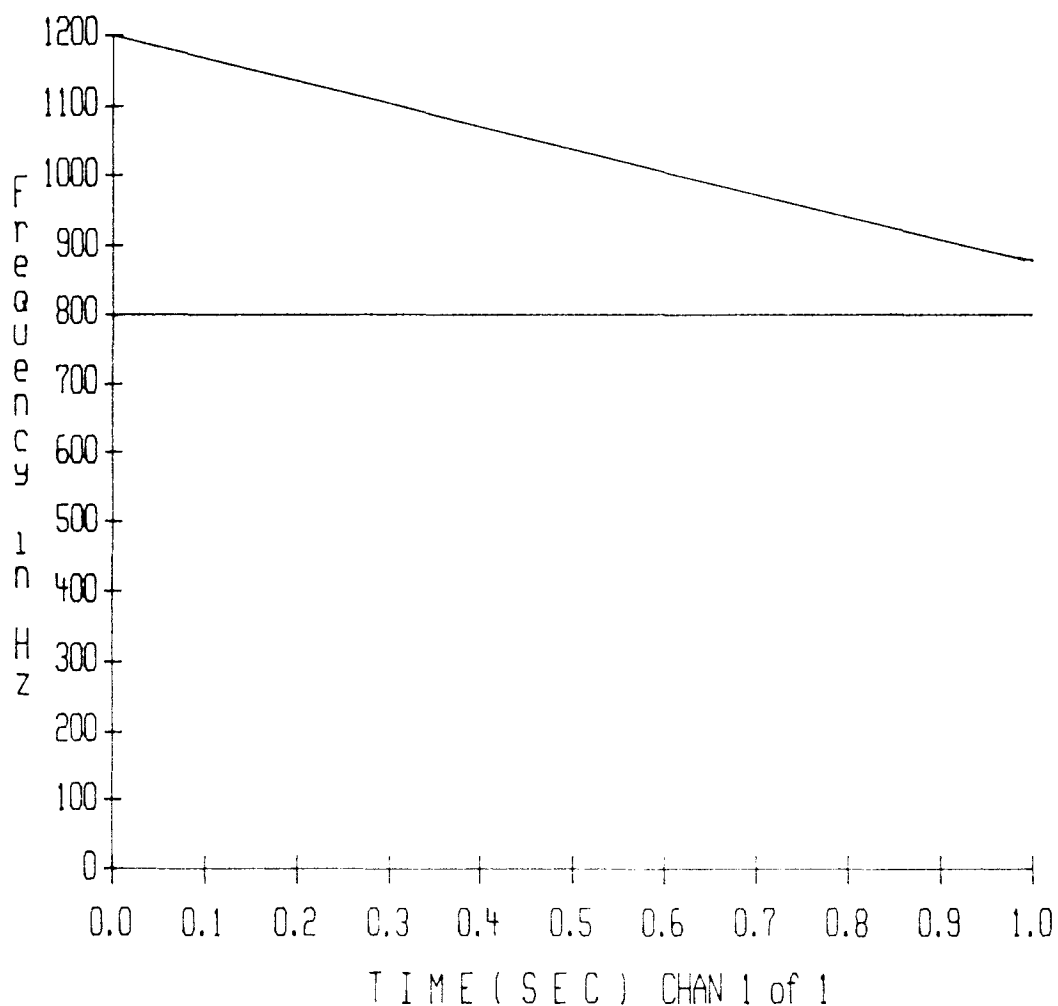


Figure 4.1: Artificial Duet Test Signal #1: Synthesis Frequencies.

---

- 2) This example contains two voices with constant frequencies a minor third apart (pitches C5 and D#5; i.e.,  $f_0 \approx 523$  Hz and  $f_1 \approx 622$  Hz) both generated using phase modulation (see Figure 4.2). This is desirable because phase modulation tones contain partials with distinctive, independent amplitude envelopes. The modulation index for the lower and upper voices are 4 and 5, respectively. Both voices are one second in duration and have equal peak amplitudes. For this combination of pitches, the sixth partial of the lower voice and the fifth partial of the upper voice are close enough in frequency to cause beating (see Figure 4.9).
- 3) This example is a simple duet containing a phase modulation voice and a voice with the same amplitude for all its partials. Note boundaries of the two voices occur at different times to facilitate evaluation of the transition capability of the frequency tracking and separation procedures and their behavior for unconnected (staccato) notes (see Figure 4.3).
- 4) Two fixed-waveform voices with time-varying amplitude envelopes were chosen to determine the behavior of the entire system in the presence of a level mismatch between voices (see Figure 4.4).

#### ACOUSTIC TEST SIGNALS

- 5) The first "real" signal is actually a contrived duet generated by additive mixing of two monophonic recordings of solo female singers. The two original recordings (prior to mixing) are available for comparison with the output of the separation procedure (see Figure 4.5). The first voice in example 5 sings an arpeggio (with vibrato), while the other

---

Voice 1: Fundamental frequency: 523 Hz  
Phase modulation synthesis,  
carrier/modulator ratio = 1:1, index = 4

Voice 2: Fundamental frequency: 622 Hz  
Phase modulation synthesis,  
carrier/modulator ratio = 1:1, index = 5

Voice 1 and voice 2 have the same peak amplitudes.

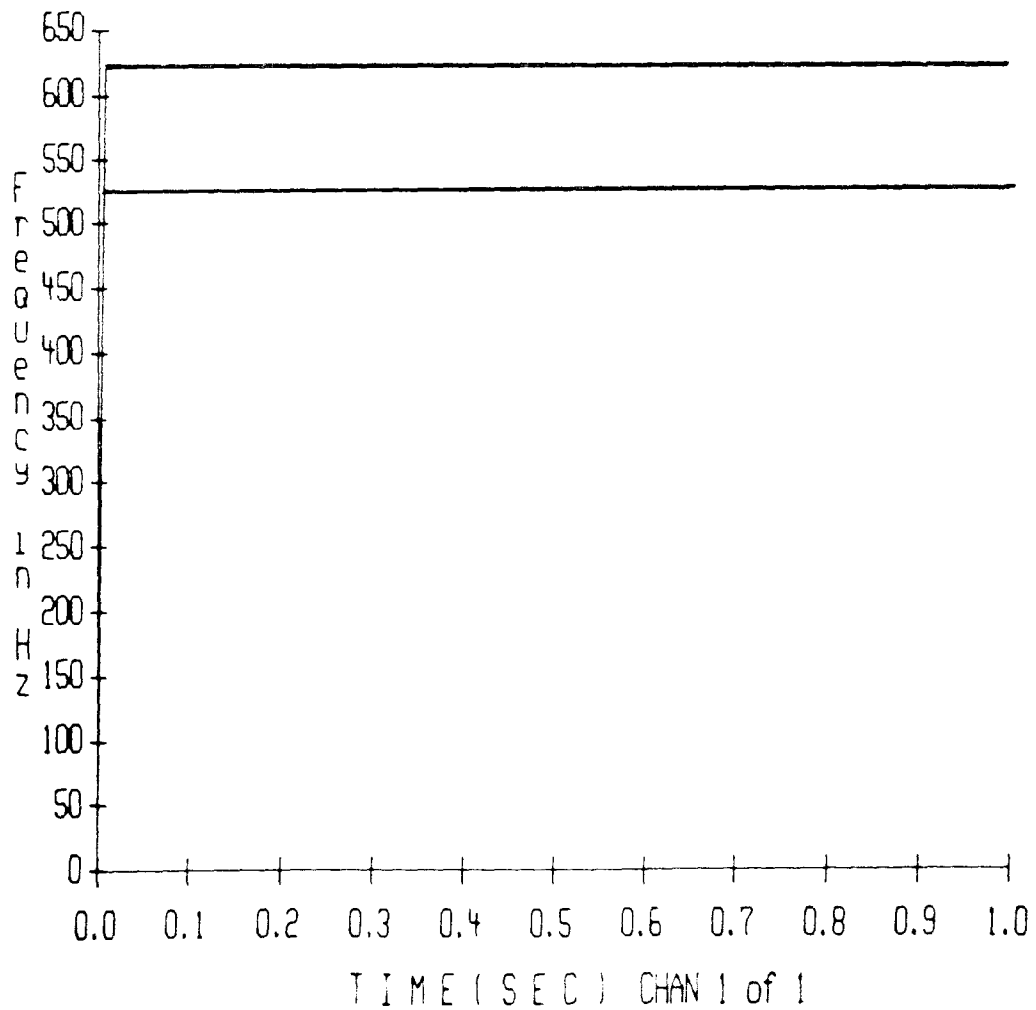


Figure 4.2: Artificial Duet Test Signal #2: Synthesis Frequencies.

---



Voice 2:

Voice 1:

(a)

Voice 1: Phase modulation synthesis,  
 carrier/modulator ratio = 1:1, index = 4  
 Peak amplitude = 15000

Voice 2: 7 equal amplitude partials  
 Peak amplitude = 10000

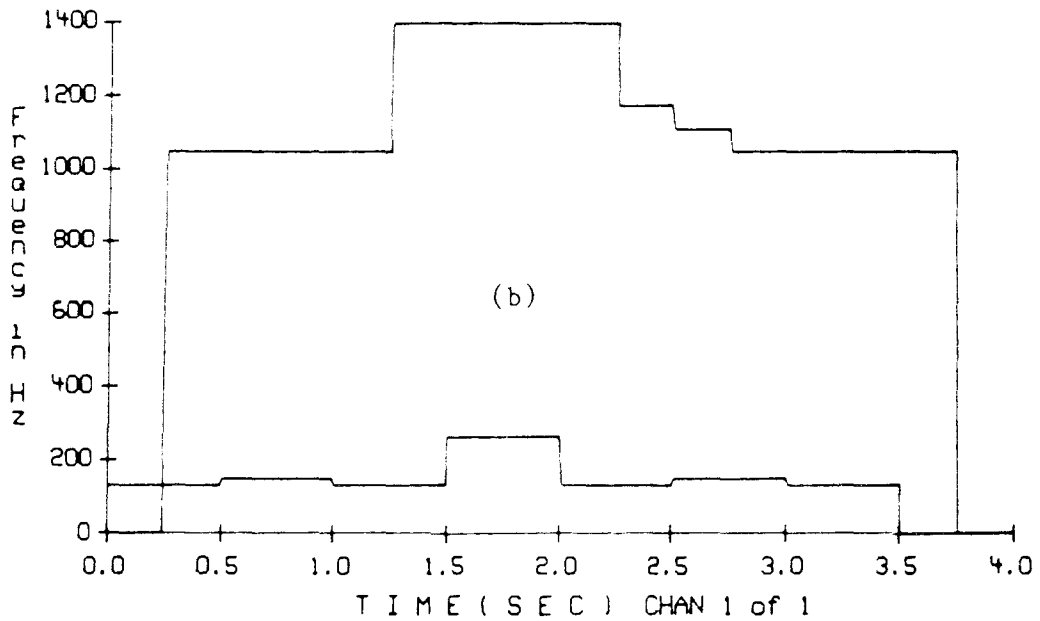


Figure 4.3: Artificial Duet Test Signal #3:  
 (a) Musical Score  
 (b) Frequency Specification

---

Voice 1: Fundamental frequency: 523 Hz  
Phase modulation synthesis,  
carrier/modulator ratio = 1:1, index = 4  
Peak amplitude = 10000

Voice 2: Fundamental frequency: 622 Hz  
7 equal amplitude partials  
Peak amplitude = 20000

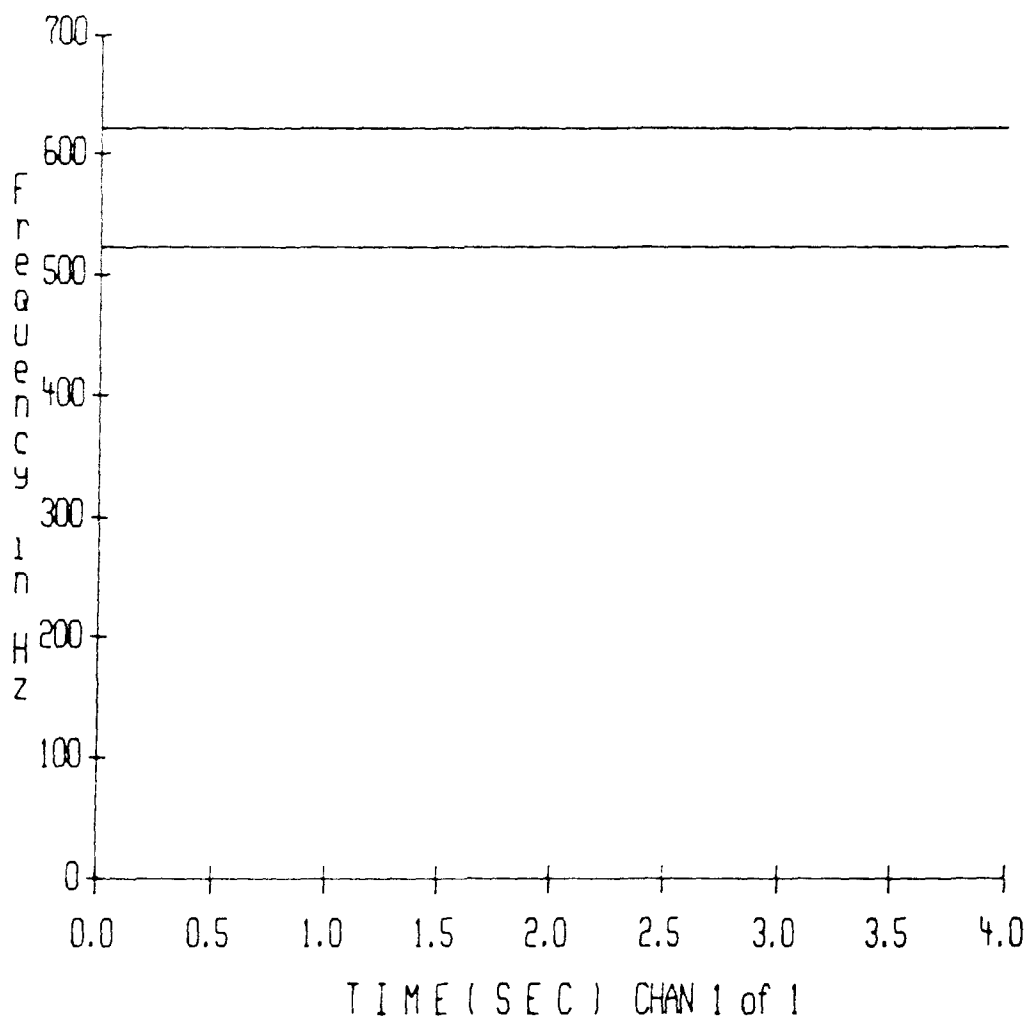
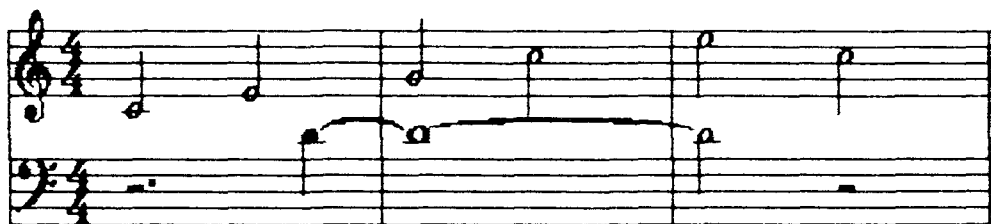


Figure 4.4: Artificial Duet Test Signal #4: Synthesis Frequencies.

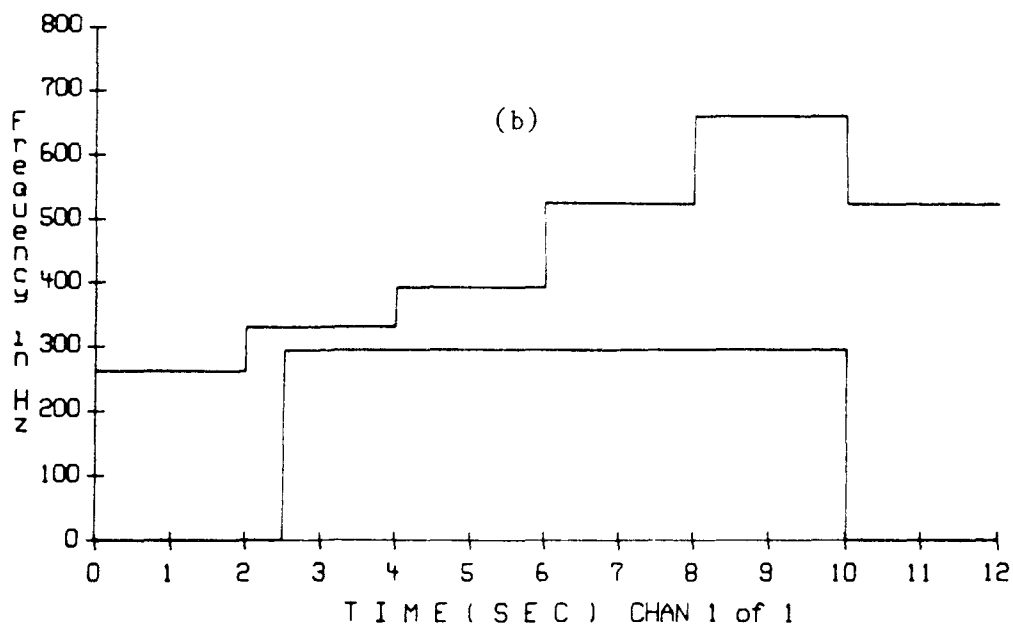
---



(a)

Voice 1: Soprano singing voice,  
arpeggio singing vowel /a/ ("ah"), with vibrato

Voice 2: Alto singing voice, vowel /a/ ("ah"), no vibrato



(b)

Figure 4.5: Duet Test Signal #5:  
(a) Musical Score  
(b) Frequency Specification

---

voice sings a constant pitch without vibrato. The frequency ranges of the two voices overlap, which violates one of the original assumptions about allowable duets. However, the range overlap occurs only for the first note of the upper voice, which is also a solo note. Thus, this test duet is an example requiring manual intervention: The duet frequency tracking process is not started until after the first solo note is over.

- 6) This example is a short segment of Duo #1 for Clarinet and Bassoon by Beethoven, obtained from an analog record album. The example was chosen to test the system in the presence of typical reverberation, surface noise and other distortion (see Figure 4.6).
- 7) A tuba and trumpet duet was chosen to check the tracking and separation process for voices widely separated in frequency. The test segment comes from an analog recording of Sonatina for Trumpet and Tuba by Anthony Iannaccone and contains background noise and reverberation (see Figure 4.7).
- 8) Test duet 8 uses the same musical score as example 7, except the recording was made using live performers in a nonreverberant room. This example is used to compare the system performance for a "clean" signal with its performance for the signal of example 7.

#### 4.2 Evaluation of Duet Fundamental Frequency Tracking

The first fundamental research question posed in the introduction was

---

DUO #1 for CLARINET and BASSOON  
L.V. Beethoven

Clarinet

Bassoon

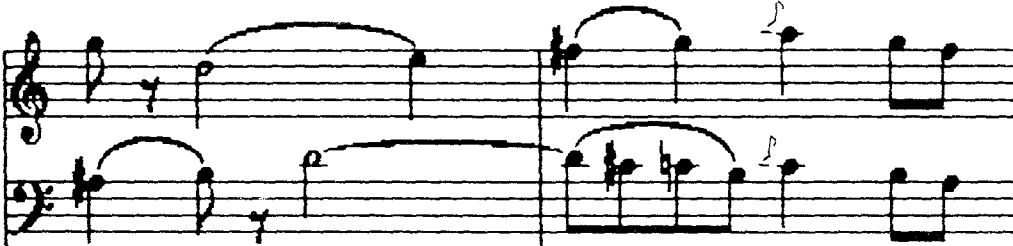
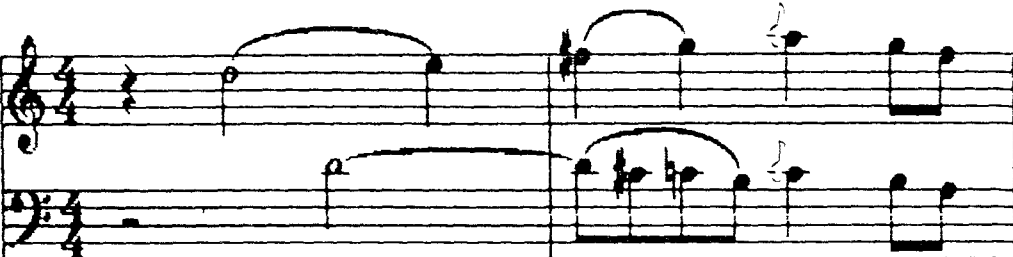


Figure 4.6: Duet Test #6.

---

---

SONATINA for TRUMPET and TUBA  
Anthony Iannaccone

Trumpet

Tuba

The image displays a musical score for a duet between a Trumpet and a Tuba. The score is organized into two systems. The first system consists of two staves: the top staff is for the Trumpet (treble clef) and the bottom staff is for the Tuba (bass clef). The time signature is 4/4. The Trumpet part begins with a whole rest, followed by a quarter note G4, a quarter note A4, and a quarter note B4. The Tuba part begins with a whole rest, followed by a quarter note G2, a quarter note A2, and a quarter note B2. Both parts then play a melodic line with eighth notes: G4-A4-B4-A4-G4 for the Trumpet and G2-A2-B2-A2-G2 for the Tuba. The second system continues this melodic line. The Trumpet part has a quarter note G4, a quarter note A4, a quarter note B4, and a quarter note A4. The Tuba part has a quarter note G2, a quarter note A2, a quarter note B2, and a quarter note A2. The score concludes with a double bar line.

Figure 4.7: Duet Used for Tests #7 and #8.

---

How may we automatically obtain accurate estimates of the time-variant fundamental frequency of each voice from a digital recording of a duet?

In Chapter 3 the two-way mismatch (TWM) duet fundamental frequency-tracking algorithm was proposed as a possible answer to this question. To evaluate this method, the frequency-tracking algorithm was applied to the artificially prepared duets. The qualitative performance of the TWM frequency tracker for the real duet examples is considered later in this chapter.

The MQ analysis and TWM frequency-tracking results for artificial example 1 are shown in Figure 4.8. The partials of the two voices are clearly visible in the MQ output, and the frequency-tracking algorithm has no difficulty following the gross characteristics of the two fundamental frequencies. However, a close examination of the fundamental frequency trace for the upper voice reveals occasional short-term errors. The frequency errors (all less than 1%) occur at points where partial collisions take place, disrupting the harmonic series of the voice. The two-way mismatch algorithm has some immunity to this problem due to its "best match" criterion, but the amplitude and frequency fluctuations inherent during a partial collision still cause some uncertainty.

Figure 4.9 shows the MQ analysis and two-way mismatch frequency-tracking results for artificial example 2. In this example the two phase modulation voices are nearly separable by track segregation alone. Therefore, the TWM frequency tracker results are nearly perfect.

Figure 4.10 shows the MQ analysis and two-way mismatch frequency-tracking results for artificial example 3. The TWM frequency-tracking results match

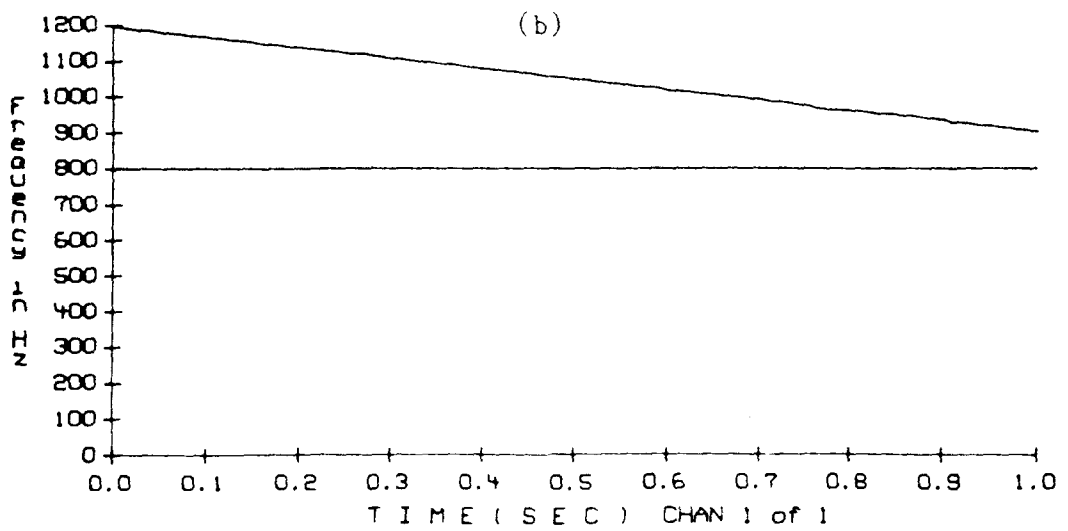
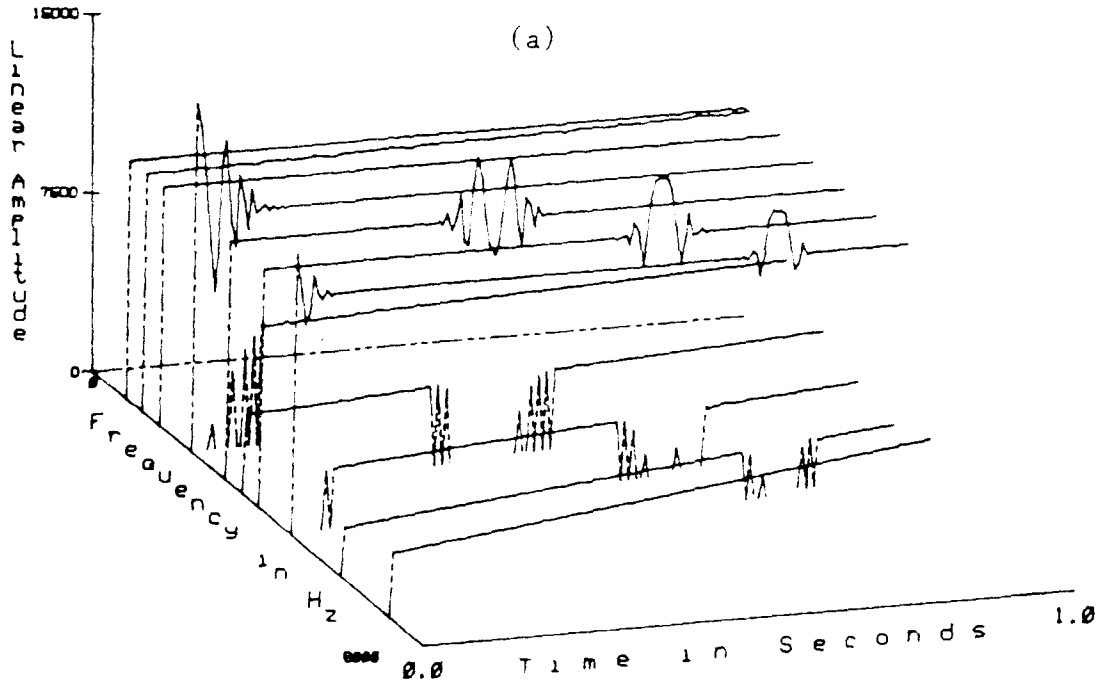


Figure 4.8: (a) MQ Analysis and (b) Two-way Mismatch (TWM) Frequency Tracking of Artificial Duet Test Example #1.



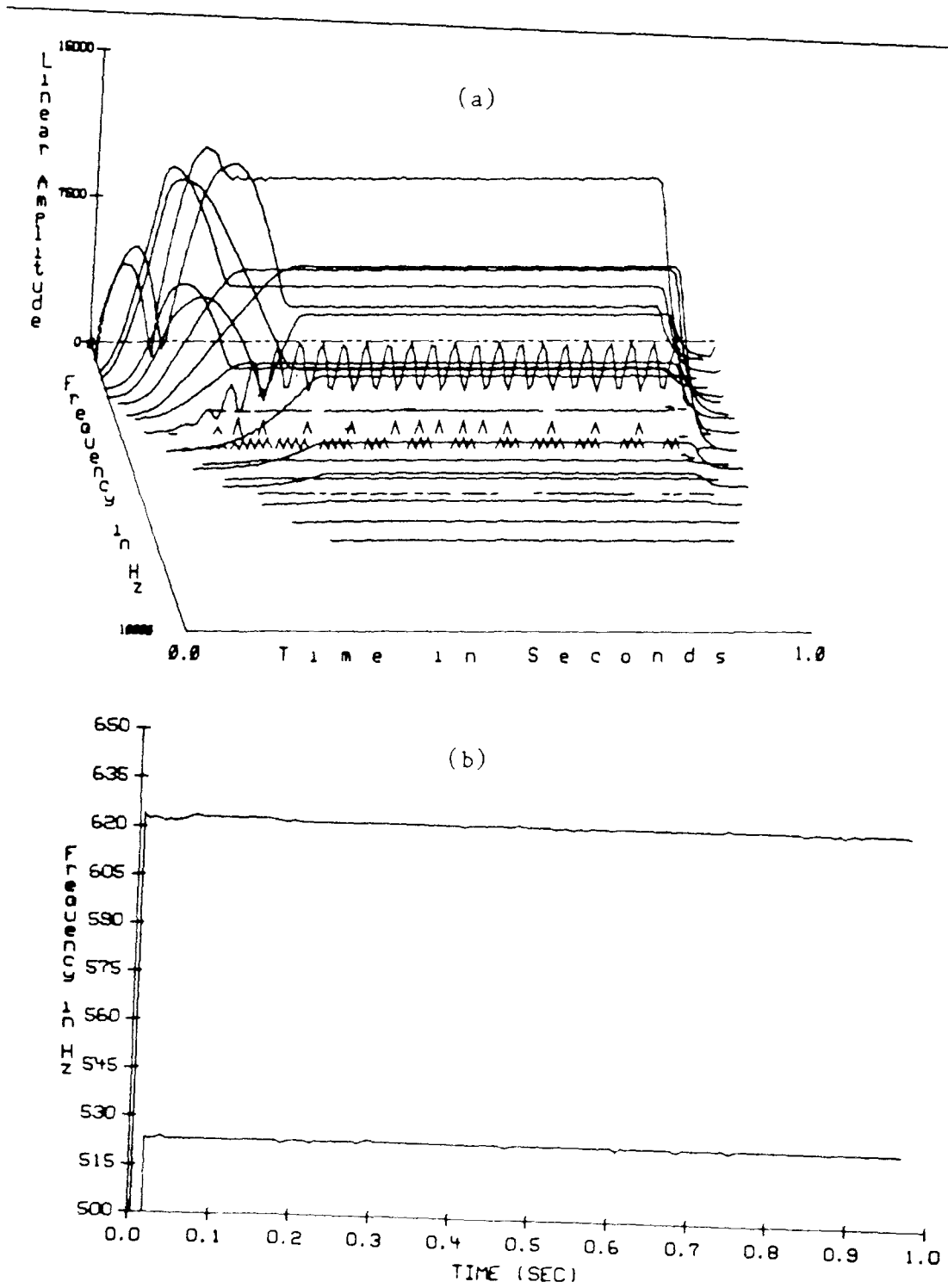


Figure 4.9: (a) MQ Analysis and (b) Two-way Mismatch (TWM) Frequency Tracking of Artificial Duet Test Example #2.

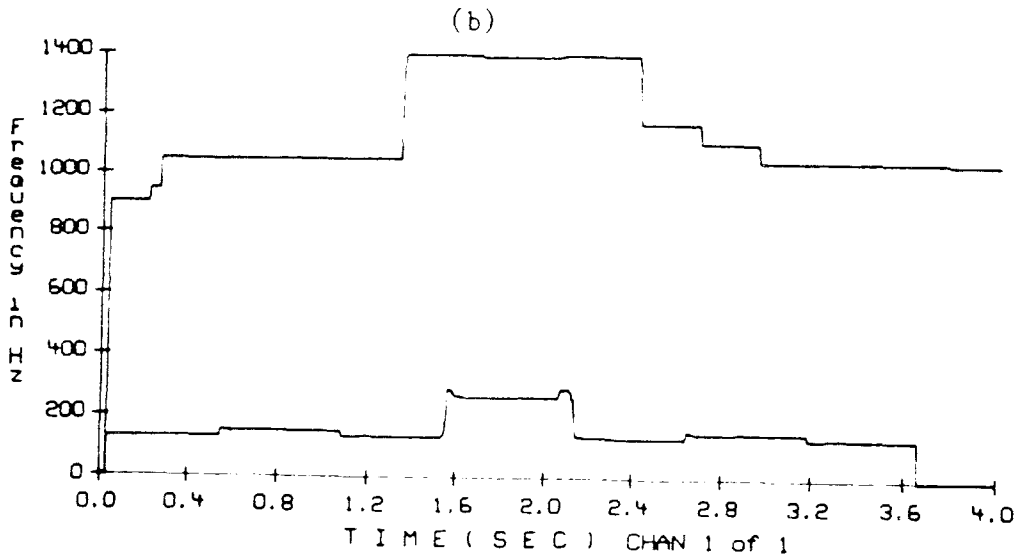
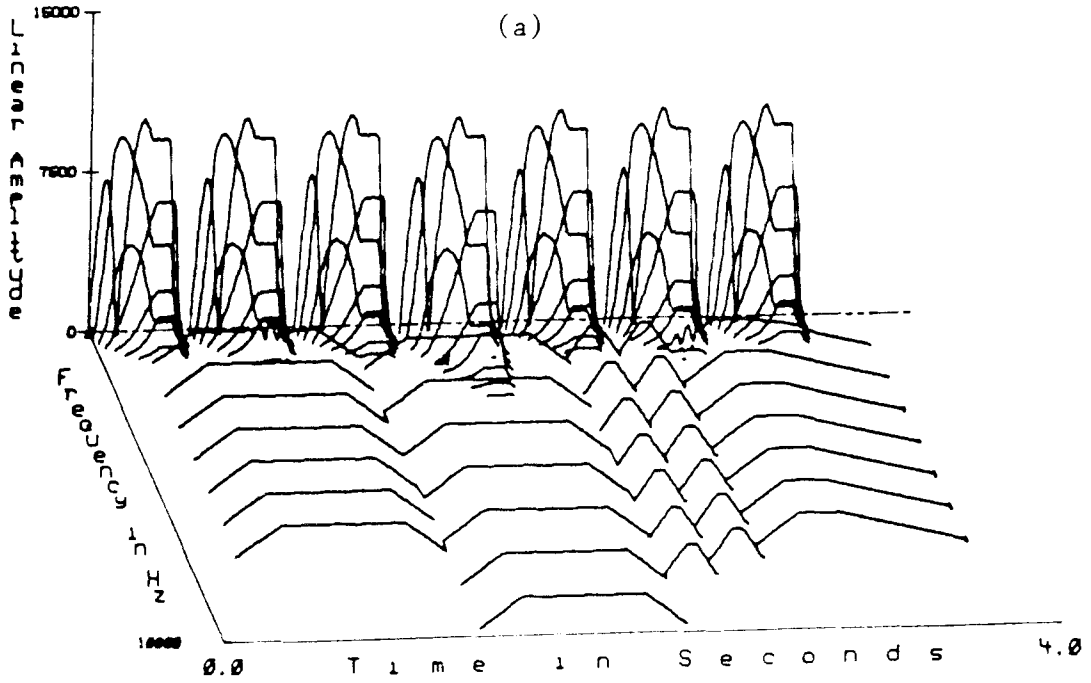


Figure 4.10: (a) MQ Analysis and (b) Two-way Mismatch (TWM) Frequency Tracking of Artificial Duet Test Example #3.

the true values very well during portions of the duet where both voices are present, but the tracker output fluctuates when only one voice is present. This problem is due to an assumption in the TWM algorithm that two sets of harmonic peaks are present in every frame of the MQ analysis data. For solo passages (or between staccato notes) only one set of harmonic peaks are found in the MQ output, so some means is necessary to choose between the normal duet tracking mode and a solo tracking mode.

An attempt to solve the duet/solo problem is included in a more recent implementation of the TWM process. The tracker compares the two-way mismatch error calculated for the "best" pair of fundamental frequencies with the error calculated for the "best" single fundamental frequency. In other words, the mismatch error calculation is performed twice on each frame, first searching for the minimum error for a pair of frequencies, then searching for the minimum error for an individual frequency. The method yielding the best match (smallest error) is considered the appropriate interpretation of the contents of the analysis frame.

Unfortunately, this solution essentially doubles the search calculation required for each frame. Also, if the fundamental of the upper voice coincides with a harmonic of the lower voice, the resulting series of peaks could look like a solo voice. A better solution would be to identify the current voicing in a less brute-force manner, but the current approach has been satisfactory for the purposes of this project.

A difference in level between voices of a duet has several implications for the frequency tracking process. For example, if the level mismatch for a

given pair of nearby partials exceeds 20 dB or so, the weaker partial may be completely obscured by the stronger partial. The frequency tracking process relies on the partial frequency estimates from the MQ analysis, so if many partials of the weaker voice are obscured, the fundamental frequency estimate for that voice may be inaccurate. The MQ analysis and frequency tracking results for the final artificial example, number 4, are shown in Figure 4.11. For this example the TWM process is able to find the frequencies of the two voices, but some fluctuation of the measured frequencies is evident, due to the amplitude mismatch between the two voices. This is despite the fact that the original signal contains only voices with constant frequencies.

The duet of example 5 is the final objective test of the TWM frequency tracker. This duet was generated from two known solo voices, so the fundamental frequency results using the TWM technique for the individual voices (Figure 4.12) can be compared with the result obtained by the duet frequency tracker (Figure 4.13). The TWM process exhibits some difficulty in tracking the two fundamental frequencies for the first few notes of this contrived duet. The principal difficulty is due to the vibrato of the upper voice when the two fundamental frequencies are close together. This is because vibrato causes the partials of the upper voice to sweep back and forth in frequency, producing a complex series of partial collisions which disturb the peak matching process. Unfortunately, no simple solution to this problem is available using the TWM approach. We must rely on the averaging property of the mismatch technique to reduce the sensitivity of the procedure.

In summary, the following conclusions can be drawn from this section on the performance of the two-way mismatch fundamental frequency tracker:

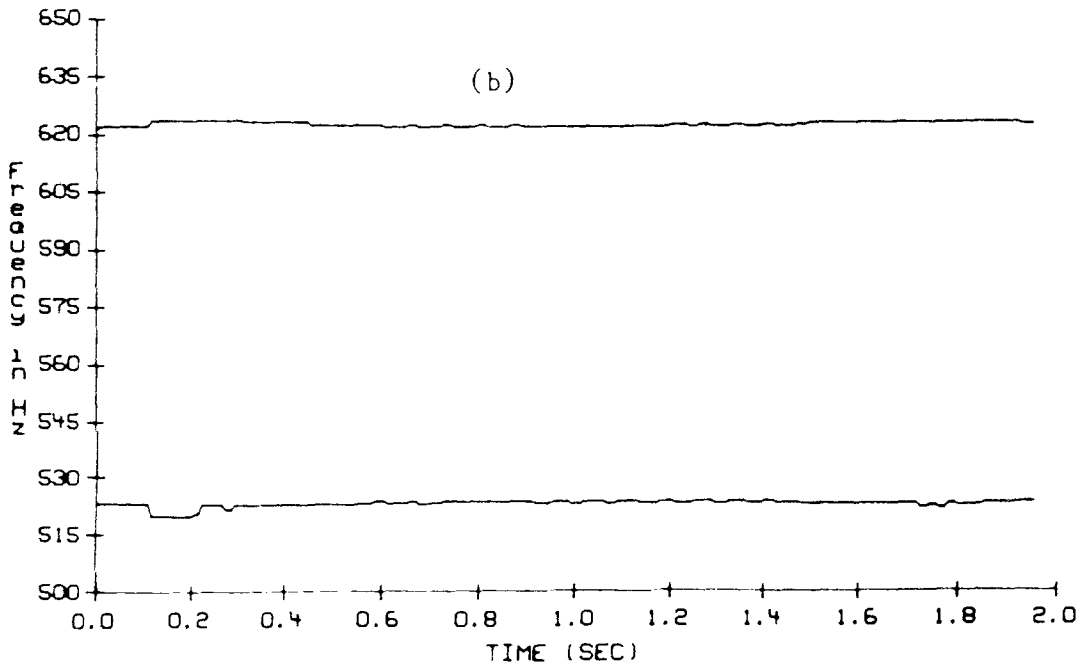
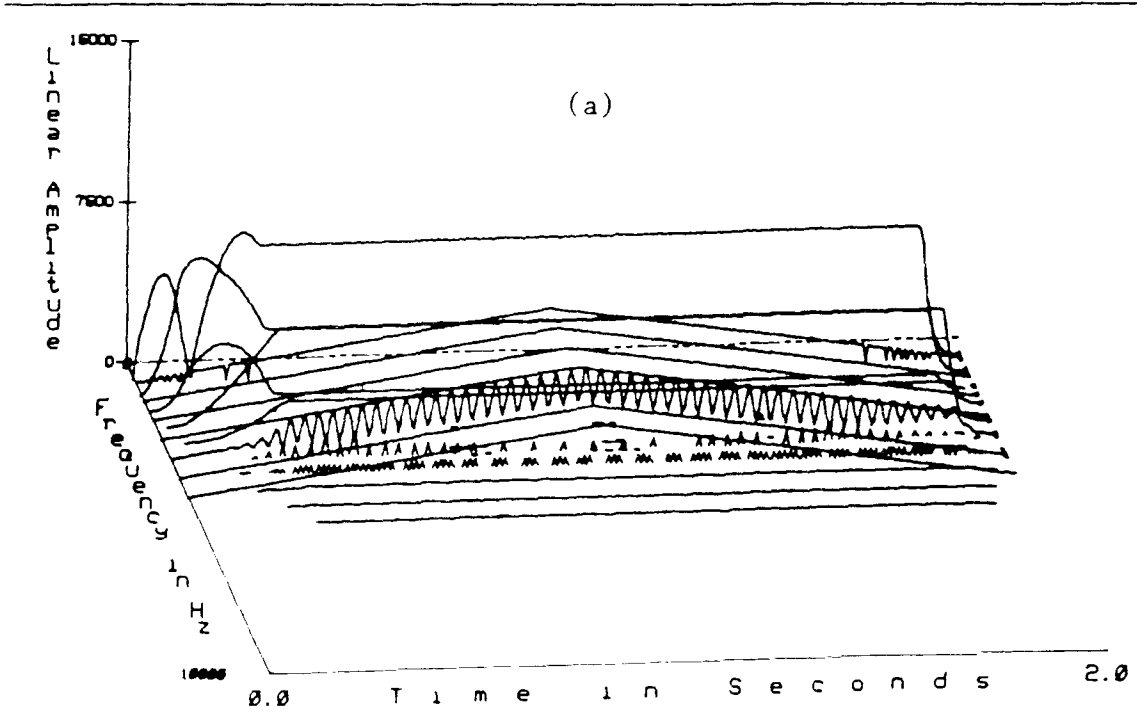


Figure 4.11: (a) MQ Analysis and (b) Two-way Mismatch (TWM) Frequency Tracking of Artificial Duet Test Example #4.

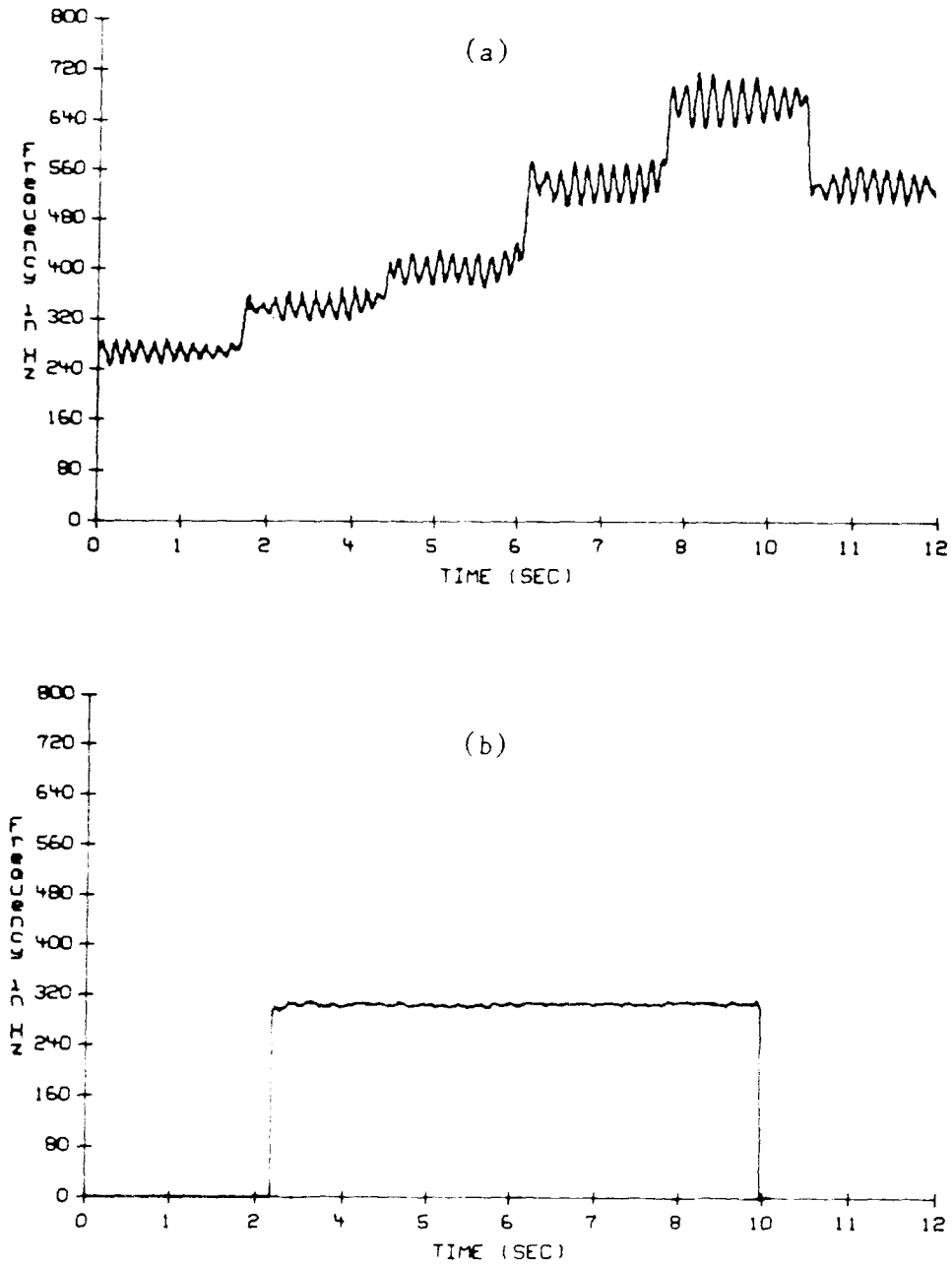


Figure 4.12: Frequency Tracking Data for the INDIVIDUAL Soprano Voices Used in Example #5.  
(a) Arpeggio With Vibrato  
(b) Constant Pitch Without Vibrato

---

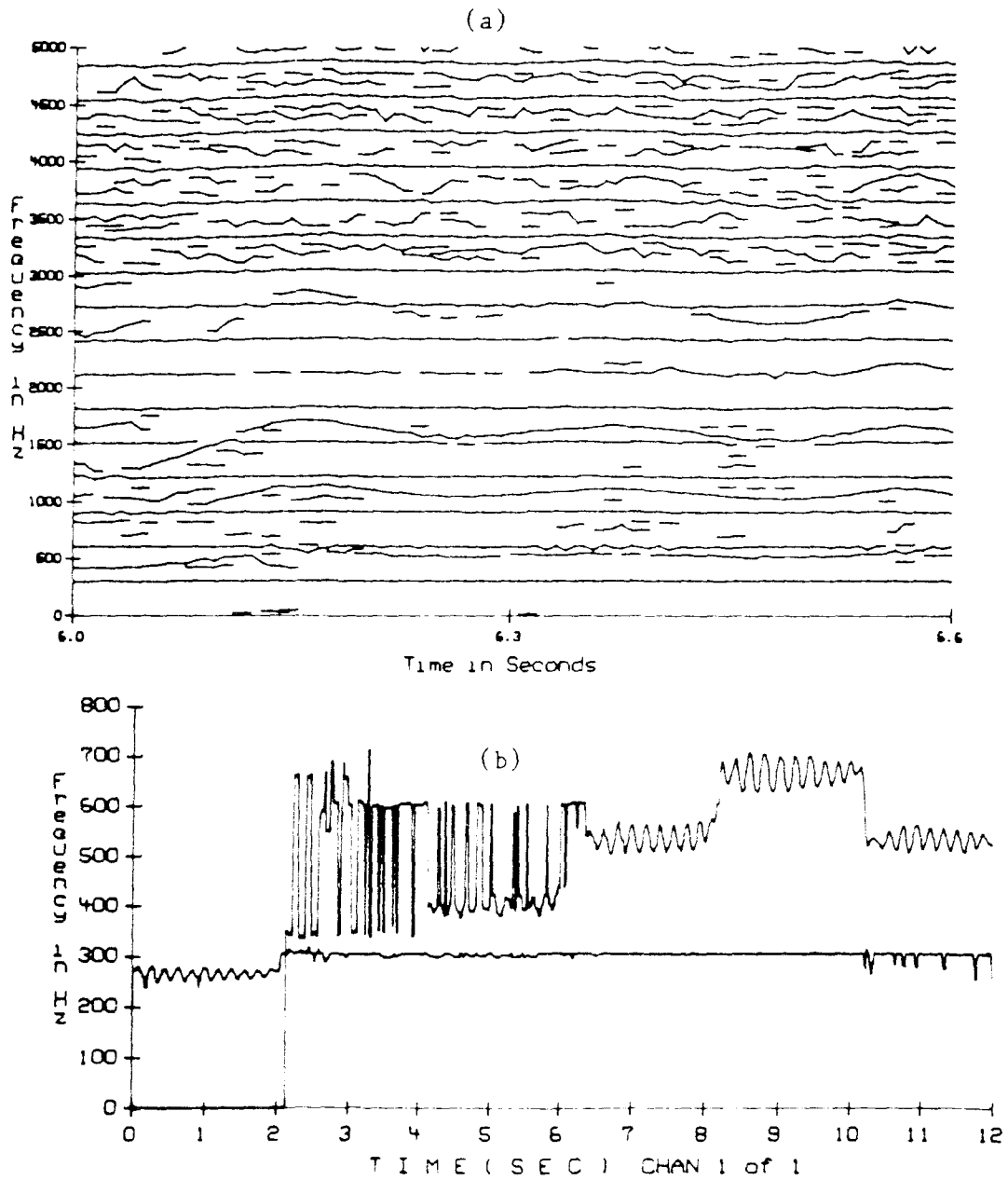


Figure 4.13: (a) MQ Analysis (excerpt) and (b) Two-way Mismatch (TWM) Frequency Tracking of Duet Test Example #5 (Note tracking problem during  $2 < t < 6$ ).

- 1) The TWM procedure performs very well for duet signals with either constant or slowly changing voice frequencies and similar peak amplitudes.
- 2) The tracking process has difficulty interpreting the input signal when staccato notes occur in one or both voices or during other transitions between the duet and solo paradigms.
- 3) The frequency tracking process may be impaired if the list of frequencies supplied by the MQ analyzer is seriously corrupted because of a level mismatch between the two voices or some other degradation.

#### 4.3 Evaluation of Voice Separation with Known Fundamental Frequencies

The second fundamental question posed in the Introduction was

Given time-varying fundamental frequency estimates of each voice in a duet, how may we identify and separate the interfering partials (overtones) of each voice?

The analysis procedure proposed in Chapter 3 is first evaluated using the same artificial duets used to evaluate the TWM pitch tracking procedure. The known fundamental frequencies are supplied to the separation procedure in order to determine its best-case performance, i.e., how well the separation task performs with "perfect" a priori estimates of the fundamental frequencies.

The first artificial test example contains one voice with a constant 800 Hz fundamental, and the other voice with a linear fundamental frequency transition from 1200 Hz to 880 Hz. The separation results using this known frequency information are shown in Figure 4.14. Note that even with perfectly



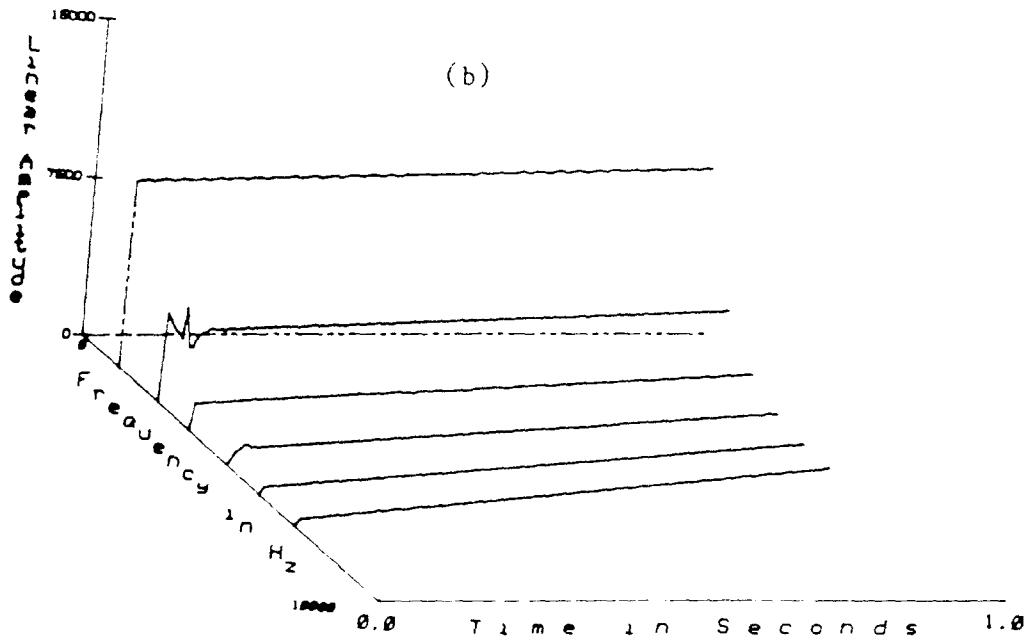
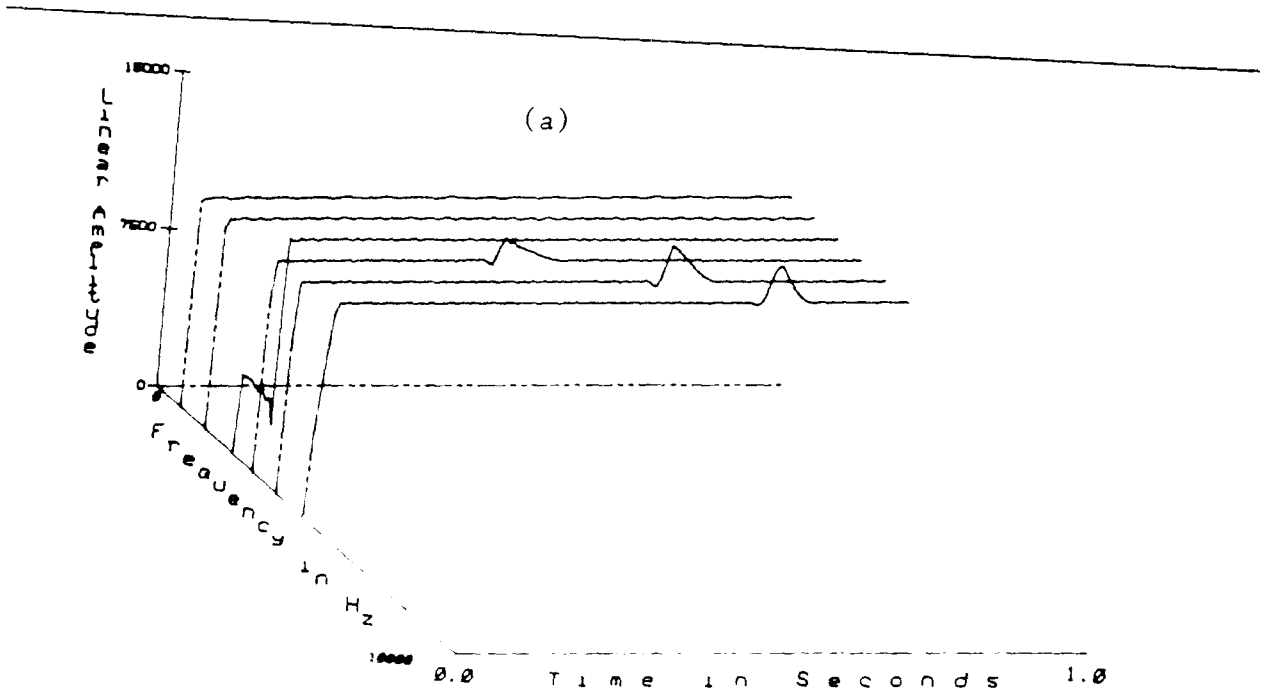


Figure 4.14: Separation Spectra of Example #1  
 Using a priori Frequency Data.  
 (a) voice 1 (b) voice 2

specified fundamental frequencies, the separation process is not perfect. The amplitude discrepancy between the extracted voices of Figure 4.14 and the constant partial amplitudes of the original voices can be traced to one of the underlying assumptions of the separation process: namely, that the peaks in the short-time spectrum are simply shifted and scaled copies of the Fourier transform of the analysis window function. This assumption is exactly correct only if the sinusoidal components comprising the signal do not change their frequency or amplitude during the interval covered by the analysis window. For test example 1 the fundamental frequency changes  $1200 - 880 = 320$  Hz in one second, or 0.32 Hz/msec. Further, the frequency sweep rate for the second partial is 0.64 Hz/msec, for the third partial, 0.96 Hz/msec, etc. The MQ analysis window used in this example is 25.6 msec in duration, yielding a frequency change during the window of 8.192 Hz for the fundamental, 16.384 Hz for the second partial, 24.576 Hz for the third partial, etc. Thus, the assumption of constant frequencies during the window duration is clearly violated in this case. Unless we were to resort to the impractical approach of explicitly predicting the short-time spectrum for every swept-frequency component identified in every input frame, the linear equation solution strategy explained in Chapter 3 is not a truly valid approach for this example.

The main effect of rapidly changing frequencies observed in the short-time spectrum is convolutional broadening of the peak corresponding to the

changing frequency component. The increased frequency extent of the spectral peak increases the likelihood of a collision between the broadened peak and any adjacent components, requiring extra care in the separation process.\*

Unlike the changing frequencies of example 1, test example 2 contains two voices with constant frequencies. As mentioned in the last section, many of the partials in example 2 are spaced adequately to avoid any collision problems. However, partial number six (3139 Hz) of the lower voice and partial number five (3111 Hz) of the upper voice are close enough to produce beating. The separation results are shown in Figure 4.15, and the collision repair is depicted in more detail in Figure 4.16. Note that Figure 4.16 includes the partials' amplitude-vs.-time and frequency-vs.-time projections for clarity. In general, the recovered envelopes are entirely appropriate for the phase modulation process used. However those which were subject to collision suffer from some amplitude and frequency perturbations.

For an arbitrary duet signal, the fundamental frequency of the upper voice may be an integer multiple of the fundamental of the lower voice. This means that every partial of the upper voice is subject to a collision with a lower voice partial. However, the situation may be salvageable if the spectrum of the lower voice happens to contain little energy in the region of the upper voice partials. In this quasi-bandlimited case, the separation strategies of Chapter 3 may still be applicable. Test duet 3, for example, contains several notes where the voices are in octave alignment, but with very

---

\*It should be noted that the underlying issue here is the fundamental time-bandwidth limitation of short-time analysis techniques (Chapter 3).

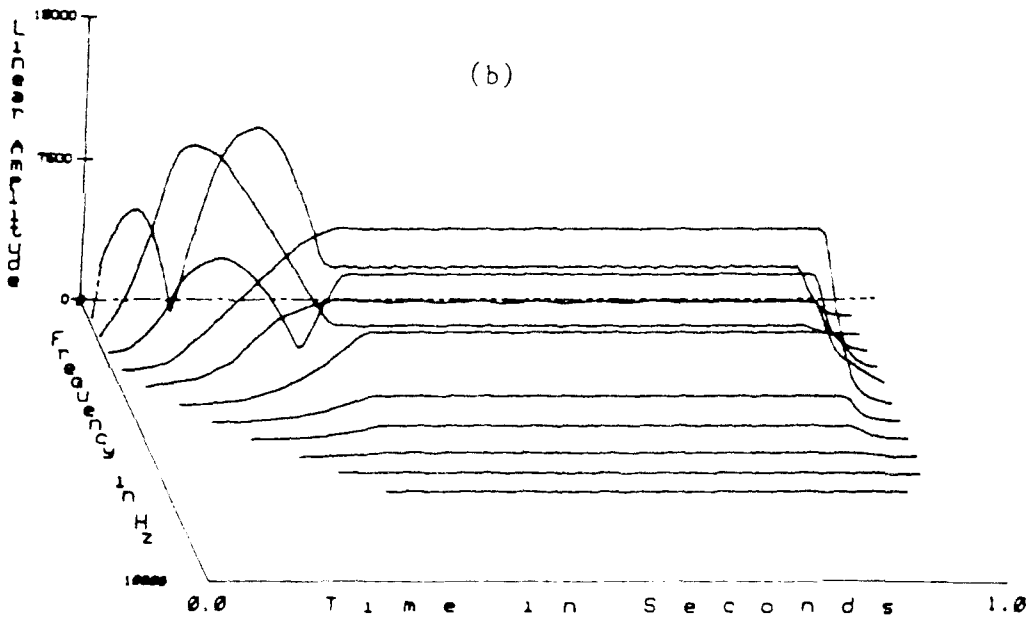
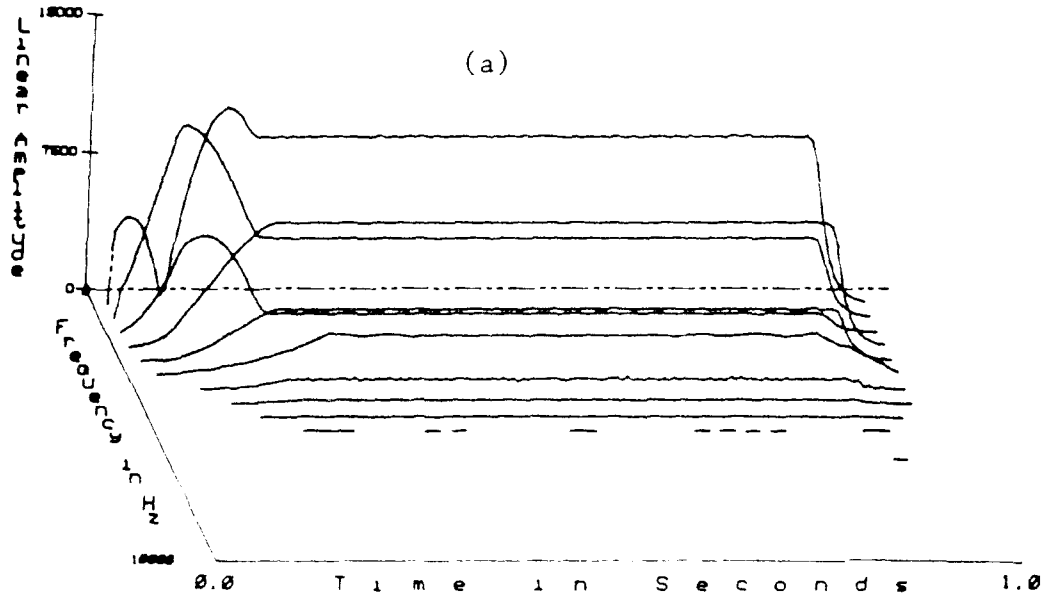
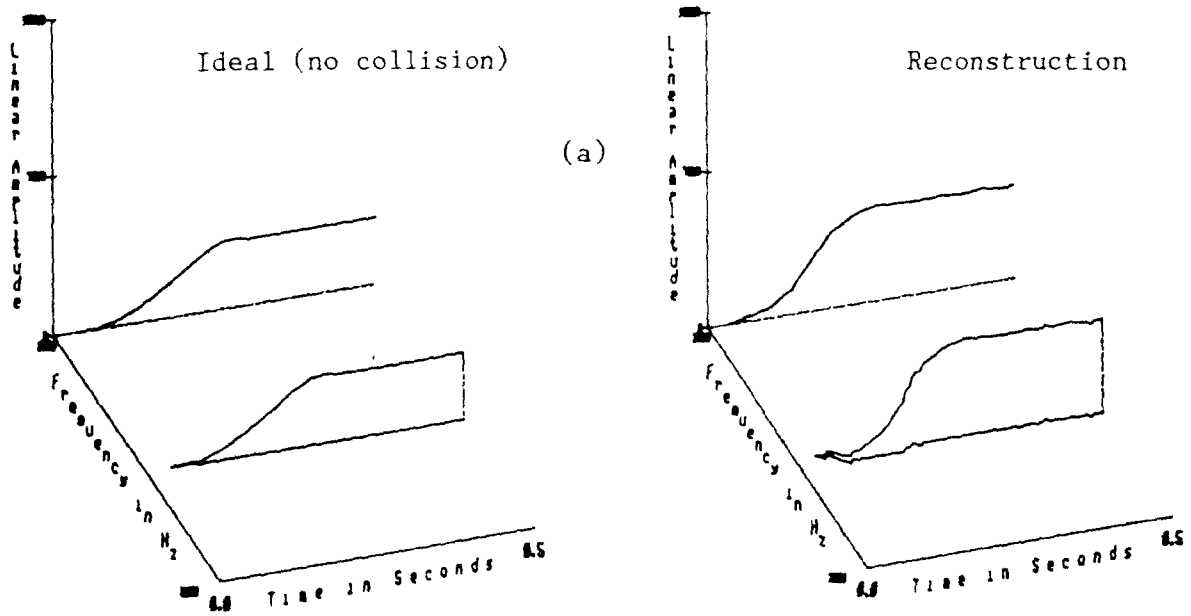


Figure 4.15: Separation Spectra of Example #2  
Using a priori Frequency Data.  
(a) voice 1 (b) voice 2

Partial 5 of Upper Voice



Partial 6 of Lower Voice

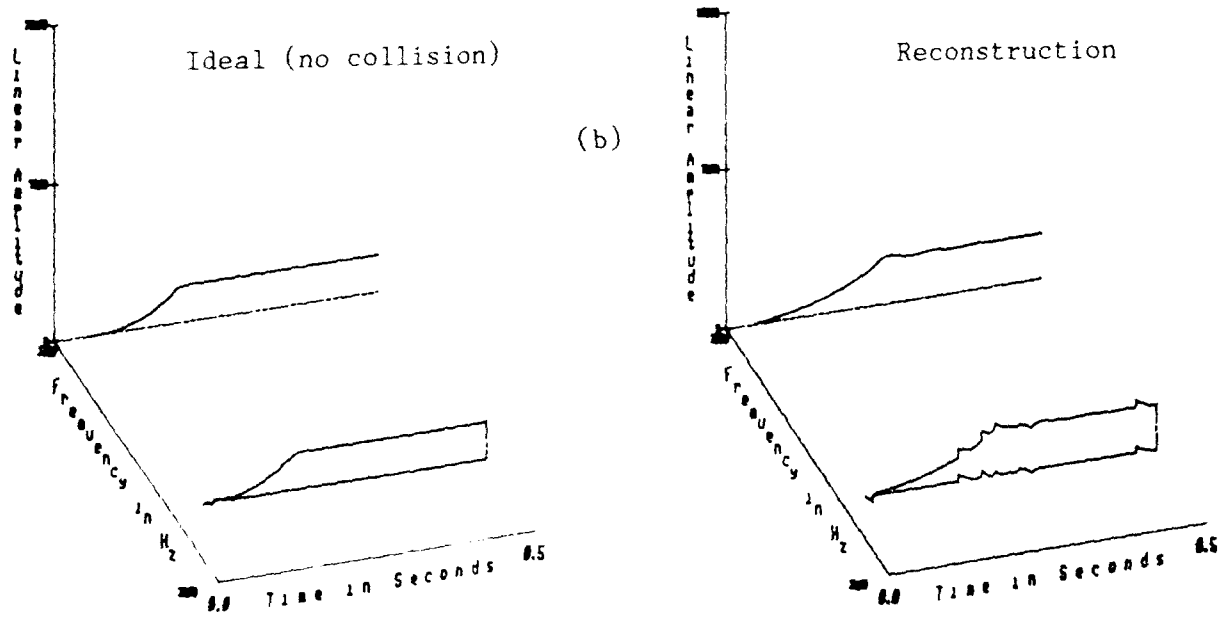


Figure 4.16: Collision Repair of Example #2.  
 (a) upper voice (b) lower voice

little spectral overlap. The separation results for example 3 are shown in Figure 4.17.

The separation output in the case of an amplitude mismatch between the duet voices is shown in Figure 4.18 for test example 4. Although the separation process has some difficulty interpreting the changing amplitude relationship between the two voices, the estimates of the colliding partials are never grossly in error.

For the duet of example 5 the separation process must deal with the vibrato present in the upper voice. A major difficulty in handling signals containing vibrato is in the transition between different separation strategies as collisions begin and end during the vibrato cycle. For example, consider a partial with a nominal frequency of 300 Hz and 5% frequency vibrato ( $300 \pm 15$  Hz) accompanied by another partial with a constant 350 Hz fundamental frequency. If a collision is defined to occur whenever the difference between two frequency components is less than 40 Hz, the two partials in this example will collide only during the portion of each vibrato cycle in which the frequency of the lower partial exceeds 310 Hz. Any discontinuity at the collision boundary, which might even go unnoticed if it were a one-time occurrence, can become very obvious when it repeats synchronously with the vibrato waveform. Some of these effects can be identified (and heard) as small, discontinuous features in the output data (see Figure 4.19).

To summarize this section:

- 1) The separation process is most effective for portions of the duet where the fundamental frequencies of both voices remain constant.

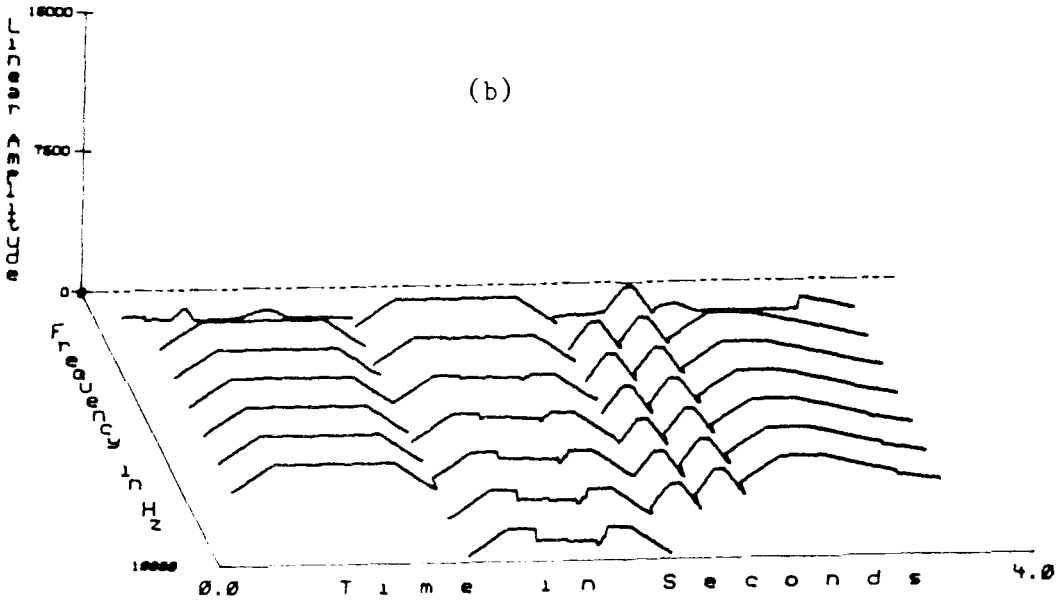
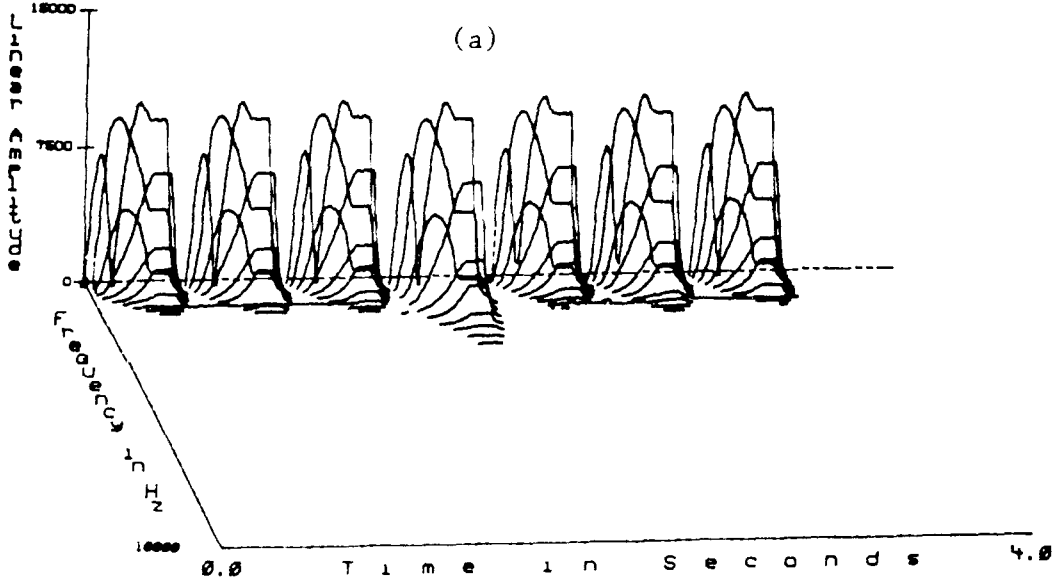


Figure 4.17: Separation Spectra of Example #3  
Using a priori Frequency Data.  
(a) voice 1 (b) voice 2

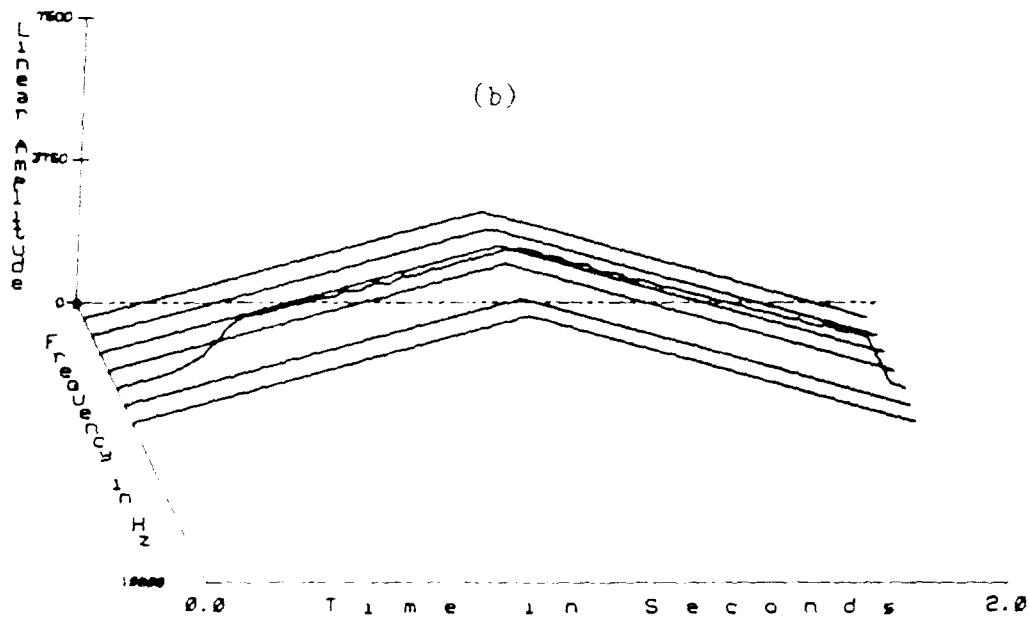
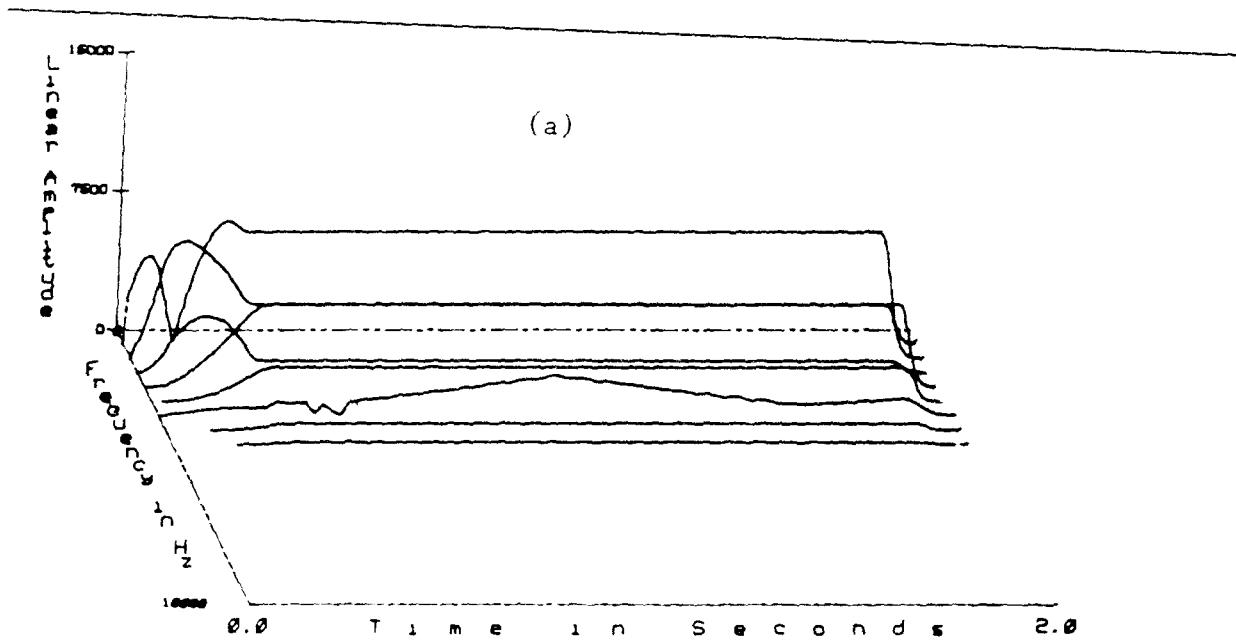


Figure 4.18: Separation Spectra of Example #4  
 Using a priori Frequency Data.  
 (a) voice 1 (b) voice 2



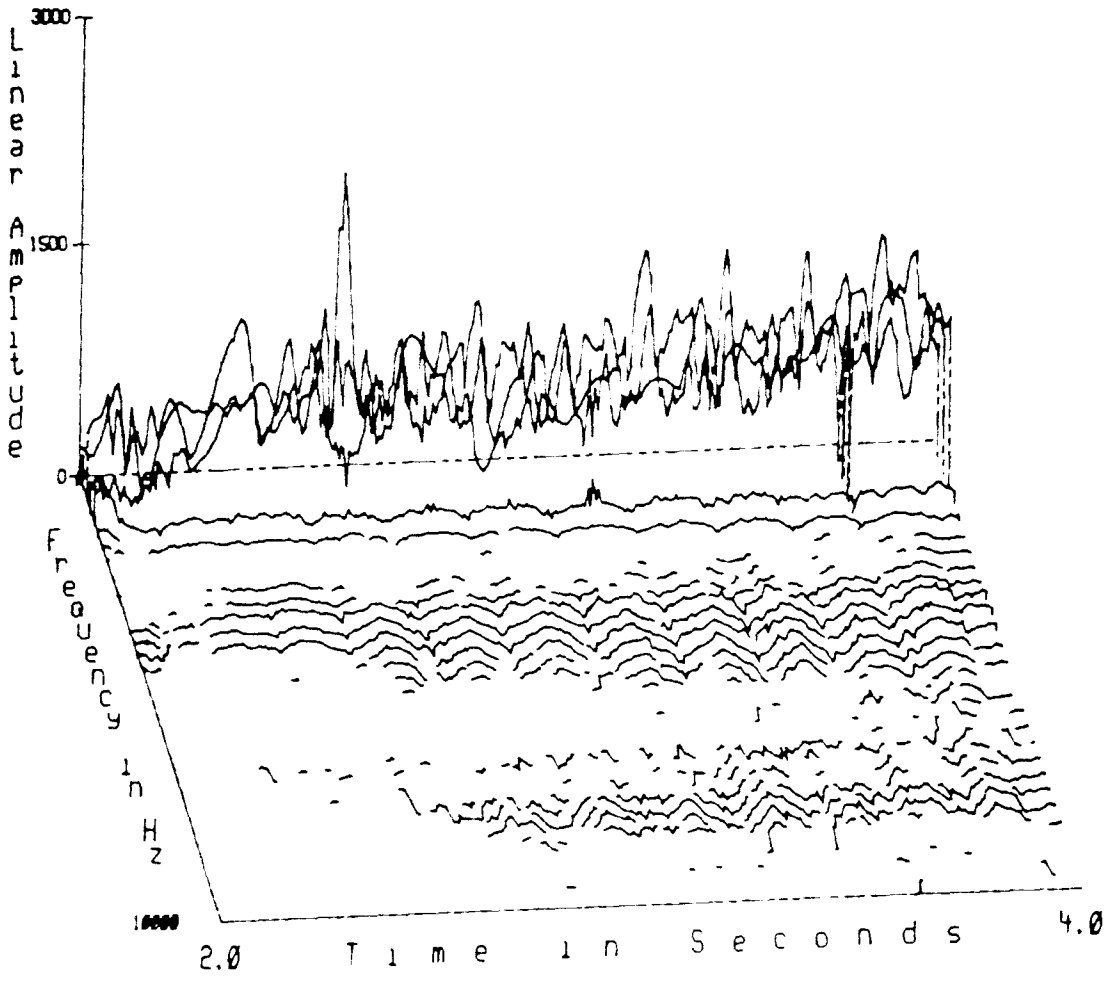


Figure 4.19: Separation Spectrum of Example #5  
Using a priori Frequency Data  
(Excerpt from Upper Voice).

- 2) The separation task may not be perfect, even in the "best case" situation of a priori knowledge of the duet fundamental frequencies. The discrepancies are primarily due to the time-bandwidth limitations of the short-time Fourier transform used in the analysis.
- 3) Reliable separation requires the frequency estimates for each partial to be within a few hertz of the true value in order for accurate collision detection and repair to be accomplished. For example, if the fundamental frequency estimate is in error by some small amount  $f_e$ , the frequency error for partial J will be  $J \cdot f_e$ . An error of this sort may cause problems for any separation strategy which attempts to extract features directly from the short-time spectrum.

#### 4.4 Evaluation of Voice Separation with Frequency Tracking

The complete automatic separation system was tested with the same test signals used to evaluate the TWM frequency tracker and the voice separation process with a priori frequency knowledge. The results for the four artificial duet examples were found to be comparable to the results obtained with a priori frequency knowledge, with the exception of some discrepancies due to inaccuracy of the TWM frequency tracking estimates. These results are summarized in Figures 4.20 through 4.23.

The separation results for the contrived duet of example 5 are shown in

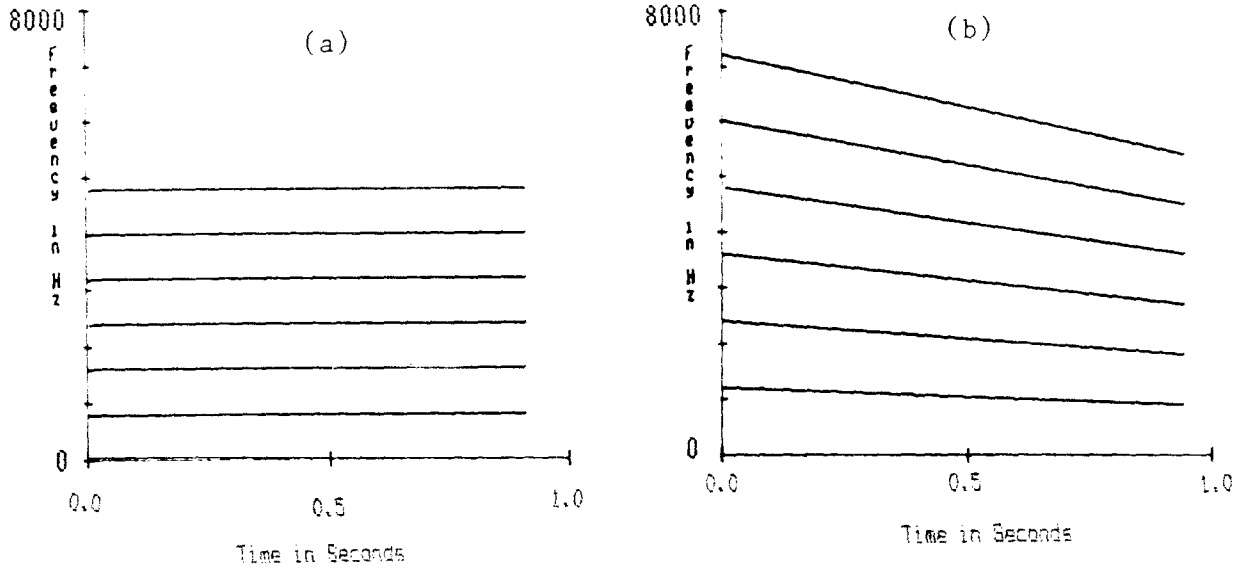


Figure 4.20: Results for Example #1 Using the Complete Separation Process. (a) voice 1 (b) voice 2

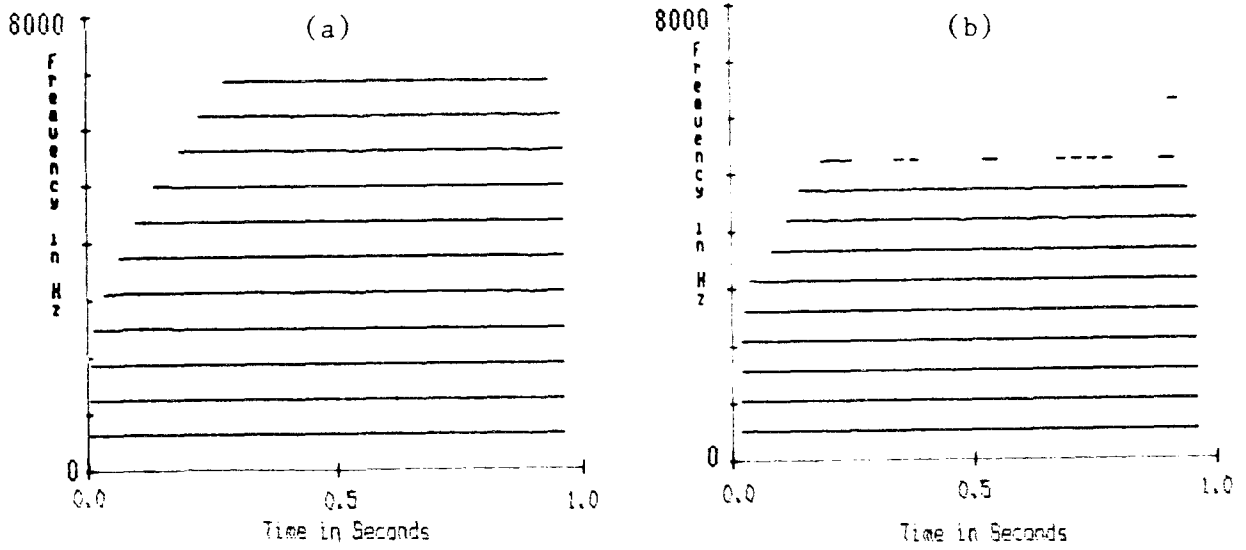


Figure 4.21: Results for Example #2 Using the Complete Separation Process. (a) voice 1 (b) voice 2

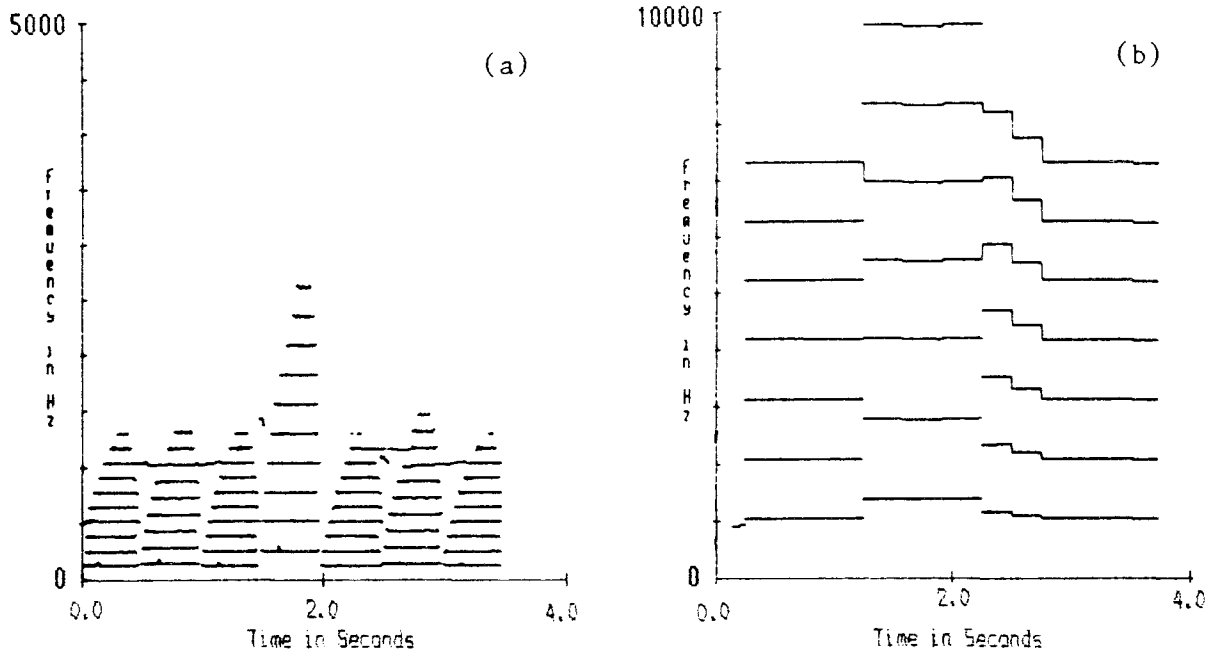


Figure 4.22: Results for Example #3 Using the Complete Separation Process.  
 (a) voice 1 (b) voice 2

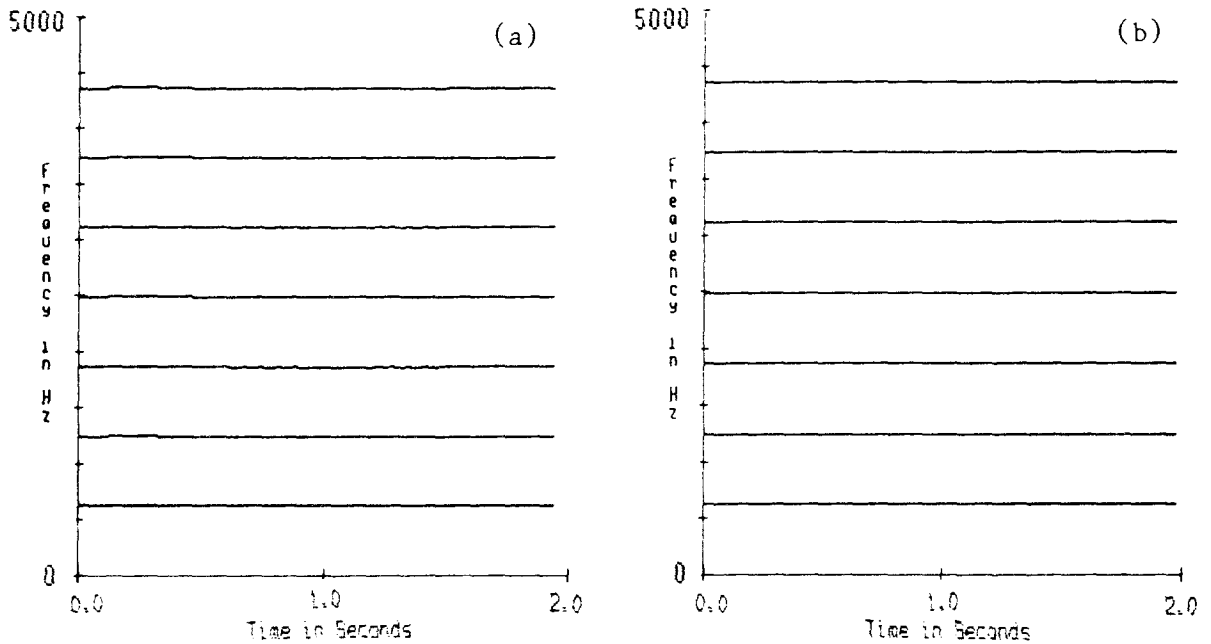


Figure 4.23: Results for Example #4 Using the Complete Separation Process.  
(a) voice 1 (b) voice 2

Figure 4.24. The uncertainty of the fundamental frequency estimates for the upper voice manifests itself in the imperfect separation results compared to the original signal (Figure 4.13), and the result with a priori frequencies (Figure 4.19).

In the Introduction, one of the stated assumptions was that only nonreverberant recordings should be processed. However, because of the pervasive nature of reverberation in recorded music, it was important to determine whether this restriction could be relaxed.

Duet test examples 6 and 7 are recorded segments obtained from analog record albums (for the musical scores see Figures 4.6 and 4.7). The reverberation present in the recordings is a source of trouble, because the duet separation procedure assumes that no more than two sets of harmonic

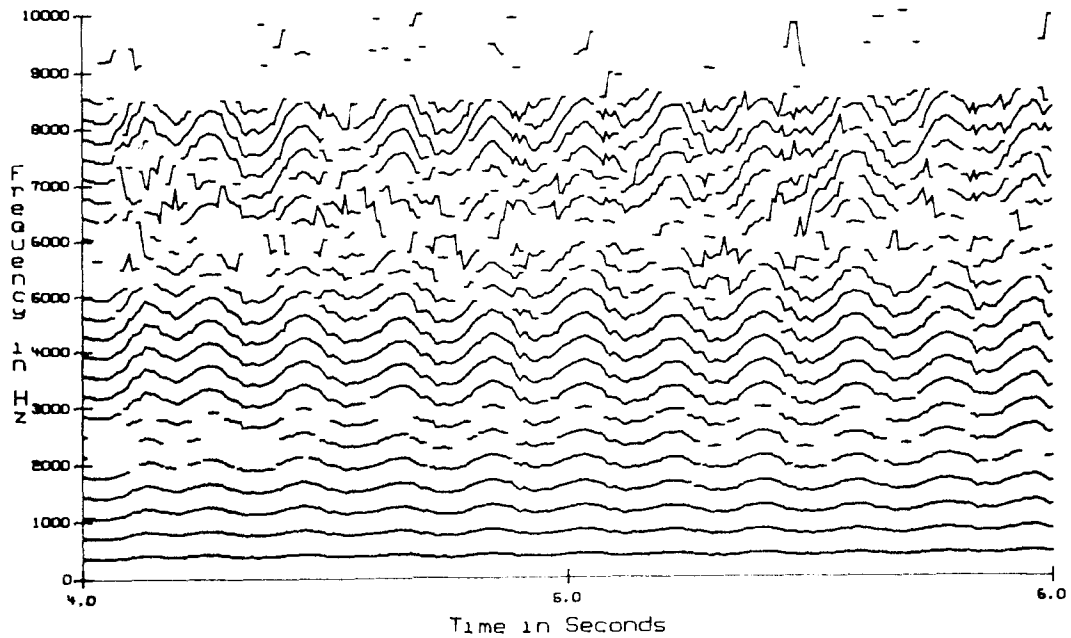


Figure 4.24: Results for Example #5 Using the Complete Separation Process (Excerpt from Upper Voice).

partials are present in a single analysis frame. With reverberation added, the implicit attack and release times of each note no longer match the explicit information available from the analysis data. This is because the frequency tracks of the current pair of notes are corrupted by the frequency tracks of previous notes due to reverberation.

When a voice changes from one note to the next, the frequency tracker may

- 1) begin to follow the new note,
- 2) continue to follow the reverb tail of the previous note, or
- 3) hop back and forth between the two choices in a random fashion.

The first case occurs when the reverb tails of the previous note cause partial collisions with the new note. The tails are NOT included in the two-voice separation process, so the collisions are NOT identified and corrected. The

second case occurs when the attack portion of the new note is missed, particularly if the new note starts at a lower amplitude than the reverberated note. When the third case occurs, the separation results are generally unsatisfactory.

The separation output for example 6 is shown in Figure 4.25. The presence of moderate surface noise (from the analog record album source) does not impair the performances of the frequency tracking and separation procedures. However, the reverberation present in the original recording is not so benign. The sound quality degradation consists of the audible presence of the reverberation tails of each note trailing over into the next note due to partial collisions. Another even more troublesome effect is the presence of the reverb tails of one voice in the separation output for the other voice!

Although the sounds of the reverb tails are desirable in the original recording, they cause undesirable artifacts in the separated voices. The reverb artifacts can be masked by the attack of the next note, however, so the separation process may be satisfactory for certain combinations of voices and reverb levels. The comments about the effect of reverberation for duet example 6 carry over to example 7.

In example 7 the frequency tracking data for the tuba and trumpet voices required manual editing. In this case the accuracy of the frequency tracking is limited by the considerable amount of reverb present in the recorded signal. After manual intervention to correct the frequency data, the extraction of the trumpet voice is quite good. This is because the sharp attacks on each note allow the frequency tracker to make a sharp transition at



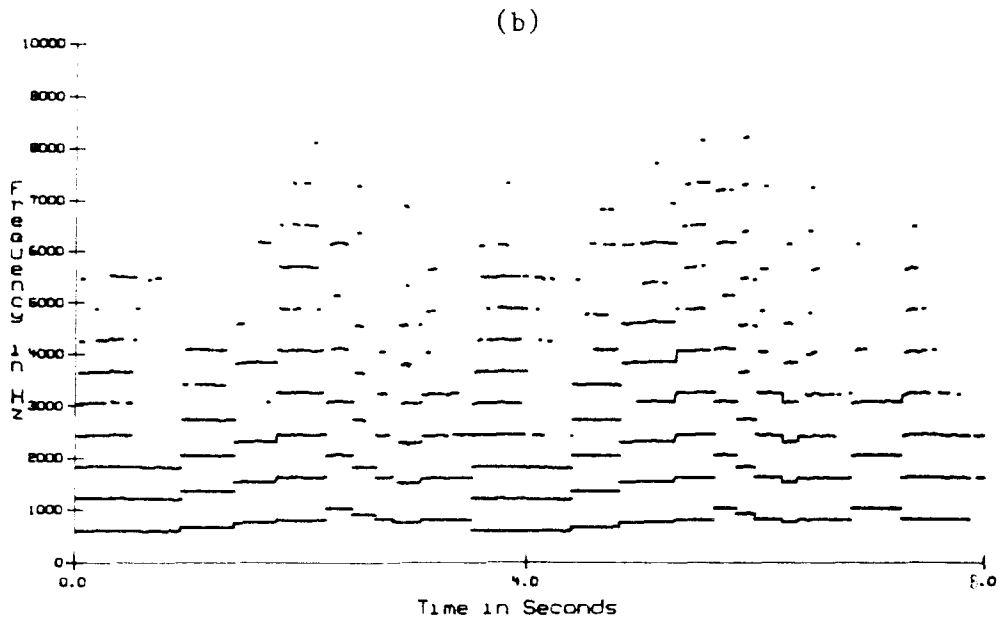
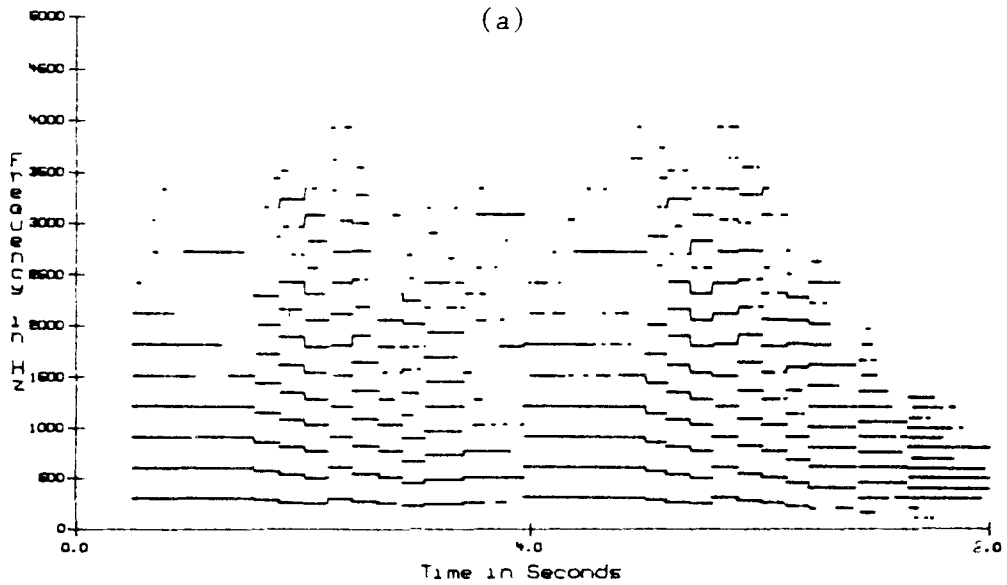


Figure 4.25: Results for Example #6 Using the Complete Separation Process.  
(a) voice 1 (b) voice 2

---

each new note. The tuba extraction is less successful, primarily due to the presence of reverb tails from several trumpet notes in the tuba data due to corruption of the closely spaced, low amplitude tuba partials by colliding partials from the trumpet. The separation results are shown in Figure 4.26.

Example 8 is a recording of the same musical passage as example 7, except that it was recorded in a nonreverberant room using two student musicians. For this example no hand intervention was required to obtain reliable fundamental frequency estimates, and the separation results are noticeably better than for the album recording. The results for example 8 are shown in Figure 4.27.

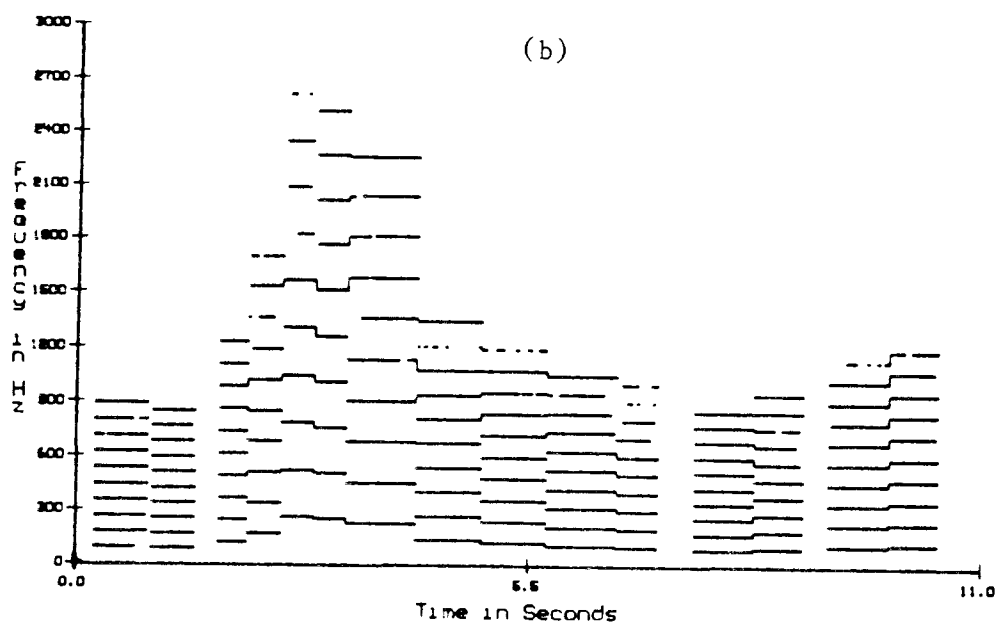
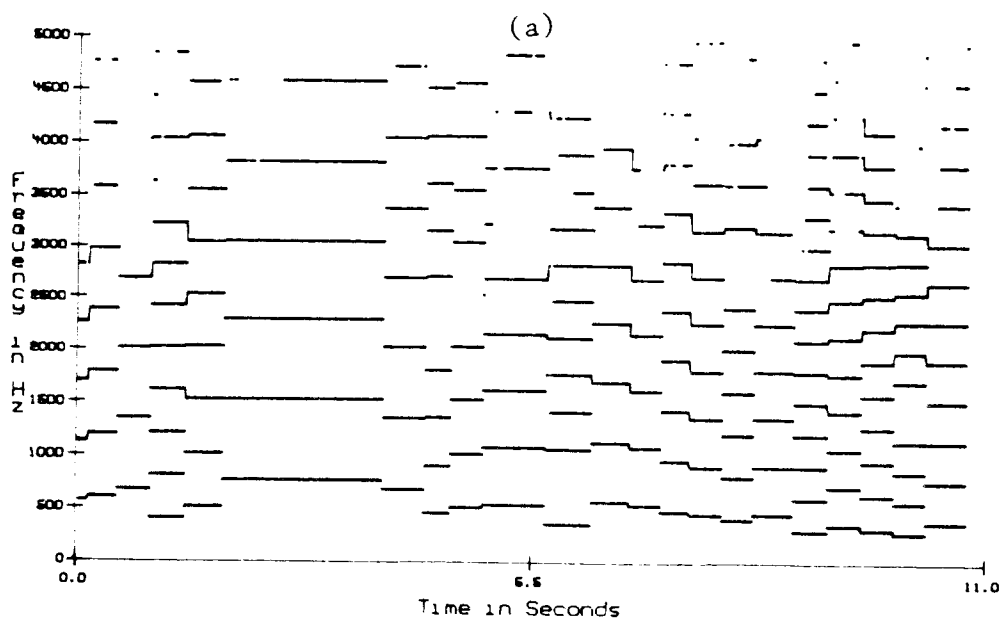


Figure 4.26: Results for Example #7 Using the Complete Separation Process (Fundamental Frequencies Manually Edited).  
(a) voice 1 (b) voice 2

---

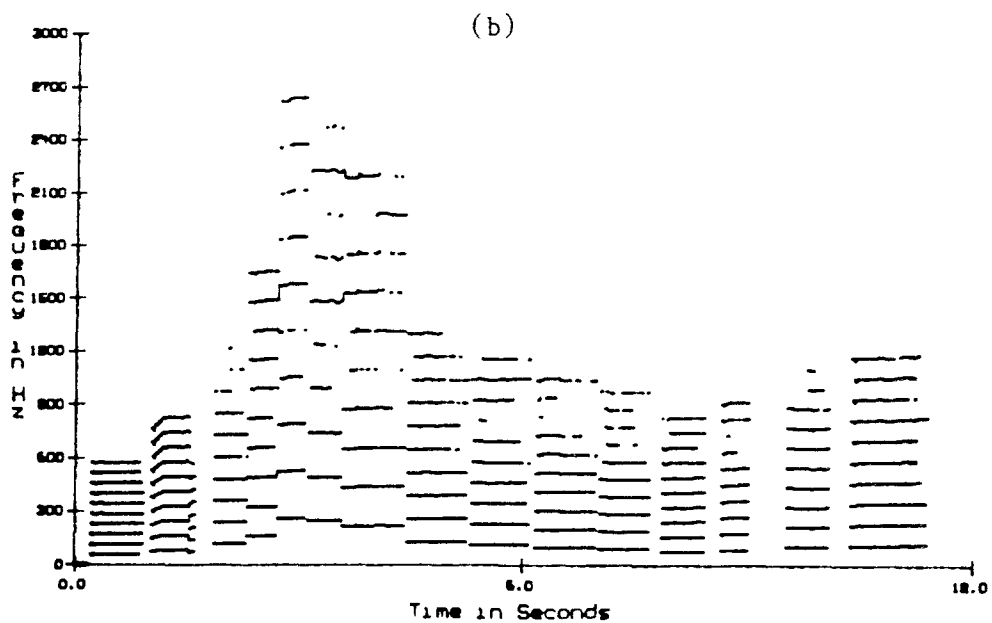
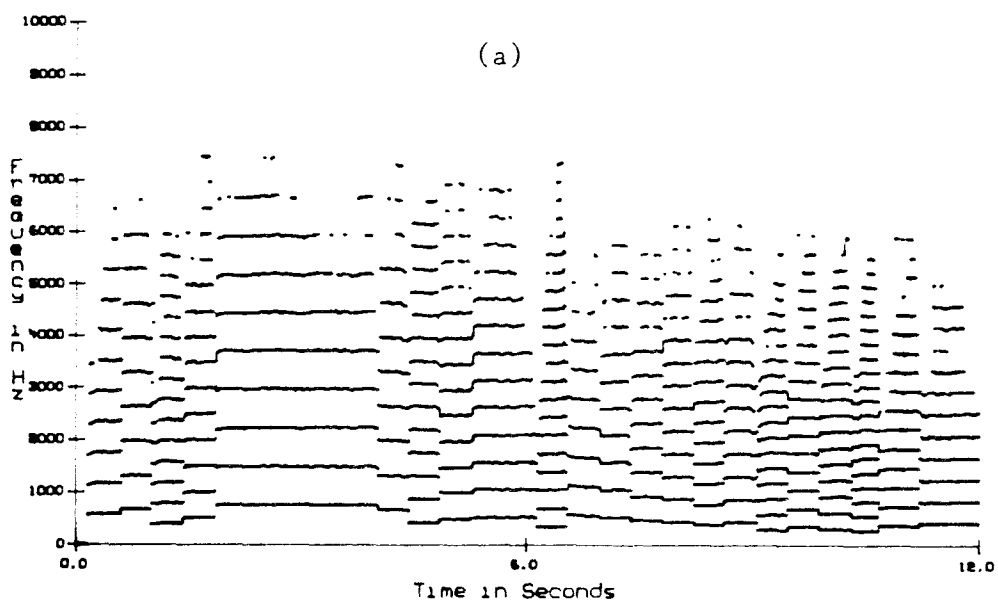


Figure 4.27: Results for Example #8 Using the Complete Separation Process.  
(a) voice 1 (b) voice 2

## CHAPTER 5

### CONCLUSIONS

Considered in this dissertation is the problem of automatic decomposition of a musical recording into its constituent signal components. With the current implementation, reasonable success in both frequency tracking and separation is assured only if recordings are restricted to duets of quasi-harmonic voices having nonintersecting fundamental frequency ranges specified in advance and performed with an absence of reverberation. The major goal of this work was to demonstrate the feasibility of composite signal decomposition using a time-frequency analysis procedure.

The analysis/separation/synthesis system was implemented in the C programming language on a small general-purpose digital computer (an IBM RT-PC Model 125 workstation). Most of the software has also been ported successfully to VAX-11 and Convex minicomputers.

#### 5.1 Summary of Findings

The research effort developed novel solutions to two fundamental problems: estimation of the two fundamental frequencies of a duet from the composite monaural signal and separation of the two voices given the pair of fundamental frequencies. The procedure deals with the following situations:

- 1) In a typical duet the partials of one voice collide with the partials of the other voice. The separation procedure determines the contribution of each colliding partial to the resulting amplitude and frequency interaction.

- 2) Level imbalances between the two voices may make the parameters of the weaker voice difficult to ascertain. The frequency tracking and voice separation procedures supply any missing information, if possible, using knowledge of previous and subsequent analysis frames.
- 3) The duet voices may either play simultaneously, one at a time (during solos), or not at all (during shared rests). The voice separation process determines the current voicing paradigm and applies the appropriate separation method.
- 4) For recordings containing reverberation or other characteristics not strictly within the guidelines set forth in Chapter 1, the performance of the frequency tracking and voice separation procedures is degraded. The acceptability of the degraded separation depends upon the particular combination of voices and reverb and the intended application for the results.

In summary, the results were excellent for combinations of voices and frequencies in which the number of collisions between partials of the two voices was small. The results were less satisfactory for frequency combinations in which one voice had most of its partials coincident with partials of the other voice, e.g., if the fundamental frequency of the upper voice was an integer multiple of the fundamental frequency of the lower voice. For typical duets the frequency relationship will change from note to note, causing the quality of the extraction to vary from note to note as well.

Several implementation details are considered in Appendix A, and the individual software modules are described in Appendix B.

## 5.2 Future Directions

The results of this dissertation indicate that the procedure merits further examination and development as it stands. However, some additional areas not addressed in this project remain intriguing topics for further research:

For a truly practical musical signal separation system, the output signal should never be perceptually "worse" than the input signal. The system must also be robust, with reasonable behavior for a wide range of input signals and minimal operator intervention. As mentioned in Chapter 2, these goals may imply a system with many levels of knowledge--from short-time spectra and pitch tracking, to note segmentation and perhaps even analysis of musical form. Moreover, such a system should be capable of adaptive behavior in response to the changing characteristics of its input signal.

The problems associated with acoustic signals need further study. In particular, the detrimental effect of reverberation encountered in this investigation needs to be resolved, since most music is recorded with natural or artificial reverberation. Further understanding of human perceptual strategies might provide necessary breakthroughs in this area.

The approach in this project has been to reconstruct the individual voices by resynthesis from modified time-variant analysis data. Another approach could be developed in which the time-variant analysis would be used for note segmentation only, with the signals themselves generated according to a prespecified artificial synthesis method. This approach has the advantage that the separated voices would have controllable timbres and would be less

susceptible to audible partial collision problems. The approach is an extension of the method discussed in Chapter 3, in which a synthesis model would be used for the replacement of partials damaged by collision. In this case all partials would be directly (or indirectly) replaced.

Finally, the MQ analysis/synthesis procedure has the potential to be a useful foundation for several interesting applications in audio signal processing. For example, the work of Serra (1986) and Smith and Serra (1987) exploited the peak-tracking and smooth-phase continuity properties of the MQ process for seamless splicing of real attack transients onto sustained synthetic waveforms. This concept could be extended to the editing situation in which an undesired "pop" or "click" must be removed from a digital recording, preferably without further corruption of the signal. The MQ procedure can also be employed for independent time or frequency modifications, such as pitch transposition, time scale compression/expansion, and sound synthesis.



APPENDIX A  
IMPLEMENTATION NOTES

The examples used in this dissertation were all obtained using an analog-to-digital converter running at a 20 kHz sample rate (monaural) and 16-bit (linear) quantization. All calculations during processing were performed using 32-bit floating point arithmetic.

The input signal was converted to floating point and pre-emphasized using a first-order fixed filter of the form:

$$H(z) = 1 - Ez^{-1} \tag{A.1}$$

with  $E=0.950$ . This high-pass pre-emphasis was included to help counteract the typical spectral rolloff of musical sounds with increasing frequency. Without pre-emphasis the low amplitude high-frequency partials are sometimes obscured by the analysis sidebands of stronger partials. The pre-emphasis also helps to equalize the partial amplitudes, compressing the internal dynamic range requirements of the analyzer.

The short-time Fourier transform (STFT) was implemented using a fixed-length Kaiser window either 511 or 1023 points in duration, corresponding to 25.55 msec or 51.15 msec, respectively. The choice of window length was made according to the lowest note expected in the signal to be analyzed: The longer window was used to resolve fundamental frequencies below approximately 100 Hz, with some loss of time resolution.

The pre-emphasized, windowed input data was then zero-padded by a factor of two (to length 1024 or 2048), and a standard fast Fourier transform (FFT) algorithm was used to obtain the discrete Fourier transform (DFT) for each data frame. The frame hop was set to a fixed increment of 128 samples (6.4 msec), corresponding to one-fourth or one-eighth of the window length.

The MQ analysis procedure (see Chapter 3) was applied to every frame of the STFT. The spectral peak selection process was limited in two ways:

- 1) A user-specified minimum peak amplitude was used as a global noise floor.
- 2) A floating threshold level 50 dB below the maximum spectral peak in a given frame was used to prevent misinterpretation of sidebands of the window transform as signal components.

Each peak value from the short-time spectrum was stored in a C language data structure, viz.

```
typedef struct _peak
{
    float mag;           /* magnitude of spectral peak */
    float freq;         /* frequency of spectral peak */
    float phase;        /* phase of spectral peak */
    short int link;     /* forward match to peak no. 'link' */
    struct _peak *prev; /* pointer to previous peak, this frame */
    struct _peak *next; /* pointer to next peak, this frame */
    struct _peak *bmatch; /* pointer to best match, previous frame */
    struct _peak *fmatch; /* pointer to best match, next frame */
} PEAK ;
```

After processing, a time domain signal was synthesized directly from the linked-list data structure using an additive procedure. Finally, the synthesized signal was de-emphasized using the inverse of the filter in (A.1),

converted from a 32-bit floating point to a 16-bit integer form, and passed through a digital-to-analog converter for evaluation.

APPENDIX B  
DESCRIPTION OF SOFTWARE MODULES

The two-way mismatch fundamental frequency tracker and the duet separation system described in this dissertation are each comprised of several separate computer programs, subprograms, functions, and subroutines.

The following is a functional summary of each of the software modules. This summary is provided to illustrate the development and evaluation approach used in this project.

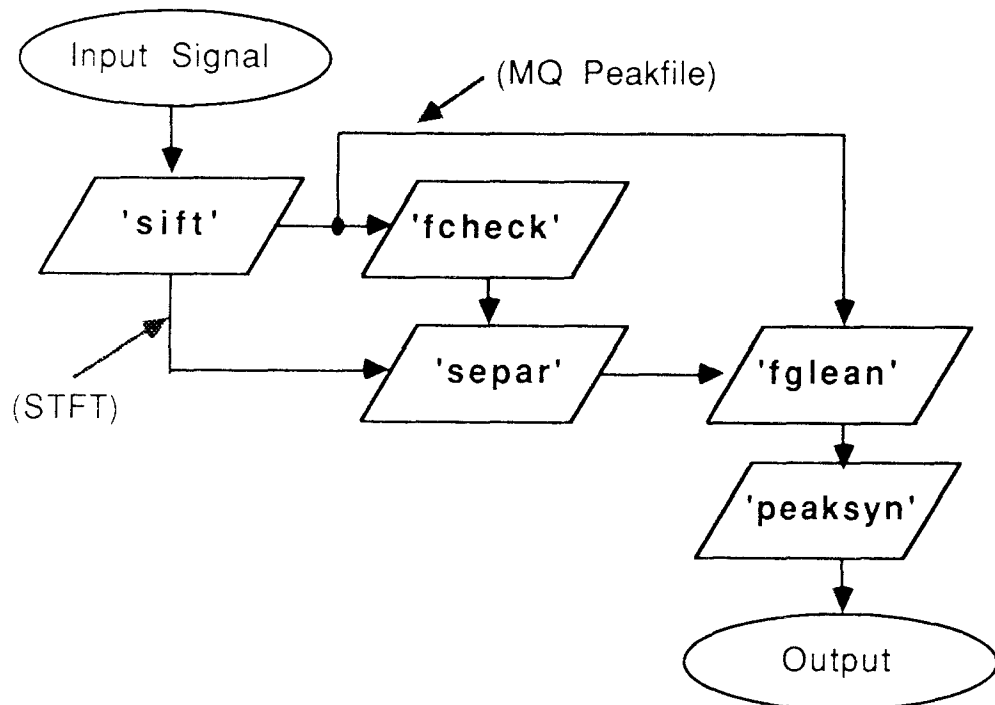


Figure B.1: Module Flow Diagram.

---

## 'SIFT'

The primary signal analysis program is named 'sift' because its purpose is to "sift out" the partials from the composite input signal.\* The 'sift' program computes the STFT of the input signal and produces an output file in the MQ analysis format (see Chapter 3). The program operates (conceptually) in pipeline fashion: while a frame of the input signal is read from a disk file, windowed, and processed by the FFT, the prior frame is processed into the MQ representation, and the STFT and MQ data of the frame before that are written to the output disk files. This form of implementation reduces the internal storage requirements of the analysis program. Also, output frames stored in the disk file are ready for examination even while processing is underway.

## 'FCHECK'

The 'fcheck' module implements the two-way mismatch (TWM) duet fundamental frequency estimation procedure. 'Fcheck' reads the MQ-format output file produced by 'sift' and produces a separate frequency output file for the two duet voices. The user supplies the name of the MQ-format file, the two nonoverlapping fundamental frequency ranges of the duet voices, and an estimate of the number of significant partials for each voice.

---

\*This name should NOT be confused with the Simplified Inverse Filter Tracking (SIFT) algorithm for fundamental frequency estimation using linear prediction techniques [Markel, 1972].

In the event that the frequency estimates produced by 'fcheck' contain obvious errors, an editor program called 'patch' can be used to manually correct the offending data.

### 'SEPAR'

The program 'separ' operates on three files containing information sources about the duet: its STFT, its MQ representation, and its frequency tracking estimates. 'Separ' uses this information to perform the track segregation and linear equations separation strategies described in Chapter 3. A pair of partials too close together for the direct separation strategies is marked for further processing. The 'separ' program is run twice: once to extract the lower voice of the duet, then again to extract the upper voice. Thus, the output of the two runs is a pair of MQ-format files ready for final processing and resynthesis.

### 'FGLEAN'

Any unresolved partial collisions contained in the output files produced by 'separ' are handled by 'fglean'. Unresolved collisions are treated by one of the secondary strategies discussed in Chapter 3, e.g., by examination of the amplitude "beat" patterns of the colliding partials.

'Fglean' also contains a set of empirical context rules designed to preserve at least first-order amplitude continuity of the partial tracks. For example, a rule to eliminate a single-frame "dropout" can be

expressed in words such as

IF: a partial track has amplitude ZERO in the current frame,  
 AND: the track is matched to a nonzero peak in the previous frame,  
 AND: the track is matched to a nonzero peak in the next frame,  
 THEN: replace the amplitude in the current frame with the average of  
 the matching peaks in the previous frame and the next frame.

The output of 'fglean' is the final MQ-format results for one of the duet voices.

#### 'PEAKSYN'

The 'peaksyn' program is used to synthesize a time-domain signal from the amplitude, frequency, and phase data for each track in an MQ-format file. The synthesized signal is converted to a 16-bit integer form for D/A conversion and playback.

In addition to the processing modules listed above, several graphics programs are used for display and evaluation of MQ-format files: 'Printpeak' prints the MQ analysis data in text form for a given range of frames (only used for debugging purposes); 'Readpeak' displays the MQ data as a frequency vs. time graph; and 'Dpeak' displays the MQ data as a three-dimensional graph of frequency and amplitude vs. time.

Program code listings have not been included in this dissertation due to their length. Interested persons may make arrangements with the author to obtain program listings and/or tape examples of this work.

## LIST OF REFERENCES

- J. B. Allen (1977), "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-25, no. 3, pp. 235-238.
- J. B. Allen and L. R. Rabiner (1977), "A unified approach to short-time Fourier analysis and synthesis," Proc. IEEE, vol. 65, no. 11, pp. 1558-1564.
- J. W. Beauchamp (1965), "Fourier series methods for analysis of a transient harmonic tone and a projected analog system for analysis and resynthesis of musical tones," Chapter 5 of "Electronic instrumentation for the synthesis, control, and analysis of harmonic tones," Ph.D. dissertation, Univ. of Illinois, Urbana, IL.
- J. W. Beauchamp (1966), "Transient analysis of harmonic musical tones by digital computer," Audio Eng. Soc. Preprint No. 479.
- J. W. Beauchamp (1969), "A computer system for time-variant harmonic analysis and synthesis of musical tones," in Music by Computer, H. von Foerster and J. W. Beauchamp, eds., New York: Wiley.
- J. W. Beauchamp (1975), "Analysis and synthesis of cornet tones using nonlinear interharmonic relationships," J. Audio Eng. Soc., vol. 23, no. 1, pp. 778-795.
- J. W. Beauchamp (1981), "Data reduction and resynthesis of connected solo passages using frequency, amplitude, and 'brightness' detection and nonlinear synthesis technique," Proc. Int. Computer Music Conf., pp. 316-323, Computer Music Assn., San Francisco, CA.
- A. H. Benade (1976), Fundamentals of Musical Acoustics. New York: Oxford University Press.
- K. W. Berger (1964), "Some factors in the recognition of timbre," J. Acoust. Soc. Am., vol. 36, no. 10, pp. 1888-1891.
- J. P. L. Brokx and S. G. Nootboom (1982), "Intonation and the perceptual separation of simultaneous voices," J. Phon., vol. 10, pp. 23-36.
- C. Chafe, B. Mont-Reynaud and L. Rush (1982), "Toward an intelligent editor of digital audio: recognition of musical constructs," Computer Music Journal, vol. 6, no. 1, pp. 30-41.
- E. C. Cherry (1953), "Some experiments on the recognition of speech with one and two ears," J. Acoust. Soc. Am., vol. 25, pp. 975-979.



- J. Chowning, L. Rush, B. Mont-Reynaud, C. Chafe, W. A. Schloss, and J. Smith (1984), "Intelligent systems for the analysis of digitized acoustic signals," Stanford Univ. Dept. of Music Rep. STAN-M-15, Stanford, CA.
- J. Chowning and B. Mont-Reynaud (1986), "Intelligent analysis of composite acoustic signals," Stanford Univ. Dept. of Music Rep. STAN-M-36, Stanford, CA.
- R. E. Crochiere (1980), "A weighted overlap-add method of short-time Fourier analysis/synthesis," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28, no. 1, pp. 99-102.
- R. E. Crochiere and L. R. Rabiner (1983), Multirate Digital Signal Processing. Englewood Cliffs, NJ: Prentice-Hall.
- R. G. Danisewicz and T. F. Quatieri (1988), "An approach to co-channel talker interference suppression using a sinusoidal model for speech," MIT Lincoln Laboratory Technical Report 794, Lexington, MA.
- A. Dembo and D. Malah (1988), "Signal synthesis from modified discrete short-time transform," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-36, no. 2, pp. 168-180.
- C. Dodge and T. Jerse (1985), Computer Music: Synthesis, Composition and Performance. New York: Schirmer Books.
- M. B. Dolson (1983), "A tracking phase vocoder and its use in the analysis of ensemble sounds," Ph.D. dissertation, California Institute of Technology, Pasadena, CA.
- M. B. Dolson (1985), "Recent advances in Musique Concrete at CARL," Proc. Int. Computer Music Conf., pp. 55-60, Computer Music Assn., San Francisco, CA.
- J. Everton (1975), "The separation of the voice signals of simultaneous speakers," Ph.D. dissertation, Univ. of Utah, Salt Lake City, UT.
- H. Fletcher (1934), "Loudness, pitch and the timbre of musical tones and the relation to the intensity, the frequency, and the overtone structure," J. Acoust. Soc. Am, vol. 6, pp. 59-69.
- S. Foster, W. A. Schloss, and A. J. Rockmore (1982), "Toward an intelligent editor of digital audio: signal processing methods," Computer Music Journal, vol. 6, no. 1, pp. 42-51.
- R. H. Frazier, S. Samsam, L. D. Braida, and A. V. Oppenheim (1976), "Enhancement of speech by adaptive filtering," Proc. IEEE ICASSP, pp. 251-253.

- M. D. Freedman (1965), "A technique for analysis of musical instrument tones," Ph.D. dissertation, Univ. of Illinois, Urbana, IL.
- M. D. Freedman (1967), "Analysis of musical instrument tones," J. Acoust. Soc. Am., vol. 41, p. 793.
- M. D. Freedman (1968), "A method for analyzing musical tones," J. Audio Eng. Soc., vol. 16, no. 4, p. 419.
- J. W. Gordon (1984), "Perception of attack transients in musical signals," Ph.D. dissertation, Stanford Univ., Stanford, CA (also Dept. of Music Rep. STAN-M-17).
- J. M. Grey (1975), "An exploration of musical timbre", Ph.D. dissertation, Stanford Univ., Stanford, CA (also Dept. of Music Rep. STAN-M-2).
- D. W. Griffin and J. S. Lim (1984), "Signal estimation from modified short-time Fourier transform," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-32, no. 2, pp. 236-242.
- D. W. Griffin and J. S. Lim (1988), "Multiband excitation vocoder," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-36, no. 8, pp. 1223-1235.
- B. A. Hanson and D. Y. Wong (1984), "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," Proc. IEEE ICASSP, vol. 2, pp. 18A.5.1-4.
- F. Harris (1978), "On the use of windows for harmonic analysis with the discrete Fourier transform," Proc. IEEE, vol. 66, no. 1, pp. 51-83.
- H. L. F. von Helmholtz (1885), On the Sensations of Tone as a Physiological Basis for the Theory of Music, tr. J. Ellis, London: Longmans (reprinted by Dover, New York, 1954).
- C. K. Lee and D. G. Childers (1988), "Cochannel speech separation," J. Acoust. Soc. Am., vol. 83, no. 1, pp. 274-280.
- D. Luce (1963), "Physical correlates of nonpercussive musical instrument tones," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.
- D. Luce and M. Clark, Jr. (1967), "Physical correlates of brass-instrument tones," J. Acoust. Soc. Am., vol. 42, no. 6, pp. 1232-1243.
- R. C. Maher and J. W. Beauchamp (1988), "Is there a single vibrato waveform?" J. Acoust. Soc. Am., Suppl. 1, vol. 83, p. S31.

- J. D. Markel (1972), "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. Audio Electroacoust., vol. AU-20, pp. 367-377.
- J. D. Markel and A. H. Gray, Jr. (1976), Linear Prediction of Speech. New York: Springer-Verlag.
- S. McAdams (1984), "Spectral fusion, spectral parsing, and the formation of auditory images," Ph.D. dissertation, Stanford Univ., Stanford CA (also Dept. of Music Rep. STAN-M-22).
- R. J. McAulay and T. F. Quatieri (1986), "Speech analysis/synthesis based on a sinusoidal representation," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34, no. 4, pp. 744-754.
- O. N. M. Mitchell, C. A. Ross, and G. H. Yates (1971), "Signal processing for a cocktail party effect," J. Acoust. Soc. Am., vol. 50, pp. 656-660.
- B. Mont-Reynaud and M. Goldstein (1985), "On finding rhythmic patterns in musical lines," Proc. Int. Computer Music Conf., pp. 391-397, Computer Music Assn., San Francisco, CA.
- J. A. Moorer (1974), "The optimum comb method of pitch period analysis of continuous digitized speech," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-22, no. 5, pp. 330-338.
- J. A. Moorer (1975), "On the segmentation and analysis of continuous musical sound by digital computer," Ph.D. dissertation, Stanford Univ., Stanford, CA (also Dept. of Music Rep. STAN-M-3).
- J. A. Moorer (1978), "The use of the phase vocoder in computer music applications," J. Audio Eng. Soc., vol. 26, no. 1/2, pp. 42-45.
- J. A. Naylor and S. F. Boll (1987), "Techniques for suppression of an interfering talker in co-channel speech," Proc. IEEE ICASSP, vol. 1, pp. 205-208.
- A. M. Noll (1966), "Cepstrum pitch determination," J. Acoust. Soc. Am., vol. 41, no. 2, pp. 293-309.
- A. V. Oppenheim and R. W. Schaffer (1975), Digital Signal Processing. Englewood Cliffs, NJ: Prentice-Hall.
- T. W. Parsons (1976), "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am., vol. 60, no. 4, pp. 911-918.
- Y. M. Perlmuter, L. D. Braida, R. H. Frazier, and A. V. Oppenheim (1977), "Evaluation of a speech enhancement system," Proc. IEEE ICASSP, pp. 212-215.

- M. Piszczalski and B. A. Galler (1977), "Automatic music transcription," Computer Music Journal, vol. 1, no. 4, pp. 24-31.
- M. Piszczalski and B. A. Galler (1979), "Predicting musical pitch from component frequency ratios," J. Acoust. Soc. Am., vol. 66, no. 3, pp. 710-720.
- M. R. Portnoff (1976), "Implementation of the digital phase vocoder using the fast Fourier transform," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, no. 3, pp. 243-248.
- M. R. Portnoff (1980), "Time-frequency representation of signals and systems based on short-time Fourier analysis," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28, no. 1, pp. 55-69.
- M. R. Portnoff (1981), "Time-scale modification of speech based on short-time Fourier analysis," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-29, no. 3, pp. 374-390.
- T. F. Quatieri and R. J. McAulay (1986), "Speech transformations based on a sinusoidal representation," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34, no. 6, pp. 1449-1464.
- L. R. Rabiner and R. W. Schafer (1978), Digital Processing of Speech Signals. Englewood Cliffs, NJ: Prentice-Hall.
- J. C. Risset and M. V. Mathews (1969), "Analysis of musical instrument tones," Physics Today, vol. 22, no. 2, pp. 23-30.
- M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley (1974), "Average magnitude difference function pitch extractor," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-22, no. 5, pp. 353-362.
- E. L. Saldanha and J. F. Corso (1964), "Timbre cues and the identification of musical instruments," J. Acoust. Soc. Am., vol. 36, no. 11, pp. 2021-2026.
- B. Sayers and E. C. Cherry (1957), "Mechanism of binaural fusion in the hearing of speech," J. Acoust. Soc. Am., vol. 29, pp. 973-987.
- W. A. Schloss (1985), "On the automatic transcription of percussive music--from acoustic signal to high-level analysis," Ph.D. dissertation, Stanford Univ., Stanford, CA (also Dept. of Music Rep. STAN-M-27).
- M. R. Schroeder (1968), "Period histogram and product spectrum: new methods for fundamental frequency measurement," J. Acoust. Soc. Am., vol. 43, no. 4, pp. 829-834.

- X. Serra (1986), "A computer model for bar percussion instruments," Proc. Int. Computer Music Conf., pp. 257-262, Computer Music Assn., San Francisco, CA.
- V. C. Shields, Jr. (1970), "Separation of added speech signals by digital comb filtering," M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA.
- W. A. Slawson (1982), "The musical control of sound color," Canadian Univ. Music Review, no. 3, pp. 67-79.
- J. O. Smith and X. Serra (1987), "PARSHL: an analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," Proc. Int. Computer Music Conf., pp. 290-297, Computer Music Assn., San Francisco, CA.
- M. M. Sondhi (1968), "New methods of pitch extraction," IEEE Trans. Audio Electroacoust., vol AU-16, pp. 262-266.
- T. G. Stockham, Jr. (1971), "Restoration of old acoustic recordings by means of digital signal processing," Preprint, 41st Convention, Audio Engineering Society, New York.
- J. Strawn (1987), "Analysis and synthesis of musical transitions using the discrete short-time Fourier transform," J. Audio Eng. Soc., vol. 35, no. 1/2, pp. 3-13.
- R. J. Stubbs and Q. Summerfield (1988), "Evaluation of two voice-separation algorithms using normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am., vol. 84, no. 4, pp. 1236-1249.
- M. Weintraub (1985), "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Stanford Univ., Stanford, CA.
- D. Wessel (1985), "Timbre space as a musical control structure," in Foundations of Computer Music, C. Roads and J. Strawn, eds., Cambridge, MA: MIT Press.
- E. H. Wold and A. M. Despain (1986), "Parameter estimation of acoustic models: audio signal separation," Proc. 1986 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, IEEE, New York.
- M. Yanagida, Y. Miyoshi, Y. Nomura, and O. Kakusho (1985), "Application of the least-squares method to sound-source separation in a multi-source environment," Acustica, vol. 57, pp. 158-167.
- U. T. Zwicker (1984), "Auditory recognition of diotic and dichotic vowel pairs," Speech Commun., vol. 3, pp. 265-277.

## VITA

Robert Crawford Maher was born on January 24, 1962, in Cambridge, England, of American parents. He attended public schools in Middleton, Wisconsin.

In 1984, Mr. Maher received his Bachelor of Science degree in Electrical Engineering, magna cum laude, from Washington University in St. Louis, and was named Outstanding Senior in the Electrical Engineering Department. While at Washington University, he held a four-year, full-tuition Langsdorf Fellowship, and was the recipient of National Merit and National Honor Society Scholarships. Also in 1984, he was awarded a National Science Foundation Graduate Fellowship for three years of graduate study in Electrical Engineering.

In 1985, he received his Master of Science degree in Electrical Engineering from the University of Wisconsin at Madison, and began doctoral work in that field at the University of Illinois, Urbana-Champaign, under Dr. James Beauchamp. In 1985-1986 and 1988-1989 he was a research assistant in the areas of digital audio, acoustics, and computer music. He also received a University of Illinois Graduate Fellowship for 1985-1986, and an Audio Engineering Society Educational Grant for 1988-1989.

Mr. Maher has active research interests in the application of engineering and signal processing methods in the audio arts, particularly in digital audio systems, music composition and performance, and acoustics. He is a member of the Tau Beta Pi, Eta Kappa Nu, and Phi Kappa Phi honor societies, and several professional organizations including the IEEE, AES, ASA, CMA, and the IMA.