Combined-Channel Instantaneous Frequency Analysis for Audio Source Separation Based on Comodulation

by

Barry David Jacobson

B.S. Electrical Engineering, Columbia University, 1984
M.S. Electrical Engineering, Columbia University, 1992

SUBMITTED TO THE HARVARD-MIT DIVISION OF HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2008

Signature of Author: .....
Harvard-MIT Division of Health Sciences and Technology
May 15, 2008

Certified by: .....
Thomas F. Quatieri
Senior Member of Technical Staff, MIT Lincoln Laboratory
Affiliated Faculty, Harvard-MIT Division of Health Sciences and Technology
Thesis Supervisor

Certified by: .....
Gert Cauwenberghs
Professor of Neurobiology, UCSD
Thesis Supervisor

Accepted by: .....
Martha L. Gray
Edward Hood Taplin Professor of Medical and Electrical Engineering
Co-Director, Harvard-MIT Division of Health Sciences and Technology

Combined-Channel Instantaneous Frequency Analysis for Audio Source Separation Based on Comodulation

by

Barry David Jacobson

Submitted to the Harvard-MIT Division of Health Sciences and Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

Abstract

Normal human listeners have a remarkable ability to focus on a single sound or speaker of interest and to block out competing sound sources. Individuals with hearing impairments, on the other hand, often experience great difficulty in noisy environments. The goal of our research is to develop novel signal processing methods inspired by neural auditory processing that can improve current speech separation systems. These could potentially be of use as assistive devices for the hearing impaired, and in many other communications applications. Our focus is the monaural case where spatial information is not available.

Much perceptual evidence indicates that detecting common amplitude and frequency variation in acoustic signals plays an important role in the separation process. The physical mechanisms of sound generation in many sources cause common onsets/offsets and correlated increases/decreases in both amplitude and frequency among the spectral components of an individual source, which can potentially serve as a distinct signature. However, harnessing these common modulation patterns is difficult because when spectral components of competing sources overlap within the bandwidth of a single auditory filter, the modulation envelope of the resultant waveform resembles that of neither source.

To overcome this, for the coherent, constant-frequency AM case, we derive a set of matrix equations which describes the mixture, and we prove that there exists a unique factorization under certain constraints. These constraints provide insight into the importance of onset cues in source separation. We develop algorithms for solving the system in those cases in which a unique solution exists. This work has direct bearing on the general theory of non-negative matrix factorization which has recently been applied to various problems in biology and learning.

For the general, incoherent, AM and FM case, the situation is far more complex because constructive and destructive interference between sources causes amplitude fluctuations within channels that obscures the modulation patterns of individual sources. Motivated by the importance of temporal processing in the auditory system, and specifically, the use of extrema, we explore novel methods for estimating instantaneous amplitude, frequency, and phase of mixtures of sinusoids by comparing the location of local maxima of waveforms from various frequency channels. By using an overlapping exponential filter bank model with properties resembling the cochlea, and combining information from multiple frequency bands, we are able to achieve extremely high frequency and time resolution. This allows us to isolate and track the behavior of individual spectral components which can be compared and grouped with others of like type.

Our work includes both computational and analytic approaches to the general problem. Two suites of tests were performed. The first were comparative evaluations of three filter-bank-based algorithms on sets of harmonic-like signals with constant

frequencies. One of these algorithms was selected for further performance tests on more complex waveforms, including AM and FM signals of various types, harmonic sets in noise, and actual recordings of male and female speakers, both individual and mixed.

For the frequency-varying case, initial results of signal analysis with our methods appear to resolve individual sidebands of single harmonics on short time scales, and raise interesting conceptual questions on how to define, use and interpret the concept of instantaneous frequency.

Based on our results, we revisit a number of questions in current auditory research, including the need for both rate and place coding, the asymmetrical shapes of auditory filters, and a possible explanation for the deficit of the hearing impaired in noise.

Committee Members

Bertrand Delgutte, Ph.D., Committee Chairman
Associate Professor of Otology and Laryngology and Health Sciences and Technology, Harvard Medical School

Tom Quatieri, Ph.D., Thesis Supervisor
Senior Member of Technical Staff, MIT Lincoln Laboratory
Affiliated Faculty, Harvard-MIT Division of Health Sciences and Technology

Gert Cauwenberghs, Ph.D., Thesis Supervisor
Professor of Neurobiology, University of California at San Diego

Louis Braida, Ph.D.
Henry Ellis Warren Professor of Electrical Engineering, MIT
Professor of Health Sciences and Technology, Harvard-MIT

George Zweig, Ph.D.
Professor Emeritus, Theoretical Physics, California Institute of Technology
Fellow, Los Alamos National Laboratory

Table of Contents

Table of Contents 5
Acknowledgments 11
Chapter 1 Introduction 17
1.1 Overview 17
1.2 Biological Approaches to Source Separation 18
1.3 Concept of Comodulation 19
1.4 Thesis Organization 25
Chapter 2 Review of Previous Work 27
2.1 General Categories 27
2.2 Sound Direction 28
2.2.1 Auditory Beamforming 28
2.2.2 Independent Component Analysis 29
2.3 Harmonic Relationships 30
2.4 Computational Auditory Scene Analysis 30
2.4.1 A Survey 31
2.4.2 Contrasts with Visual Scene Analysis 38
2.5 Approaches based on Spectral Estimation 39
2.6 Sinusoidal Modeling 40
2.7 Coherence-Based Approach 41
2.8 Super-resolution Methods for Sine Estimation 41
2.8.1 Linear Prediction 42
2.8.2 Capon's Method 43
2.8.3 Pisarenko's Method 44
2.8.4 MUSIC (Multiple Signal Classification Algorithm) 44
2.8.5 ESPRIT (Estimation of Signal Parameters via Rotational Invariance Techniques) 45
2.8.6 Comparison 46
2.9 Recent Work on Time-Frequency Analysis 47
2.9.1 Norden Huang's Method 47
2.9.2 The Reassigned Spectrogram 48
2.9.3 The Local Vector Transform 49
2.10 Conclusion 50
Chapter 3 Aspects of Comodulation 53
3.1 Introduction 53
3.2 Application to Music 53
3.2.1 Timbre 54
3.2.2 Sample instruments 55
3.2.3 Altoflute 57
3.2.4 Violin 59
3.2.5 Trumpet 61

3.2.6 Piano 63
3.2.7 Bass Clarinet 66
3.2.8 Oboe 68
3.3 Relevance to Speech 69
3.3.1 Source-Filter Model 69
3.3.2 Formants 69
3.3.3 Vowels 70
3.3.4 Consonants 70
3.3.5 Continuity in Speech 71
3.3.6 Radiation Resistance 71
3.3.7 Singing 71
3.3.8 Amplitude Comodulation and Speech 73
3.3.9 Frequency Comodulation and Speech 74
3.4 Comodulation: A Priori or a Posteriori 74
3.5 Reduction of Ambiguity 75
3.6 Phase Comodulation: Scaling 81
3.6.1 Amplitude Comodulation: Vertical Scaling 81
3.6.2 FM Comodulation: Horizontal Scaling 82
3.6.3 Phase Comodulation: Time-Domain Perspective 84
3.6.4 Waveforms: Cycle to Cycle Comparison 85
3.6.5 Why Imperfect Scaling Occurs 87
3.7 Summary 89
Chapter 4 A Mathematical Analysis of Amplitude Comodulation 91
4.1 Introduction 91
4.2 Initial Intuition 91
4.3 Source Representation 93
4.4 Determination of Number of Sources 94
4.5 Source Identification 94
4.6 Uniqueness Theorem for Non-Negative Matrix Factorization 95
4.7 Proof 96
4.8 Interpretation 100
4.9 Multidimensional Case 100
4.10 Simple Proof for Square Case 101
4.11 Examples 102
4.11.1 Example 1: Non-Unique Non-Negative Decomposition 102
4.11.2 Example 2: Unique Non-Negative Decomposition 104
4.12 Solution by Inspection 104
4.12.1 Method 104
4.12.2 Formal Proof 106
4.12.3 Implications 107
4.13 Alternating Iterative NNMF Algorithm 107
4.14 Results: Amplitude-Modulated Harmonic Sets 109
4.14.1 Description of Signals 109
4.14.2 Explanation of Results 110
4.14.3 Evaluation of Performance and Sources of Error 110

4.15 Results: Mixture of Clarinet and Oboe 112
4.15.1 Description of Signals 112
4.15.2 Explanation of Figures and Results 112
4.15.3 Evaluation of Performance and Sources of Error 112
4.15.4 Tremolo 113
4.16 Results: Mixture of Violin and Oboe 115
4.16.1 Description of Signals 115
4.16.2 Evaluation of Performance and Sources of Error 115
4.17 Results: Out of Phase Harmonic Sets 118
4.17.1 Description of Signals 118
4.17.2 Evaluation of Performance and Sources of Error 118
4.18 Relation to Auditory Scene Analysis 121
4.18.1 Onsets 121
4.18.2 Comodulation Masking Release 121
4.18.3 Future Refinements 122
4.19 Phase 122
4.20 Other Issues 125
4.20.1 Requirements for Real-Time Operation 125
4.20.2 Effect of Noise 125
4.20.3 Other Work on Non-Negative Matrix Factorization 126
4.21 Summary 126
Chapter 5 Estimating Instantaneous Frequency in Mixtures of Sinusoids 129
5.1 Introduction 129
5.2 Resolution Issues in Source Separation 130
5.3 Considerations in Filtering: The Speech Case 132
5.3.1 Spectrograms 132
5.3.2 Waveforms 134
5.3.3 AM vs. FM Characteristics under Filtering 136
5.3.4 Constructive and Destructive Interference (Beating) 140
5.3.5 Comparison of Higher Harmonics 140
5.3.6 Summary 143
5.4 Spectral Estimation Based on Local Maxima 143
5.4.1 Temporal Processing for Frequency Resolution 144
5.4.2 Estimation of Parameters 145
5.4.3 Behavior of Local Maxima in Mixtures 145
5.5 Iterative-Subtraction Algorithm 150
5.5.1 Single Channel—Introduction and Motivation 150
5.5.2 Algorithm 151
5.5.3 Results 152
5.5.4 Multiple Channels 155
5.5.5 FM Signals 160
5.5.6 Multiple Signals 164
5.5.7 Mathematical Simulations of Actual Filters 169
5.5.8 Discussion 171
5.6 Peak-Locus Algorithm 171

5.6.2 Algorithm 182
5.6.3 Discussion 182
5.7 Simultaneous-Equation Approach 186
5.7.1 Determining Number of Sources 188
5.7.2 Algorithm 189
5.7.3 Discussion 190
5.8 Filter Design Considerations 191
5.8.1 Rectangular Filter Design: A Motivating Case 191
5.8.2 Filter Bank and Signal Specification 195
5.9 Comparisons and Evaluations 199
5.9.1 Weighted case: Iterative-Subtraction Method 201
5.9.2 Weighted case: Simultaneous-Equations Method 202
5.9.3 Weighted case: Peak-Locus Method 203
5.9.4 Filtered Case: Simultaneous-Equations Method 204
5.9.5 Analysis of Results 205
5.10 Summary 205
Chapter 6 Combined-Channel Analysis of Modulation, Noise and Speech 207
6.1 Introduction 207
6.2 Filter Parameters 207
6.3 Harmonic Mixture-Delta 1 Hz 208
6.3.1 Results with Simultaneous-Equations Algorithm 208
6.3.2 Results Using FFT 209
6.3.3 Source of Errors in FFT 211
6.4 Modulated Sinusoids 217
6.4.1 AM Ramp-Modulated sine 217
6.4.2 FM Ramp-Modulated Sine 217
6.4.3 AM Sine-Modulated Sine. 219
6.4.4 FM Sine-Modulated Sine 225
6.5 Noise tests 231
6.5.1 10 dB SNR 231
6.5.2 0 dB SNR 232
6.5.3 -10 dB SNR 234
6.5.4 Comparison of Noise Test Results with FFT 235
6.6 Speech Tests 237
6.6.1 Male Speech 237
6.6.2 Female Speech 239
6.6.3 Mixed Speech 242
6.7 Analysis 247
6.7.1 Instantaneous Frequency 247
6.7.2 Comodulation and Sidebands 250
6.8 Transient Response 251
6.9 Summary 257
Chapter 7 Analytical Approaches 259
7.1 Introduction 259
7.2 Derivation 260

7.3 Solving for the Nullspace 268
7.3.1 Relation between Frequencies and Phases 268
7.3.2 Background 268
7.3.3 Solution for Frequency 269
7.3.4 Determining Phase 271
7.3.5 Computing Nullspace from SVD 273
7.4 Explicit Formula for Nullspace 274
7.4.1 Nullspace of Q and Z Matrices Separately 274
7.4.2 Method of Dai-Jones 276
7.5 Graphical Bounds on Frequency 277
7.6 Summary 281
Chapter 8 Summary and Conclusion 283
8.1 Summary 283
8.2 Future Work 285
Appendix 289
Comparison of Independent Component Analysis and Comodulation 289
References 295

Acknowledgments

The author is grateful for the help of many people over the years. There is no such thing as a self-made man. So much time and effort were invested into this work and my upbringing and education by numerous individuals. First and foremost are my parents Dr. Myron Jacobson, and Dr. Janice Jacobson Sokolovsky, who in addition to attending to my physical needs, tried their best to provide a stimulating home environment and a solid foundation for future accomplishment. My father received a B.S. from McGill University, and an M.D. from the University of Chicago, and my mother received a B.A. and a J.D. from the University of Chicago, and a Ph.D. in Economics from Columbia University. My aunt Edith Benjamin, and my grandmother Rose Mark, of blessed memory would often take me to the outstanding museums in Chicago's South Side when I was a child. My favorite was the Museum of Science and Industry. There was always something interesting to see, and always many hands-on activities and demonstrations to develop a child's curiosity. These two people played a great role in my upbringing with their encyclopedic knowledge of English and Jewish literature. Their stories kept my brothers and I fascinated for many hours on end. My grandmother on my father's side, Fanny Jacobson, also took great pride in her grandchildren's achievements, was an expert in the business world and well-read in many other areas. My grandfather, Solomon Jacobson, whom I was only privileged to know for a short time, was talented musically and played the violin. My mother's husband Avram, was kind and warm-hearted, and very much a Renaissance man with far-reaching interests in philosophy and the arts. My father's wife, Miriam, was kind enough to provide hospitality during our visits to NY.

I would like to thank my brothers, Mike and Danny and their friends for all the good times we had, and for all their help with many things along the way. I enjoyed coming back from Boston for all the family gatherings over the years. Mike is married to Rachel Hankoff and they have three wonderful girls, Emma, Juliett and Eliana.

My uncle, Sandy Jacobson, is a science teacher *par excellence* and he instilled in me a fascination for scientific devices and concepts from my earliest years. He encouraged me in my ham radio hobby, which led to a desire to understand electronics and signal processing theory on a deeper level. I remember all the chemicals and radio parts in my grandmother's basement that he and my father used to tinker with, and the numerous experiments and demos he showed us, including how to make home-made flairs out of strontium and magnesium metal (I seem to recall) in the backyard. Another time they mixed sulfuric acid and sugar. I have said publicly that most people get upset if their car breaks down, but my uncle is just the opposite, he views it as a challenging and enjoyable project to keep him occupied on a Sunday afternoon. I would like to thank his wife Phyllis and my cousins Susan, Sarah and Julie and their spouses and children for their friendship and support.

When a builder wants to show somebody his life's work, he can take the person to one of his buildings and point out his accomplishments. However, when a teacher wants to show somebody his life's work, he can only point to his students. It is incumbent on us to realize that every single brick within us was put there by our parents and teachers, especially during our

formative years. I want to thank all of my teachers from nursery through high school by name for their warm and complete dedication to us, their concern for all of our needs, their meticulous attention to every detail in the classroom and their love of the subject matter, so that each child could learn and develop to his or her maximum.

I owe much to my nursery teachers at Rodfei Zedek in Chicago, where my hearing problem was first diagnosed by the attentive teachers at the school, They noticed I would often turn up the volume on the record player. (I was a long-time expert on record players by age 4.)

Before entering 1st grade, my family moved to Skokie, Illinois, where I attended the Hillel Torah North Suburban Day School under the guidance of the Principal, Rabbi Bruckensein, and Assistant Principals, Rev. Noah Wolf and Rev. Van Lewin. I would like to thank all my teachers from grades 1-6, and hope I can recall each of their names. Mrs. Huss, Mrs. Glass, Mrs. Betzalel, Mrs. Sandman, Mr. Cohen, Mrs. Katzman, Mrs. Berlman, Mrs. Golub, Mr. Betzalel, Mrs. Eskin, Mrs. Rosenzweig, Rabbi Kushner, Rabbi Michael Myers, Rabbi Irwin Pollack, Mrs. Shamir, Mrs. Barak, Mrs. Goldberg, Mrs. DiPrizio, Mrs. Shore, Mrs. Wagner, Mrs. Epstein—loved teaching music, and Mr. Mishkin. Thanks to my long-time friend Alan Allswang for help with the names.

I had numerous friends from that time, but the families that I remember the best and have maintained friendship with are the Loeb, Karp, Allswang and Socol families, paragons of kindness and scholarship. Mrs. Sarane Loeb was full of life and enthusiasm, despite a crippling illness for much of her life. Danny Loeb, Ph.D. (MIT, 1989), provided some interesting insights on uniqueness for some of the work in Chapter 4. A warm sense of community was engendered by the kind members of Congregation Or Torah, led by Rabbi Eliezer Berkovits.

Before entering 7th grade, my family moved to West Hempstead, NY, and I enrolled at the Hebrew Academy of Nassau County, where I studied until high school graduation. Under the guidance of the Dean, Rabbi Meyer Fendel—a model of warmth and ethics, and the Principals, Mrs. Sally Riemer, Rabbi Moshe Gottesman and Rabbi Schonbrun, and the Secretary, Mrs. Ruth Provda, I was privileged to continue my education with their dedicated and warm staff. I would like to thank my teachers Rabbi Binyomin Schubert, Mrs. Aaronson, Mrs. Dinner, Mrs. Strumpf, Mr. Glassman, Seymour Silbermintz of blessed memory—famous choir director with many recordings, Rabbi Jacob Heisler, Rabbi Lain, Rabbi Jacob Wehl of blessed memory—who never wasted a single moment from his Talmudic studies and writings, Mr. Schwell, Rabbi Mordechai Besser, Prof. Berlinger—taught FORTRAN including computer graphics, Mrs. Faila, Mr. Joseph Disimone—possibly the funniest math teacher who ever lived—(fail early, avoid the June rush), Rabbi Shlomo Wahrman—noted author and one of the world's greatest Talmudic scholars, Mrs. Judith Lynch, Mr. Hammer, Mr. Rosen—one of the most brilliant physics teachers, Mrs. Lipskin, Mrs. Rozenzweig, Rabbi Kenny Davis and Hillel Lichtenstein. Thanks to my old friend Everett Goldin for helping me with the names.

I am fortunate to have numerous lifelong friends from my later childhood in West Hempstead, whom I speak to regularly, till this day. The one who has been a constant source of humor and cheerfulness, and who insists on regular get-togethers is Aaron Berger. My study partners and close friends from elementary and high school, Duv Fendel, Robbie Bechhofer, Menachem Gold, Shragai Botwinick, Reuven Halpern, Charlie Rudansky, David Feinberg and Marc Rosenbloom have all gone on to great achievement. I ended up marrying Marc's cousin. The

Zivotofsky family provided much friendship and guidance over the years, due to their expertise in many areas of science and engineering, and application to Judaic law. The Young Israel of West Hempstead synagogue warmly welcomed our family, and we derived much guidance from leaders Rabbi Sholom Gold and Rabbi Yehudah Kelemer, and youth directors Yussie Weiser and Mel David. Later we also benefited from the West Hempstead Synagogue led by Rabbi Baras and his very kind and outgoing family.

Our cheerful neighbors, the Rayman, Kanowitz, Falk, Simpson, Sudwerts, Finkel, Weinberg, Baruch, Silberberg, Reinberg—ham radio experts, Derech Chaim, Kirzner, Brody, Moskowitz, Finkelstein, Rabin, Schondorf, Einhorn, Feder and extended Goldstein families—cousins of noted auditory researcher Julius Goldstein, made life enjoyable in the various places we have lived.

This thesis would not have been possible without the concerted efforts of Prof. Bill Peake of the Department of EE and CS at MIT, and the Eaton-Peabody Lab. When I first explored graduate options, my family and I were on vacation in Boston over Memorial Day. I was given the name and number of Bill Peake by my supervisor, Dr. Shyam Khanna of the Fowler Memorial Lab at Columbia University. Although it was a holiday weekend, Prof. Peake warmly invited me to come immediately to his office where we spoke for quite a while. He also set up appointments with others at MIT. Because of his efforts and the efforts of Prof. Nelson Kiang (thanks also to Nelson for the book *The Blind Doctor*, perhaps the best book I have ever read) who is truly a multi-talented individual, a scholar of immense proportions in many diverse fields and the founder and first Director of the Speech and Hearing Biosciences and Technology program, I was admitted that fall. Over the years, they continued to give much individual guidance, and taught inspiring courses within the program.

My advisors, Tom Quatieri and Gert Cauwenberghs, both noted authors, gave of their time and vast knowledge in audio signal processing and related fields. They were patient enough to understand the broad aim, and to see the project through until fruition, despite the interminably slow pace of long-running iterative algorithms. They granted flexibility to try almost anything I wanted, despite the fact that there was often little to go on other than raw intuition. I am deeply gratified that I was able in some measure to return their investment, and eventually flesh out many of the core ideas with concrete results and theoretical backing. They saw that although the way was very rough, and the methods unconventional, I was genuinely striving for elegance, not a kludge that was cobbled together, and they had enough faith to let me persevere.

I was fortunate to have on my committee Profs. Louis Braida and Bertrand Delgutte who are the current co-directors of the program. They both have extremely broad and in-depth knowledge of many facets of the field, including biological, physical and mathematical modeling and the theory of electronic devices. I was honored to have Prof. George Zweig serve as a reader, with his world-class expertise in physics (one of the two scientists who originally posited the existence of quarks) and in mathematical modeling of the auditory system.

I was extremely fortunate to have as a classmate, Leonid Litvak, who became one of my closest friends. His brilliance and understanding of so many complex issues made me look forward to our numerous discussions, where we would exchange ideas and define and hammer out many possible approaches to a difficult problem. He has a clarity of thought that is rare, and an ability

to hone into the crux of the matter. Aside from his vast theoretical knowledge, he is an expert in so many different aspects of the practical side of doing science.

The fellows in Gert's Lab at Johns Hopkins were extremely helpful and knowledgeable. Special thanks to Marc Cohen, Milutin Stanacevic, Shantanu Chakrabartty, Roman Genov and Yunbin Deng for all their help and friendship. My labmates at Lincoln Lab, Nick Malyska, Daryush Mehta and Dave Messing provided much assistance with many good ideas and coding tips in difficult situations. Thanks to Nick for numerous late rides home after the last van left.

Despite all the hardships, the Harvard-MIT experience has no parallel—the expertise, the labs, the demos, the atmosphere, the bone-crunching work. Hopefully, I was able to absorb something from the experts, but more importantly I observed their way of thinking and approaching a problem, and in breaking down difficult subject matter so it flows easily from first principles.

During our time in Boston, many families helped us out in so many ways and with so many difficult situations. Our close friends, the Bachrach family, helped us acclimate to our new surroundings. We are deeply indebted to our children's schools, the Chabad New England Hebrew Academy run by Rabbi and Mrs. Ciment, an exceptionally warm environment which radiates sunshine, Torah Academy with its excellent and devoted teachers, and Maimonides Day School with its outstanding facilities. The Young Israel of Brookline provided a warm sense of community, and treated our family with true kindness. Rabbi Gershon Gewirtz, the spiritual leader, allowed us to regularly share Torah thoughts with the community, from which we greatly benefited. I have fond memories of the Talmud study group that met after Bobby Wolff's 7 am Sat. service. Participants included Prof. Nelson Lande, Marvin Weiner, Dr. Yehuda Cern, Andy Ledewitz, Alan Pollack, Boston Globe columnist Jeff Jacoby, and Ethan Berger. My children and I gained a great deal from the father-son programs at the Boston Kollel, headed by Rabbis Bier, Eisenstein and Leff, and the family programs at the congregations of Rabbis Moskowitz, Horowitz, Rabinowitz, and Schafer, and R. Wolosow's Camp Gan Israel. Thanks to my early-morning and late-night study partners, Rabbis Grossman, Miller and Donner.

The Wasserman, Nussbaum, Rosenbaum, Sundel and Levenberg families provided enormous help with our kids, as did our loving babysitters, Bella Gutnick, and Lyudmila Leytman.

Thanks to Seth Dank (Contemporary Hearing, NY), Lee Mark (Professional Hearing, MA), and Steve and Yvonne Diaz (Earmold Concepts, FL) for their valiant efforts in maximizing my hearing. Thanks to Drs. Jonathan Coleman and Robert Motzer and staff at Memorial Sloan Kettering for successfully getting me through a very difficult time (good as new).

We mention this very long list because there are so many people who assist a person his entire life through teaching, friendship, and in many other ways, and one can't accomplish anything on his own. We apologize if any names were inadvertently left out, and note that this list could actually run into the thousands.

I especially want to thank my devoted wife Shira (nee Greenberg). She has worked extremely hard both in and out of the home to support us, raise our family, run the household and also to assist in the community at large. She served as Sisterhood President of the Young Israel of Brookline for many years, for which we were honored at the annual dinner in 1994. She was willing to allow me to achieve my potential, despite numerous hardships along the way.

Her parents, Rabbi Dr. Stanley and Annette Greenberg passed away in 2002, and would have been very proud to be at the graduation. They both strongly valued education, and supported us in many ways. They loved their grandchildren dearly. It is sad that they have missed out on many of the children's accomplishments. May their memory be for a blessing.

Thanks to her siblings and families, Ron and Linda Greenberg (and Ari Zalman and Annette), Naomi and Michael Levi (and Yehuda, Ahuvah, Yoni and Chana), and Debbie and Reuven Addi (and Avigayil) for their help and closeness.

May we merit to see our children Yehudis Michal, Akiva Moshe Elimelech, Naftali Shlomo Mordechai, and Raizel Faiga Bracha grow to be a source of pride to us and the community at large, with upright character, kindness of heart and great achievement.

I would be remiss without thanking this wonderful country we live in, the good old USA, for providing limitless opportunities for education and growth. The freedom to pursue any calling, the kindness and fairness with which all citizens are treated, and the care that the government takes of the less fortunate creates a positive environment conducive to entrepreneurship and the creation and development of new ideas and technology at the highest levels. This winning formula is quite possibly the explanation for so much of the success our country has seen in so many diverse sectors, and the reason why places like Harvard and MIT and so many other excellent institutions of higher learning and of trade and commerce were founded and continue to flourish in this country. May the USA remain strong and free, and with the help of its servicemen, it citizens, its scientists and its leaders, we will overcome the serious challenges we have faced in the past few years, and continue to make the world a better place for all.

Thanks to the MIT, Harvard, Northeastern, Lincoln Lab, Hewlett-Woodmere, Hofstra and Stony Brook libraries for generously allowing me the use of their fine facilities.

Finally, I wish to thank the Almighty for his kindness to me, and for all of the preceding. May He continue to enable us to accomplish all that we desire.

Chapter 1

Introduction

1.1 Overview

The source separation problem of which the multiple talker situation is one example has proved formidable despite the wide attention and sizable literature it has generated. Fueling interest are the many conceivable applications which could benefit from this type of technology. Better performance of speech recognition and identification systems in noise, recovery of a speaker of interest from a single channel tape recording of simultaneous speakers, transcription of a musical score, improved resistance of critical communications systems to jamming, and recovery of sonar signals from background clutter are all applications in which an audio signal of interest must be isolated from the others.

We cite from the back cover of (Divenyi, 2005)

> "The cocktail party effect—the ability to focus on one voice in a sea of noises—is a highly sophisticated skill that is usually effortless to listeners, but largely impossible for machines. Investigating and unraveling this capacity spans numerous fields including psychology, physiology, engineering and computer science."

To this we could add acoustics, biochemistry, neuroscience and applied mathematics. While at first, it may appear that biology is an inexact science compared to the hard sciences of physics and chemistry, in reality, the inexactness is often in our inability to properly model the behavior and interactions amongst the unfathomable number of components which comprise many biological systems; each component, nonetheless, obeying precise and exact laws of chemistry and physics. In this work, we attempt to set forth a few organizing principles and to share our

way of thinking about the problem which can possibly be of use to the next generation of researchers in unraveling the mystery of source separation.

The speech enhancement problem can be broadly categorized into two forms. The first case is where the interfering signal has a well-defined structure, such as in a competing speech signal or other signal possessing some regularity. The second case is where the interfering signal is random or noiselike in structure. The focus of our work is the former case, where assumptions on distinctive properties of the signal of interest versus those of the background may lead to separability, although extension to the latter case may be possible, and we do some evaluations of our systems in random noise, as well.

1.2 Biological Approaches to Source Separation

In general, there are various cues available to a listener to segregate sounds according to their probable source. An important cue is the spatial relationship among the sounds in the listening environment. If certain sounds appear to be emanating from one direction, and others from a second direction, that likely indicates that there are multiple sources, and could be used as a basis of separation. However, in general, determining the location of a sound source is a binaural task, requiring two ears. The sound level and time of arrival differ slightly between the right and left ear depending on the source direction, and these differences are used by the brain to compute the location of the source. In this work we concentrate primarily on cues that can be used in a monaural, or single channel system where spatial information is unavailable. In addition, the binaural separation problem is itself rather complicated despite the availability of additional cues, and certain techniques that we will develop for the monaural case may prove useful for the binaural case, as well. To see why this might be so, consider that if two sources coming from different locations have overlapping frequency spectra, one will not be readily able to discern time of arrival and level differences from a simple comparison of the waveforms measured at the right and left ears, as they will not simply be scaled or delayed versions of each other, as would be the case for a single source. At the very least, further processing will be needed to dissect the component mixtures within the two channels, and only then can comparisons of level and time delay be made. Possibly, some of the methods we will develop

for analysis of waveforms in general, may be of use for this initial binaural processing step, as well.

Another cue for grouping sounds is pitch. Many sounds, including vowels in speech and the notes of various types of musical instruments such as the string and wind instruments consist of sets of frequency components which are harmonically related to a common fundamental frequency (i.e., are integer multiples of the fundamental). One could then classify all components that form a common pitch as belonging to the same source, and distinct from other sources of differing pitch. One difficulty with this approach is the fact that many types of sounds are aperiodic and do not exhibit a harmonic relationship between their components. These include unvoiced consonants, the percussion instruments, and many other kinds of noiselike sources. In addition, even in cases where the sources are perfectly harmonic in nature, there are other difficulties which will be discussed in Chapter 2.

1.3 Concept of Comodulation

The separation approach upon which we focus involves tracking comodulated components. Comodulation refers to the property that for a given source, there are likely to be relationships among its spectral components, such that they will start/stop at the same time and will rise/fall in amplitude and increase/decrease in frequency at the same rate. These are due to the physical mechanisms of generation of the original sound. Evidence that the auditory system is sensitive to these types of cues in some form or another, is abundant in the literature. The extent to which common frequency modulation is useful in source separation was explored by (McAdams, 1984) (Bregman, 1990) and others. Others have looked at the effect of mistuned (Lin and Hartmann, 1998) or mistimed harmonics (Darwin and Ciocca, 1992), and found that if certain limits are exceeded, then they are perceived as separate sources.

The simplest way to harness the property of comodulation would be to pass sounds through a filter bank similar to the human cochlea, and look for those frequency bands whose envelopes appear to be related. These are likely to emanate from an identical source, and to be distinct from envelopes produced by other sources. The difficulty with such a scheme is that in those frequency bands which contain energy from more than one source, the envelopes will be the resultant of the amplitudes of all sources containing energy in those bands. This makes a direct

comparison of band envelopes misleading. For example, if a source which is rising in amplitude and a source which is falling in amplitude both contain energy in a particular band, the envelope of that band may appear flat, and then be misinterpreted as belonging to a nonexistent third source, since it is dissimilar both to those bands whose envelopes are rising and to those whose envelopes are falling.

Even worse, if the frequencies of two sources overlapping within a band do not precisely match, or are not perfectly coherent with respect to each other, then interference will occur, causing the envelope of the band to fluctuate, and making determination of the modulation characteristics of either source very difficult. For these reasons, many source separation algorithms include as a caveat in their description that there must be a certain minimum frequency separation between the harmonics of the respective sources.

Still another difficulty is how to determine whether a rise or fall in the magnitude of the envelope of a source is caused by true amplitude modulation, or rather by frequency variations which cause the signal to sweep along the passband of the filter and to be attenuated differently at different points in time, in accordance with the frequency response of that particular filter, thus making the output appear to vary in amplitude.

For the case of perfectly coherent, constant-frequency, amplitude-modulated sources, we have formulated a set of matrix equations that describe the sound mixture. We prove that under certain conditions there exists a unique factorization which isolates the contribution of each source, and we develop algorithms for finding that solution when it exists. We graphically demonstrate these on recordings of musical instruments.

From the form of the solutions to our matrix equations, it turns out that having unique onset times is a major factor in separability. This corresponds well to what is known about the cochlear nucleus, which is the first auditory processing center in the brain. The cochlear nucleus contains cells called onset detectors which signal the beginning of a stimulus. We speculate that their outputs may be input to a network of coincidence detectors (cross-correlators) which will compare and group bands which share similar onset/offset times and rise/fall characteristics. The cells in this network might be synchronized by means of clocking signals produced by another group of cochlear nucleus cells, called choppers, which provide rhythmic timing pulses.

To handle the more general case of incoherent and frequency-varying sources, we have developed methods for tracking instantaneous frequency, phase and amplitude of mixtures of sinusoids. These methods operate under the premise that by looking at the behavior of a signal in multiple overlapping and closely spaced bands, we can get an unambiguous view of the underlying components of the signal, something we cannot do from a single band alone. By dissecting each band output into its constituent sinusoids, we avoid the previous problems mentioned. Instead of looking at band outputs, we look at what set of sinusoids likely gave rise to those band outputs. Each sinusoid can then be examined for amplitude or frequency variation individually, without contamination from other nearby components. Furthermore, we liberate each sinusoid from the coloring effects of the filters, thus providing a pristine view of the true modulation characteristics of each component. We will see that these methods are sensitive enough to resolve individual sidebands of speech harmonics, giving a discrete line spectrum instead of the usual spread of energy across a range of frequencies.

The general goal is to find a reversible mapping between source components and filter bank outputs. Given a set of filter outputs in filter space, we want to find a transformation back to source component space. We would also like to know what is the minimum number of filters necessary to achieve this. Finally we would like to prove that this mapping is in fact unique, and that no other set of sources can have the same image in filter space. While work on some of these issues is not complete, we believe that we have enough supporting evidence to validate the general framework.

We note that there are some philosophical difficulties in terms of how to define a component. For example, an amplitude-ramped constant-frequency sinusoid is probably perceived by a listener as a single-pitch pure tone of increasing loudness. However, Fourier analysis suggests that there are actually multiple, constant-amplitude, constant-frequency sinusoids combining in particular phase relationships to produce that waveshape. There is energy in upper and lower sidebands surrounding the original sinusoid. Should these sidebands be considered components, as well? These types of thought problems serve to confuse matters, as it is certainly more intuitive to define a component as a single sinusoid with time-varying parameters, rather than multiple closely spaced sinusoids with constant parameters. We live in a world of change. It seems unnatural to resolve changing signals into sets of constant components. We might

initially try to distinguish between frequencies that appear to arise from purely mathematical constructs, as opposed to those that can actually be related to the physical mechanisms of sound generation, such as the natural modes of vibration of the source in question. We might consider the latter to be the true spectral components of a signal, with the instantaneous amplitude described by a modulation term corresponding to the strength of the excitation at each time. Instantaneous frequency could, likewise, be defined as the frequency of those modes corresponding to the instantaneous shape of the source configuration. This avoids the previous ambiguity, at least conceptually, but we will find that there is probably a more accurate answer when we analyze test results on modulated signals in Chapter 6.

We note that we can't escape the Fourier viewpoint entirely, since it is essential for the description of the linear systems and filter banks that we will use in our processing, but we propose that it is possible to improve on its shortcomings in describing the experiences of a changing world by using additional methods of analysis to obtain more precise and intuitively useful information.

The basic approach used in this suite of algorithms is to construct filter banks with overlapping exponential frequency response shapes. The reason for this choice will become clear later. By comparing temporal features of individual channel outputs which will vary from filter to filter due to the different weighting of frequency components, we can merge the various views into a composite picture of the individual sinusoids comprising the audio mixture. Appropriate subsets of sinusoids from this mixture can then be grouped on the basis of a choice of possible criteria into separate audio streams to complete the separation process.

Our long term goal is to create a parallel, biologically plausible implementation of our algorithm, both for its use as a model of neural processing, and for the increased computational efficiency that this would provide. We look broadly to the auditory system for guidance in attacking this difficult problem, but we can do no more than guess at the many possible avenues of complex processing within its myriad pathways which might reasonably agree with physiological data. Nevertheless, when we see parallels in our way of thinking with physiologically plausible mechanisms, we will feel free to note the comparisons.

The most important temporal features for the purpose of this work are the local maxima or peaks of the waveforms of the individual band outputs. We have concentrated on this feature, because it is a clear landmark that stands out within the complex twists and turns of an auditory signal. It also has the advantage of providing amplitude as well as frequency and phase information. This is in contrast to zero-crossings, for example, which may provide frequency and phase information under limited conditions (Logan Jr, 1977), but do not readily provide amplitude information. In addition, from looking at data from actual neural responses in the auditory nerve, we believe that identification of local maxima is a key mission of early auditory processing. It is well accepted that the probability of a spike firing is related to the instantaneous amplitude of the wave cycle, with increasing probability as one approaches the crest of the waveform. This is referred to as phase locking or spike synchrony (Johnson, 1980). However, visual examination of the data seems to indicate that at the actual peak, there are more spikes than can be accounted for by the sound level alone. The period histogram seems to become very pointed at the instant of the peak due to the disproportionate number of spikes at that point. Figure 1 which is taken from (Popper and Fay, 1992) illustrates the peakedness of the neural response. Figure 2 from (Geisler, 1998) illustrates phase locking in the case of a mixture of sinusoids. The response seems to track peaks of the sum.

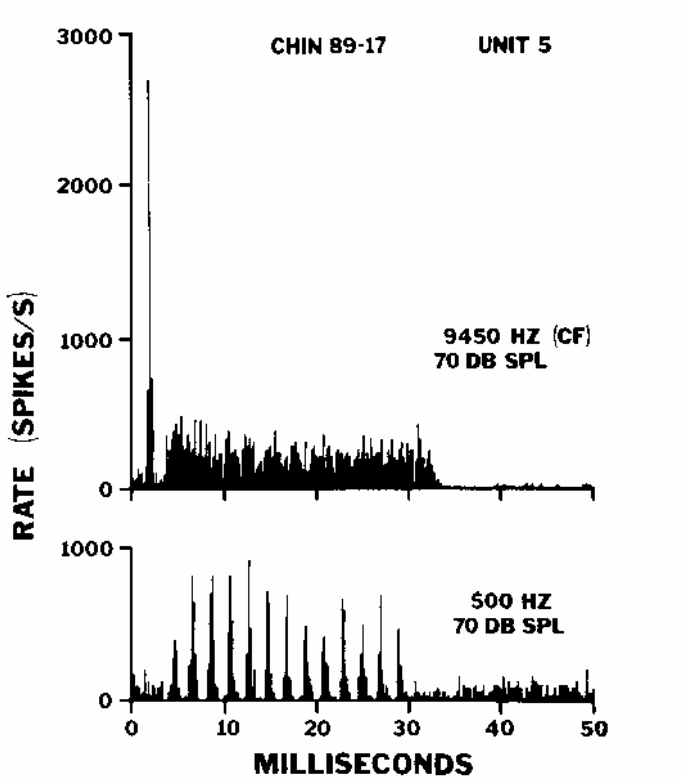

Figure 1. The increased neural response at peaks in the input waveform. Figure shows poststimulus time histograms for the reponses of a single chinchilla cochlear neuron to 9450 Hz (top) and 500 Hz (bottom) tone pips. Stimuli had rise and fall times of 4 ms, lasted 25 ms, and were presented every 150 ms. Histograms are average of responses to 200 stimulus repetitions. The neuron's CF was 9450 Hz and it had a CF threshold of 3 dB SPL and spontaneous activity of 59 spikes/s. Data from (Shivapuja, 1991) reproduced in (Popper and Fay, 1992).

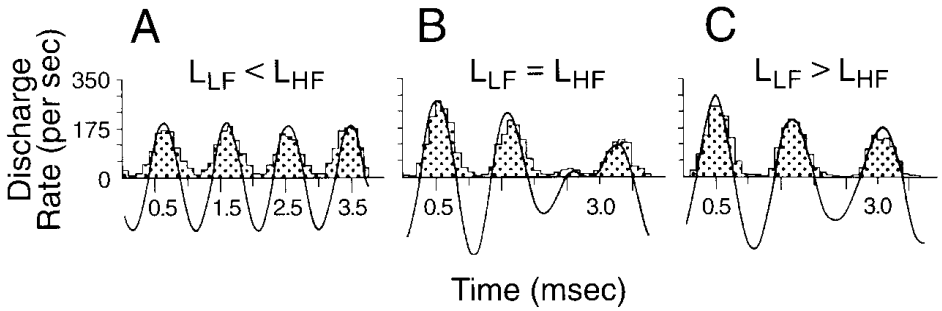

Figure 2. Neural response to a mixture of sinusoids appears to be synchronized with local maxima of resultant. Responses to two-tone stimuli were recorded from a primary afferent neuron in the squirrel monkey. The two frequencies—798 Hz (LF) and 1064 Hz (HF)—had a 3:4 ratio. When the sound level of the high-frequency tone was greater (by 20 dB), the spikes virtually synchronized to that tone alone (A). When the low-frequency tone was more intense (by 10 dB), the spikes largely synchronized to that tone (C). When the tones were of equal strength, the spikes synchronized strongly to both tones simultaneously (B). Each period histogram (plotted on the time base of the 266 Hz fundamental frequency) is fitted with a curve that is the sum of the two sine waves, arbitrarily adjusted in phases and amplitudes to achieve the best fit of its top (positive) half with the data. From (Brugge, Anderson, Hind *et al*, 1969) reproduced in (Geisler, 1998).

We also use for motivation a personal observation of this author that dynamic range seems to be extremely important for intelligibility of speech in noise. If transmission through hearing aids is impeded, even through an ill-fitting earmold, it becomes much more difficult to hear clearly in noisy situations. This is despite the fact that one can turn up the volume of the aid to compensate for the reduced loudness. The explanation seems to be that increasing the volume (multiplicative gain) causes increased peak clipping, since the maximum power output (MPO) is fixed. The loss of peaks may be detrimental to separation tasks. The author found strong support for this notion from an experiment performed by (Young Jr and Goodman, 1977). They wanted to test the notion that if speech were mixed with stronger noise (competing speech), it might be possible to improve intelligibility by clipping peaks of the mixture, thus equalizing the energies of the two. In fact, they found that the opposite was true; the intelligibility of the desired speaker was severely degraded by this type of processing. This is exactly as expected from this author's own experience.

1.4 Thesis Organization

This thesis is organized as follows: In Chapter 2 we review and categorize some of the work done by other researchers in the field of source separation over the years. There is so much literature that we can't possibly cover or even be familiar with all of it, but we attempt to include representative papers from some of the major directions that have been pursued. Chapter 3 presents an introduction to the concept of comodulation, and examines the extent to which various types of sound sources may be considered to be comodulated. It also contains examples of situations in which the use of multiple channels helps to unravel ambiguous situations, such as separating the contribution of AM and FM to the trajectory of a signal component. Chapter 4 describes a matrix approach for separation of amplitude-comodulated, constant-frequency musical signals under certain constraints, using the method of Non-Negative Matrix Factorization. Chapter 5 describes work on separation of sinusoids suitable for the more general case of frequency-modulated and amplitude-modulated signals. It contains three algorithms for signal separation all of which seek to achieve the same goal of instantaneous parameter estimation using local maxima of multiple bands, but which accomplish this by different methods. Representative results are shown for each, and the strengths and weaknesses of each are discussed. Chapter 6 describes tests on actual AM and FM

signals including noise and speech signals and raises some surprising issues in interpreting the results. Based on this we further discuss the question as to the conceptual definition of instantaneous parameters, and how they can be of use in source separation. Chapter 7 discusses an analytical approach towards computing a closed-form solution to the problem of parameter estimation using local maxima of multiple bands. Since the algorithms of Chapter 5 are numerical in nature, work is needed to build a mathematical foundation upon which these algorithms can rest. The primary question is the uniqueness question, i.e., are there other combinations of signals that would have the same distribution of local maxima across multiple bands. Chapter 8 is a forward-looking summary and conclusion and contains suggestions for future work.

Chapter 2

Review of Previous Work

2.1 General Categories

A number of approaches have been developed for identifying the number of signals in an environment and separating them according to their probable source of origin. These can be categorized according to the particular feature of the sound sources which drives the separation engine.

As we have seen in Chapter 1, separation schemes can be broadly categorized as either monaural or binaural depending on the number of channels used to record the sources. Since in many situations only a single channel is available for analysis, we will concentrate primarily on monaural cues. In addition, we believe that understanding the monaural case is fundamental to the design of any source separation system. We therefore mention briefly some of the multi-microphone approaches that have been studied, but devote the bulk of this review to the monaural literature.

We note that the focus of our work is the separation of audio sources. However, the methods we develop ultimately lead back to the problem of separating closely spaced pure sinusoids. We therefore include relevant material on both of these interrelated topics.

We note that many of the key papers related to auditory perception and source separation of both speech and music are extremely well-reviewed in three recent theses by (A. L.-C. Wang, 1994), (Ellis, 1996) and (Scheirer, 2000) which we will summarize as part of Section 2.4. Between these three works, one has a veritable cornucopia of knowledge from which to acquire a strong grounding in the classical literature. We therefore concentrate most of our effort on papers which have appeared since that time.

2.2 Sound Direction

The appearance of sounds from multiple locations indicates the existence of multiple sources, and classification of sources according to acoustical properties dependent on direction can serve as a basis for source separation. In general, determining the location of a sound source is a multi-dimensional task, requiring multiple microphones. The sound level and time of arrival differ slightly between the sounds picked up by each microphone depending on the source direction, and these differences may be used to compute the location of the source.

2.2.1 Auditory Beamforming

One method which makes use of these phase differences is auditory beamforming, which enhances sounds from one direction at the expense of sounds from other directions. Some of the earliest work in this area borrowed from theory developed for radio antennas. Early work on these systems was published by (Widrow, Glover, McCool *et al*, 1975), (Griffiths and Jim, 1982) and (Haykin, 1986).

The basic principle upon which these methods work is the use of multiple microphones to receive the sound at different locations which are offset from each other by some fixed distance. The sound, therefore, reaches the various microphones at different times, leading to phase differences between the signals at each microphone. In order to combine the signals, one applies the corresponding phase shift of each microphone as would be present in a signal from the desired direction so that all inputs are in phase with each other. If one then averages the inputs from each microphone, the desired signal will be in phase in all of the channels, but signals from other, random directions will be out of phase with each other, thus tending to cancel.

A multiple microphone array hearing aid worn on the chest was demonstrated by Bernard Widrow from Stanford University at the 2001 meeting of the Acoustical Society of America with encouraging results (Widrow, 2001). The device consisted of a rather large plastic case which was shaped to fit comfortably on a user's chest. The largest dimension measured about 8 or 10 inches. It was anchored by a loop worn around the neck which also served as an induction coil to couple the signal to the T-coil of subject's hearing aids. A key attribute was that it made weak or distant sources that were within its angle of reception sound as if they were much closer than they actually were. Low whispers in a very crowded room became easily audible.

The major disadvantage of this particular system, was that it should have been developed to be worn on the head, as a headset. This would allow one to turn his head to face the desired source, which is more natural and convenient than turning one's body. In addition, because the inductive loop was relatively far from the aids it introduced noise and attenuation, effectively reducing usable dynamic range. A head mounted system would allow one to inductively couple the signal from much closer range, or alternatively to directly couple the signal via the hearing aid's accessory shoe, providing a lower noise route.

Another general disadvantage of these types of systems is the loss of binaural information of the usual kind, and the loss of all auditory stimulation from the back and sides. (Greenberg and Zurek, 1992) have tried to address these types of deficiencies by partitioning the spectrum in such a way that part is used to give directional gain, and part to give binaural information. Other issues have been studied by (Peterson, 1989), (Peterson, Wei, Rabinowitz *et al*, 1990) and others such as figuring the optimal number of microphones which should be used in these arrays.

In general, these systems may be classified as fixed or adaptive, depending on whether the weighting of the microphones is adjusted in real-time. The advantage of adaptive systems is that they attempt to compensate for changes in the acoustic environment. However, reverberation tends to confuse these algorithms dramatically.

2.2.2 Independent Component Analysis

Similarly, another multiple microphone approach which has gained favor in the last decade is Independent Component Analysis (ICA) first developed by (Bell and Sejnowski, 1995), which in its most common form, attempts to use statistical similarities and differences between channels to sort out a mixture of n signals recorded by n sensors. The linear combinations of the various signals picked up by each sensor are all represented by a single mixing matrix. The basic premise is that due to the Central Limit Theorem, distributions of mixtures should have a more Gaussian nature than distributions of separate sources. One therefore looks for the inverse (or unmixing) matrix which produces the most non-Gaussian set of sources. Various criteria are used by different authors to make this decision. Current algorithms do not permit movement between the subjects and the various microphones (which would make it impossible for a single mixing matrix to describe), and also do not perform well in reverberant environments. A good

introduction is provided by (Hyvarinen and Oja, 2000). In an appendix we will have more to say about ICA, and the similarities and differences with our work.

For the remainder of the chapter we review monaural methods which are the main focus of the thesis for reasons described earlier.

2.3 Harmonic Relationships

An early approach to the separation problem used pitch as the primary cue, as we briefly mentioned in Chapter 1. This method has been explored by a number of researchers with varying degrees of success. (Shields, 1970) proposed comb-filtering the target speech around its harmonic energy. (Lim, Oppenheim and Braida, 1978) implemented an adaptive comb-filtering system, whereby a signal is added to a delayed version of itself, causing the resultant frequency response to become periodic in frequency. (Parsons, 1976) described a pitch tracking algorithm which tracked two speakers simultaneously by grouping spectral components according to sets of harmonic relationships. First, the harmonics were determined from the FFT using a peak-picking method. Then the fundamentals were calculated, using a scheme suggested by (Schroeder, 1968) in which all possible submultiples of each harmonic are computed, and a histogram is made. Those bins which have the most values are fundamentals of harmonic sets. Harmonically related sets sharing one fundamental frequency can then be separated from sets sharing another fundamental. The method assumes that the fundamentals are far enough apart to be separable under the resolution limitations of the FFT. Among the difficulties facing these methods is the fact that it can be difficult to identify competing sources due to the crossing of pitch tracks. Furthermore, small errors in pitch estimation lead to high levels of distortion of the output signal (Nishi and Ando, 1998).

2.4 Computational Auditory Scene Analysis

These approaches use various relationships between spectral components in attempting to group and segregate them according to probable source. Important contributions include both perceptual studies to identify cues which are likely used by humans in auditory scene analysis, and computational algorithms that attempt to perform source separation using these cues. The following paragraphs review work of both types.

2.4.1 A Survey

(McAdams, 1984) synthesized a set of 10 harmonics of a 220 Hz fundamental, with the resulting sound strongly resembling an oboe. He then applied 10% vibrato at 4 Hz to the even harmonics, and kept the odd harmonics constant. He found that listeners perceived two instruments. One was a clarinet-like instrument with a pitch of 220 Hz. The other was a soprano-like sound with pitch of 440 Hz. He also found that sets of harmonics in which all components were modulated coherently sounded more fused than ones in which the components were modulated incoherently.

This author performed some similar informal experiments on harmonic sets in which the even harmonics and odd harmonics were amplitude-modulated with different modulating frequencies and phases. He, too, found that the percept of separate sources was obtained, despite the harmonic relationships among all components. This was even stronger if the onset times were different for the even and odd harmonics.

The conclusion from perceptual experiments of this type appears to be that both amplitude and frequency comodulation play an important role in neural auditory processing.

(Bregman, 1990) conducted many experiments on listeners in an attempt to understand which aspects of a signal are important in identifying the number and types of sources present. As an example, he studied a situation in which frequency tracks approach each other and then move away. He wanted to know whether listeners perceive them as crossing over and continuing, or as bouncing and receding from each other. Bregman's work is interesting as far as identifying potentially useful cues in auditory scene analysis, but as a psychologist, he does not provide algorithms for harnessing these cues.

(Cooke, 1991) and (Brown, 1992) looked at spectral components that follow related trajectories in a time-frequency space, and attempted to separate those with dissimilar trajectories. However, in cases where the signals overlap, the trajectories become difficult to follow and to resolve. In a broad sense, our work is related to the ideas in their work, but as we have stressed, the use of features from an auditory band as a whole without further attempts to get at the underlying signals that give rise to those features may easily be misleading.

Avery Wang (A. L.-C. Wang, 1994) used Harmonic Locked Loops which are linked Phase Locked Loops that are tuned to harmonics of a common fundamental, in order to separate FM signals from mixtures. These methods are based on recognition of the fact that there is often frequency comodulation among components of a signal, and hence changes to frequency are reflected proportionately across all harmonics. His method required *a priori* information on the initial fundamental frequency of the signal in question, but could track changes, thereafter, to the fundamental and harmonics. As his methods preserve phase, they allow for separation by subtraction of that harmonic set from the remaining components.

Wang's thesis, in general, contains an excellent discussion on the differences between parametric methods and nonparametric methods of spectral analysis, with many examples of each. His knowledge of advanced mathematical spectral analysis techniques is quite impressive. He defines Fourier analysis as non-parametric, and contrasts with parametric methods which attempt to directly solve for the frequency variable ω in fitting data to an equation of the type $a\sin(\omega t+\phi)$. As we will see, the methods we will develop in utilizing the local maxima of waveforms should be classified as parametric, and this will become evident when we look at the actual equations. However, we note that mathematical methods in general have a way of being consistent across disciplines, and although Fourier analysis may appear to be based on a set of equations which is removed from the appearance of the original signal, however, since those equations are derived from the orthogonality properties of trigonometric functions, and since these follow from the properties of trigonometric integrals, which in turn follow from trigonometric derivatives which follow from trigonometric limits and finally from the basic addition formulas, Fourier methods are actually grounded in the same place. This will be seen later in Chapter 6, when we discuss the dichotomy between the instantaneous view of signals and the combinatorial view of infinite length, constant components. Basic trigonometric identities can give the same results as Fourier analysis.

In short, we believe that the shortcoming of Fourier analysis is not that it is non-parametric, but rather in the fact that it takes a long signal-length to give sufficient frequency resolution, and in the fact that the length of the window influences whether orthogonality holds. If the window is not perfectly matched to the signal frequency, then the orthogonality condition is not fulfilled, as the integral of trigonometric functions over an incomplete cycle of a waveform is not limited

to either *1* or *0*, as orthogonality requires, but will depend on the start point and endpoint of the interval.

Dan Ellis's thesis (Ellis, 1996) contains much analysis of Bregman's work, in addition to that of many others. He distinguishes between data-driven and prediction-driven source separation. These are alternately referred to in the literature as bottom-up and top-down processing, respectively. The data-driven school of thought holds that to do separation, one must first operate on the raw data at the acoustical and signal-processing levels to produce separate audio streams, and these are then passed up to the higher brain centers where they can each be interpreted or ignored, at will. The prediction-driven approach which he advocates holds that the brain doesn't need to separate the entire stream, but rather it looks for elements which it can recognize in the mixture.

Ellis's computational scheme involved breaking up a scene into various sound elements such as periodic components, noise components, and burst-like components. Based upon what is likely to be the upcoming sound type, a hypothesis is continually reevaluated about the origin of each. The argument for this type of system is that humans are known to be able to fill in gaps and missing syllables which are covered by noise. In the opinion of this school of thought it is the extraction of understanding from the mixture that is the key, and not the actual separation of waveforms.

We would like to respectfully disagree with this idea, and side with the data-driven school. The reason is that it would seem that there need to be clear templates upon which the brain can operate in order for it to be able to recognize information from one source or the other. A jumbled and mixed audio stream will have features that differ from those of either source. It seems to us unlikely that the brain's pattern recognition engines can recognize anything familiar in a mixture until it is separated and sorted. We can, however, see the opposing point of view in an instance where certain harmonics of the voice are covered up by noise, while certain others are not. In that situation, possibly the brain may recognize enough information from the clean harmonics to be able to piece together the ones that are covered. We can accept that perhaps there is enough redundancy in speech to decipher a message with only a portion of the harmonics free from interference. This would seem to be true only if there were enough unaffected harmonics to make out the general formant structure. Alternately, if the noise is

variable so that at certain time instants only the desired source is heard, even though noise frequently punctuates at other time instants, we can agree that in this case, as well, it may be possible to piece together enough information without formal separation to understand the message.

Eric Scheirer's thesis (Scheirer, 2000) does not deal directly with source separation, but concerns itself with extracting key information from a musical recording that describes the character of the piece.

In his words:

> "New models are presented that explain the perception of musical tempo, the perceived segmentation of sound scenes into multiple auditory images, and the extraction of musical features from complex musical sounds. These models are implemented as signal-processing and pattern-recognition computer programs, using the principle of *understanding without separation*."

From his encyclopedic knowledge of the auditory processing literature, going back to many of the classical papers in the field, there is one particular point which we consider extremely relevant to our work. We cite the following:

> "(Summerfield, Lea and Marshall, 1990) drew an explicit contrast between "conjoint" grouping strategies, in which energy from each correlation channel is split up and assigned to several sources, and "disjoint" strategies, in which the channels themselves are partitioned between channels [*sic*?, B.D.J.]. Their method was a disjoint method; they do not provide psychoacoustic evidence for this decision, but base it on the grounds of physical acoustics ("when sounds with peaked spectra are mixed, energy from one or other source generally dominates each channel.") (Bregman, 1990) argued for a disjoint model, which he called the principle of *exclusive allocation*."

In our opinion this paragraph goes to the heart of our entire effort. We believe that there is no basis to the disjoint model, and that it is the major stumbling block in designing a successful source separation system. There is no way to short-circuit the absolute requirement to analyze individual bands and dissect them, correctly allocating energy to each source. The difficulty in

doing this is what has prompted many approaches to explicitly warn against situations in which pitch tracks become closer than about 25 Hz. Our methods attempt to confront this issue head on and to correctly perform this allocation.

We note that Scheirer's review of Ellis's work mentions the fact that both top-down and bottom up approaches are used in his approach to interpreting audio scenes, and not exclusively the top down approach, as we described earlier.

A further interesting point raised by Scheirer is in the interpretation of the McAdams oboe experiment.

> "At one time, there was general agreement that the auditory grouping for these sort of stimuli was governed by coherent frequency modulation. This is the explanation promoted by McAdams in his presentation of experimental results using these stimuli. However, this agreement no longer maintains; in particular, (Carlyon, 1991; 1994) has argued on the basis of more extensive psychophysical testing that the actual basis of auditory grouping in these stimuli is the harmonicity of the signal. That is, as the even harmonics move away from exact harmonicity with the odd harmonics, a pitch-based grouping mechanism selects them as part of a different auditory group.
>
> "The present [Scheirer's auditory-segregation] model does adhere to the viewpoint that grouping is based on common modulation in these stimuli. Future work should examine more closely the stimuli developed by Carlyon and others to distinguish the harmonicity hypothesis from the common modulation hypothesis."

In 2003 an invitational workshop was held in Montreal, Canada to assemble many of the researchers at the forefront of the Speech Separation field. There were 20 presenters, and a book edited by (Divenyi, 2005) was compiled containing expanded articles by these researchers. The book contains a rich collection of work on the current state of the field. We note, however, that a number of those approaches are multichannel in nature, and hence outside the scope of this work. The remainder of this subsection contains short summaries of relevant papers in Divenyi's book.

Claude Alain (Alain, 2005) focused on the importance of mistuned harmonics in source separation on the basis of physiological evidence from brain wave recordings showing sensitivity to such occurrences. In general, experiments on mistuned and mistimed harmonics are among the body of perceptual evidence that indicates the importance of proper grouping in source separation.

Peter Cariani (Cariani, 2005) discussed neural timing networks for extracting precise temporal information which could possibly be used for separating periodic sounds from noiselike sounds, and which would also explain the sensitivity to mistimed harmonics. He cites (Kubovy, 1981) who demonstrated that abrupt changes in phase and/or amplitude of a harmonic can cause it to "pop out" perceptually from a mixture. We note that it is has been a motivating factor in our work that although phase seems to be unimportant in the perception of a single sound, it may play a major role in separation, as we will discuss in Chapter 4.

Te-Won Lee (T.-W. Lee, 2005) described recent efforts to adapt the ICA formulation developed for multiple channels to single channel use. We mention this, as our work relates to single channel processing. We have made such an attempt ourselves to adapt ICA to this problem by considering the envelope of each band to be analogous to a single microphone. We describe further in an appendix.

Paris Smaragdis (Smaragdis, 2005) discussed the concept of redundancy reduction as a unifying and motivating factor in the development of ICA. We note that our methods make extensive use of redundancy, both with regard to the fact that speech and music contain multiple harmonics, and with regard to the consolidation of similar information from multiple channels.

Sam Roweis (Roweis, 2005) described the use of Hidden Markov Models (HMM's) for grouping similar harmonics. He, too, emphasized the idea of redundancy, and included as an example a concept that we discussed earlier in the context of Ellis's work, that a speech message may be understood on the basis of a subset of the original harmonics, if the remainder are obliterated by noise.

Richard Stern (Stern, 2005) discussed the problem of classifying which areas of a spectrogram might be missing or corrupted by noise, and how to adapt processing methods to ignore or reconstruct these missing features. His approaches are both single and multi-channel. In single

channel approaches, he discussed the abundant psychophysical evidence that the trajectories of harmonics are extremely important in grouping, including minor fluctuations "micromodulations" in frequency and amplitude. He cited the experiments of Chowning (unreferenced) and Bregman to bolster this. This is at the core of our work on comodulation. We note that so many researchers have recognized the importance of common modulation in one form or another, however, we emphasize that it is extremely difficult to detect due to the interfering effects of sources on each other. Overcoming this problem is the focus of the major part of our work.

DeLiang Wang (D. L. Wang, 2005) discussed the issue of goals and performance measures for CASA systems. What are reasonable standards, and how should we evaluate performance? He mentioned that the goal of actually generating two separate source streams may be unrealistic, but suggested an alternate approach called the Ideal Binary Mask. The idea is that given a spectrogram or other Time-Frequency representation, the aim is to label each cell with a *1* if it is likely to come from the source of interest, or a *0* if it is likely to come from noise or an interfering signal. Even if reconstruction is not attempted or possible, an accurate mask should be considered a worthy goal. This will be relevant for our method of Chapter 4, where we perform a graphical separation of sources. Wang mentioned work by (Weintraub, 1985), and further extended by (Brown and Cooke, 1994) and (D. L. Wang and Brown, 1999) who have devised methods for actual reconstruction based on a successful binary mask if gammatone filter banks are used for the original T-F distribution. He also cited work by (Slaney, Naar and Lyon, 1994) on methods for reconstructing phase from spectrograms and correlograms.

Wang also discussed an evaluation method based on quantifying the extent to which the score of an automated speech recognition system is improved by the use of the CASA system in question. Since the difficulty ASR systems have with noise is one of the motivating factors in developing CASA systems, in general, this might be a natural benchmark of performance.

Wang listed some alternate goals to which various researchers have aspired, including the study of biological mechanisms, and successful modeling of neurobiological data. To an extent, our work attempts to do this, as well, as we have tried to explain various features of auditory neurophysiological data, including the shape of auditory filters, and the function of temporal phase-locking.

Malcolm Slaney (Slaney, 2005) argued strongly against the idea that the brain can separate sounds into separate streams. He brought evidence from a number of interesting phenomena, including.

1) The ability of a native speaker of a language to fill in a part of a word obscured by a cough, something the non-native speaker cannot do.

2) The McGurk effect in which visual perception of the speaker's lips affects the listener's auditory perception (McGurk and MacDonald, 1976).

3) Experiments in which audio cues cause visual motion perception.

He also described the history and use of the correlogram, popular with many researchers and featured prominently in the works of Ellis and Scheirer, earlier, which is a simultaneous plot of the auto-correlation of each channel, resembling a spectrogram in certain respects, but with enhanced temporal information. The time between maxima of a row corresponds to the period of the underlying channel output. He discussed the possibility of separating a correlogram of a mixture into two partial correlograms, and then further reconstructing the waveforms of each. He briefly addressed the issue of how to compute the proper phase for each channel in the course of this reconstruction. However, he believes, as before, that all this is not strictly necessary, but rather that the brain can decode the message without reconstructing the streams.

2.4.2 Contrasts with Visual Scene Analysis

We note that there are some basic differences between auditory scene analysis and visual scene analysis. In the most common visual scene analysis situation, one is presented with various objects in a scene, in many different orientations, distances and sizes, some of which are partially obscured by other objects. The task is to identify the objects in the scene, even though only a partial view may be available for each. While this is not a trivial problem, it does differ from auditory scene analysis in certain aspects that make it less difficult by comparison, in our opinion. In VSA, each pixel in the image generally comes from only one object. The task is to figure out where the borders lie between the objects, and to recognize the whole from the part. One does not need to decompose pixels into separate objects, with a certain percentage of energy allocated to one or another. In ASA, on the other hand, one has multiple sources

superimposed on each other, and it is much harder to recognize what percentage of energy comes from each source, and where each source begins and ends in time. An exact visual analogy to ASA would occur in the uncommon situation in which one is analyzing a picture taken as a double or multiple exposure. Most often, this kind of image occurs only in error, and is discarded. The analysis of blurred images may also bear some similarity to the ASA problem in that energy from multiple pixels must be properly allocated, but this involves a deconvolution operation, and differs from the separation of additive sources.

2.5 Approaches based on Spectral Estimation

Spectral Estimation includes short-time Fourier magnitude and phase-based approaches introduced by (Parsons, 1975), (Hanson and Wong, 1984), and further developed by (Naylor and Boll, 1987) and others. For reasons we will explain shortly, these approaches typically are limited in relying on strict stationarity conditions, the interferer to be much larger (e.g., 6-16 dB larger) than the target speech, and/or parametric models of the speech spectrum.

In a broad sense, these approaches may be seen as a modified form of spectral subtraction. Spectral subtraction is a technique which was originally developed to separate speech from random noise. If the noise spectrum could be estimated, then one could seemingly subtract it and be left with the speech alone. Two well known problems with this method are that noise changes over time, so that a past snapshot of the noise spectrum quickly becomes outdated; and that many spectral subtraction algorithms ignore phase information. If one misestimates or uses outdated information about the noise spectrum, one will be attenuating the wrong frequencies and corrupting the speech. It is for this reason that stationarity is required, the noise spectrum is assumed not to change. If one misestimates phase, one may in fact be adding noise, rather than subtracting. Symptoms of these types of errors include "musical noise" in the reconstructed waveform, in which spectral energy has been added that was nonexistent originally.

In order to extend the concept of spectral subtraction to the multiple talker problem and to treat a competing speaker as noise, one must contend with the additional difficulty of estimating the interferer's spectrum at each interval due to the nonstationarity of speech. The counterintuitive requirement mentioned above that the interferer be stronger, not weaker than the desired (target) speaker is due to the fact that the better we can estimate the interferer's spectrum, the

more accurately we will be able to subtract the unwanted energy and leave the target speech unaltered. Hanson and Wong concluded that magnitude estimation alone of the harmonics from the interfering speech spectrum was adequate to reconstruct intelligible speech, even though phase estimates were obtained from the noisy speech mixture as a whole. While their estimates of the interfering speech spectrum were separately obtained (using *a priori* knowledge of the interfering spectrum, clearly not useful in realistic situations), their key contribution was that the process of harmonic magnitude suppression was sufficient to preserve intelligibility.

Naylor's extension uses the assumption of harmonicity combined with robust pitch estimation to identify those regions of the spectrum that will likely have energy based on the fundamental frequency. He tested four different methods of pitch estimation: 1) Cepstral Estimation (Noll, 1967); 2) Maximum Likelihood Estimation (Wise, Caprio and Parks, 1976); 3) Harmonic Matching (Lim and Griffin, 1985); and 4) the Auditory Synchrony Model (Seneff, 1984). The authors found that best results were obtained using the Maximum Likelihood estimator. In cases where the interfering speech was controlled to be steady state, they report good separation, even when estimation of the interfering speech spectrum is performed on the actual mixture. One point which they note in their findings is that for real speech, harmonic lines are not exact integral multiples, which is similar to what we have found in tests of our methods on actual speech in Chapter 6.

2.6 Sinusoidal Modeling

(Quatieri and Danisewicz, 1990) used a sinusoidal representation of the speech signal to separate closely spaced frequency components by using a least squares solution to resolve the combined envelope formed by the two interfering speakers. The speech waveforms of two talkers are modeled as harmonic sets. Each is broken up into frames of 20-30 ms long. These frames are each multiplied by a Hann window. Since the speech is assumed to be a sum of sinusoids, the net effect in the frequency domain is a sum of scaled and shifted transforms of the window sequence. The value at each frequency component yields the amplitude and phase for any corresponding harmonic in that frame. Interpolation is used to match up values across frame intervals, with appropriate phase unwrapping to account for time evolution. The authors demonstrate that if all frequencies are known *a priori*, then one can solve a set of linear

equations to separate overlapping window transforms, and obtain separate amplitude and phase parameters for each component. In addition, even if the two fundamentals alone are known, one can still perform the separation using the assumption of harmonicity to obtain the frequencies of higher order harmonics in each set. In the case where no *a priori* frequency information was available at all, the algorithm used a gradient descent method to minimize the error between the summed reconstructed waveforms and the original signal using the two unknown fundamental frequencies as parameters.

In all cases, if any frequencies become closer than about 25 Hz from each other, then the matrix becomes ill-conditioned, and an analytic solution cannot be found for that frame. This could occur when pitch tracks cross, or even if a lower order harmonic of one speaker happens to lie near a higher order harmonic of the second speaker. To overcome this, the authors suggest using the previous or next frame as a guide to the trajectory of the pitch tracks so that current frequency locations can be estimated. In the latter case where use was not made of *a priori* frequency information, results were satisfactory only where pitch tracks were non-overlapping and where the sound level of the desired speaker was approximately equal to that of the interferer. Results were better for the first two cases. Because this method bears some similarity to our approach which will be presented later, we have gone to greater lengths in our description.

2.7 Coherence-Based Approach

(Cauwenberghs, 1999) used a wavelet formulation to exploit characteristic jitter among source components as a basis for source separation. This led to an iterative time-domain correlation-based search algorithm. The underlying assumption is that individual sources will never be perfectly coherent with respect to each other.

2.8 Super-resolution Methods for Sine Estimation

Because of the limits of classical Fourier-based methods due to the time/frequency resolution tradeoff, alternative methods have been developed to shorten the data-lengths and recording times required for satisfactory parameter estimation. The following sections review some of the modern methods for the separation of sinusoids. In virtually all of these approaches, the $n \times n$

covariance matrix formed by using a subset of n consecutive points of the data is mathematically manipulated in some fashion or appears as one of the factors in a set of linear equations which is then solved to yield the desired parameters.

2.8.1 Linear Prediction

One of the most popular methods is the use of particular assumptions on the nature of the source to attempt to find the best fit (in the least squares sense) to a particular model and to view the parameters obtained as being representative of the source. Among these are subsets in which the source is viewed as having been generated by either a train of impulses in the case of voiced speech, or by white noise in the case of unvoiced speech, which is then passed through a filter which may be one of three general types. It may be a moving average (MA) type, in which the output depends only on the current and past input samples. It may be an autoregressive (AR) filter, in which the output depends on the current input and past output values, or it may be a combination of both in which the output depends on the current input, and on past input and output values (ARMA filter). In the frequency domain, by taking z transforms of the respective difference equations, these take one of the following forms: The numerator is a polynomial in z and the denominator is a scalar, (all zero model), the numerator is a scalar and the denominator is a polynomial in z (all pole model) or both the numerator and denominator are polynomials in z (pole zero model). In general, an all pole model can well recreate sharp peaks in the spectrum, but not so well deep notches, and conversely for an all zero model, but in practice, the all pole model has gained wider acceptability. The reasons are that the mathematics is simpler and better understood, and that the lack of zeros can actually be satisfactorily compensated for by increasing the number of points (the model order or number of poles). The scalar term accounts for the overall strength of the signal.

Because the autoregressive or all pole model assumes that one can predict future values of the data from past values (with appropriate fixed weights), it is also known as Linear Prediction. It turns out that the same equations arise based on slightly different assumptions in the course of separating a sinusoid from noise using the Maximum Entropy Method (MEM), thus these methods are closely related. To determine the weights, one solves a set of linear equations called the Yule-Walker equations (also known as the normal equations) in which the covariance matrix multiplies the unknown weight vector to yield the vector of autocorrelations. While

traditional methods of solution work fine, such as Gaussian Elimination, due to various symmetries and redundancies which are properties of the autocorrelation function, there are quicker and more efficient algorithms which have been developed, among them the methods of (Levinson, 1947) and (Durbin, 1959). A byproduct of the Levinson recursion, is the computation of intermediate terms which have a physical interpretation in the course of acoustic modeling of the vocal tract as a set of tubes of differing cross sectional area.

Because the length of the data record is necessarily finite, at the beginning and end of the record, the autocorrelation values are undefined and taken as 0, as they would depend on nonexistent past or future values in each case. This could introduce inaccuracies in finding the best-fit solution. To get around this, the set of equations can be truncated by eliminating rows where there are any zero entries and then solved. This variation is known as the covariance method. In this case, the resulting equations are similar to those derived under a different set of assumptions known as Prony's method (Prony, 1795). Prony's method attempts to model a signal as a sum of exponentials, and to find the FIR filter that best cancels it. For short data records, the covariance method produces better results. If the signal is a true sum of exponentials, Prony's method produces exact results.

Increasing the model order generally increases accuracy, but can lead to false peaks in the spectrum. In an attempt to separate the effect of the signal from noise on the shape of the spectrum, (Tufts and Kumaresan, 1982) proposed the use of the Singular Value Decomposition (SVD) to separate the contribution of the higher amplitude components which are ostensibly due to the signal from the lower amplitude components which are ostensibly due to noise. By selecting those eigenvectors corresponding to the largest eigenvalues and reconstituting, one can compute a lower rank and more accurate approximation of the contribution to the inverse of the autocorrelation matrix due to the signal alone. This variation is known as Principal Component MA estimation.

2.8.2 Capon's Method

Capon's estimator (Capon, 1969) has been described as measuring the power output from a set of adaptive filters that best null the noise process (minimum variance) at the output of each filter. Whereas the periodogram uses the same filter shape for each frequency, in Capon's method the individual filters can change their shape to reduce response to energy outside the

band of interest. As previously, the autocorrelation matrix appears in the defining equation, and the spectral estimate is given by the suitably scaled reciprocal of the quadratic form of a vector containing successive harmonic frequency terms and the inverse of the autocorrelation function.

(Kay, 1988) shows that Capon's method can be shown to be equivalent to an averaging of AR estimates of all orders up to the order chosen, and the inclusion of the lower order estimates negatively impacts the resolution of the final result. Citing (Lacoss, 1971) he states that it is better than the Bartlett method (a traditional periodogram based approach), but worse than the AR method.

2.8.3 Pisarenko's Method

Pisarenko's method (Pisarenko, 1973) is based on a property of all positive-definite Toeplitz matrices, of which every covariance matrix is an example, that it can be decomposed into a sum of sinusoids plus white noise. In order to find the frequencies of these sinusoids, one computes the minimum eigenvalue and its corresponding eigenvector. The elements of this eigenvector turn out to be the coefficients of a polynomial whose roots are the frequencies of the sinusoids. The noise power is equal to the value of the minimum eigenvalue. The amplitudes of the sinusoids are then computed by back-substituting the frequency and noise terms into the equation for the covariance matrix and a simple linear system of equations results for the amplitude terms.

A drawback of this method is that it relies on perfect knowledge of the covariance matrix, while for actual data the values can only be estimated. Because of this, its accuracy suffers, and in an evaluation by Kay, he found that its performance is much worse than all the other methods of this section. Errors in the calculated frequencies occur even when there is no noise.

2.8.4 MUSIC (Multiple Signal Classification Algorithm)

In order to correct the deficiencies of Pisarenko's method, an alternate approach was suggested by (Schmidt, 1986). This method avoids the factoring step to obtain the frequencies. Instead, one makes use of the property that the signal subspace is orthogonal to the noise subspace. One again computes the eigenvalues of the covariance matrix, but rather than assuming that the noise subspace is entirely characterized by the minimum eigenvector, one chooses a threshold value above which the eigenvectors are classified as the signal subspace and below which they

are relegated to the noise subspace. One then computes an equation in which the denominator is effectively a product of the signal subspace and noise subspace. When this is minimum for a particular frequency (maximally orthogonal) the reciprocal determines the power of that frequency.

A disadvantage is that a search must be made to find the frequencies, and some *a priori* knowledge is required in terms of how many frequencies to look for, or the value of the cutoff point which determines the dimensionality of the signal subspace as opposed to the noise subspace. The overall accuracy of the method is reported by Kay as being acceptable if the SNR is above 16 dB.

2.8.5 ESPRIT (Estimation of Signal Parameters via Rotational Invariance Techniques)

While the MUSIC algorithm is considered to produce good results, as above, the required multidimensional search is computationally intensive. A method for reducing this burden was developed by (Roy and Kailath, 1989) and is highly recommended in the book and accompanying lecture notes by (Stoica and Moses, 1997) and in the lecture notes of their frequent collaborator, Juan Li. The ESPRIT method exploits symmetries in phase among time samples at earlier and later times in a period of the waveform. When applied to radar direction of arrival (DOA) problems, it requires the use of symmetrical arrays with a known spacing. In either case, it requires *a priori* knowledge or an educated guess as to the number of sources d. An eigen-decomposition is performed on the combined arrays, and the columns corresponding to the d largest eigenvalues are retained. This is then partitioned into two submatrices, one containing all rows except for the last, and the second containing all rows except for the first. Because of the properties of the particular symmetry between the arrays which is effectively a rotational invariance, it turns out that the best fit of the data to the impinging signals is given by the total least squares (TLS) solution. As explained by the authors, in usual least squares calculations of $\mathbf{Ax} = \mathbf{b}$, the matrix $\mathbf{A}$ is considered fixed, and the best fit vector $\mathbf{x}$ is calculated. However, in this case, there may be noise with equal probability in either of the two submatrices, therefore the method of TLS is employed which finds the best fit to the data with the option of adjusting either. The net result is equivalent to matrix multiplication of the first submatrix by the inverse of the second. The eigenvalues of this TLS solution matrix are then

complex and their phase angles can be mapped to the digital frequencies of the sources. The computational burden is much lower than for MUSIC, as the frequencies arise through an analytic approach, and not through maximization via an exhaustive search over one or more simultaneous parameters.

2.8.6 Comparison

For a superb head-to-head comparison of the results of eleven spectral analysis methods on the same data set in both easy to read tabular and graphical formats, one should consult (Kay and Marple Jr, 1981). They used three sinusoids and a region of colored noise. One immediately sees that FFT based methods can only resolve frequencies further apart than the reciprocal of the data set length in seconds (which we derive in Section 6.3), although two of the test sinusoids were deliberately placed closer than this. Some of the other methods were able to resolve all of the sinusoidal inputs, but only one, a variation of Prony's method, was able to correctly provide the amplitudes. This method produces results in the form of a line spectrum, as it assumes a set of sinusoids, and tries to fit the data in a least squares sense. Because of this, the noise was not well-represented, although some lines appeared in that region. Another advantage of Prony's method is that it provides phase information, while AR methods do not. His method actually bears the most resemblance to our work, but a direct comparison is difficult, since our work requires a filtering operation followed by an instantaneous analysis based on local maxima. The data length benchmarks are therefore difficult to compare. In a single filtering operation, we can analyze the frequency and amplitude progression of time-varying signals, while his method assumes constant sinusoids. In, addition the length of time necessary to resolve a signal with our method depends on its frequency, since the higher the frequency the closer the maxima. Finally, our work depends on the sampling rate, as we need to precisely locate the maxima in time, but precise specification of that relationship is difficult to quantify, and has been a matter of judgment. Further complicating the matter is our use of interpolation which artificially increases the number of data points, but introduces some noise.

The main drawback of Kay and Marple's excellent summary is that it is now dated, and additional methods have cropped up since, some of which we have mentioned. We are not aware of a similar, more current review at this time.

2.9 Recent Work on Time-Frequency Analysis

2.9.1 Norden Huang's Method

Norden Huang (Huang, Shen, Long *et al*, 1998) recently proposed a curious method for separating a signal into functional shapes which capture different characteristics of the behavior of the original signal. One forms the upper and lower envelopes of a signal by means of two sets of splines, one interpolating between the local maxima and the other interpolating between the local minima. One then computes the mean of these two curves and subtracts from the original signal. One repeats the process until (a) the residual has the same number of zero-crossings as extrema, or the numbers differ by at most 1; and (b) the mean of the upper and lower envelopes is zero, i.e., they are symmetrical about the x axis. When this occurs, one has obtained the first Intrinsic Mode Function (component) of the data. One then subtracts this and repeats the process to find the next IMF. The method differs from the work we have done, in that our work is focused on analyzing and separating sinusoidal components on the assumption that the signals of interest contain harmonic structure, such as in music and voiced speech. Huang's method does not assume this, and is actually not suited for this. His method may be useful in analyzing fluctuations in nonlinear oscillators for example. However, when it comes to separating two constant-frequency sinusoids, it fails, as he illustrates in an example. His results seem to produce two signals, one of which corresponds to the fine structure, and one to the overall envelope. But in fact, the reality is that two closely spaced sines are actually causing the pattern. In this author's opinion, this is a fundamental weakness with his method, in that it tries to fit shapes in a superficial manner while ignoring the root cause of these shapes.

In addition, the method has many *ad hoc* criteria for when to stop the sifting (subtractive iterations). Too much sifting will actually degrade the results.[1] Our methods are concerned with uncovering the underlying sinusoids that make up a complex signal, not with trying to fit the shape with various splines or processes based on them, and the like.

[1] This author finds the fact that Huang's results occasionally degrade with increasing iterations to be a major drawback, as it should be axiomatic that all results of iterative algorithms are only valid at convergence. One should not arbitrarily report results from earlier iterations, if later iterations become unsatisfactory or if convergence is not achieved.

2.9.2 The Reassigned Spectrogram

(Fulop and Fitz, 2006) describe a method which has been proposed, forgotten, and reinvented a number of times in the literature, but which has not received widespread popularity, possibly due to confusion about correct implementation, proper interpretation of results, and possible advantages. They collected three versions of the algorithm and attempted to compare results against each other and against conventional spectrograms.

The basis of the method is that possibly better precision along the frequency axis, and better alignment along the time axis can be obtained by moving various time-frequency points (cells) to other locations in the spectrogram. The basis for doing this is to compute the instantaneous frequency for each channel (channelized instantaneous frequency or CIF) and the group delay at each time (localized group delay or LGD) according to the following two equations, respectively:

$$\tilde{\omega} = \frac{\partial}{\partial t}\arg[X(\omega,t)]$$
$$\tilde{t} = t - \frac{\partial}{\partial \omega}\arg[X(\omega,t)] \tag{2.1}$$

If, for example, in a given channel the CIF calculation indicates that a certain frequency is present which differs from the center frequency of the filter for that channel (i.e., the transform of the window sequence shifted to the center frequency of that channel), then the cell is relocated to the new position given by the CIF, and not the center frequency of the band. Similarly, if the group delay at a given time shows the occurrence of a particular frequency at a time value other than the normal reference point at the center of the window, then the cell is relocated to that time.

It would help if the authors included additional elucidation of the motivation behind this alternate system of referencing frequency and time. Do these shifts occur over distances far enough so as to place the new location outside the range of the original window? If so, how does one reconcile the apparent presence of energy within the passband of the original filter with the claim that this energy is actually located at some other frequency bin which is calculated by the CIF? Similarly, if temporal adjustments are allowed that are large enough to shift energy to a completely different windowed segment of the signal, how does one explain

the appearance of energy in the original segment, if there was none at that time? Using the CIF and LGD to obtain refined estimates within a single time-frequency tile would seem to be a useful innovation that agrees with intuition, but using them to rearrange tiles would seem to benefit from further explanation.

In addition, the authors need to make the assumption that only one source has energy in a given band, otherwise the instantaneous frequency will be invalid. As we have discussed, this goes against our experience that overlap within a band is common.

These questions notwithstanding, the authors do raise a number of good points in their discussion of the shortcomings of conventional analysis, especially with regard to the "smearing" of modulated signals, and how the use of additional processing may clarify matters. In this general manner, their goal is similar to ours, however, the techniques we have developed are completely different.

The results do show sharper traces in certain cases than obtained with conventional spectrograms, but in other cases, they seem to fail to resolve harmonics that are visible conventionally. We will have more to say about the interpretation of instantaneous frequency in Chapter 6.

2.9.3 The Local Vector Transform

(Ito and Yano, 2007) present a method for improved pitch and amplitude tracking of mixtures of signals in the presence of nonstationarities. They begin by noting the deficiencies of conventional methods including the need for local stationarity within the analysis window, the dearth of accurate methods for determination of amplitude, and recurrent dissimilarities between the reconstructed waveforms and the originals. To improve, they use a formulation in which phases and amplitudes of component sinusoids are each expressed as 3-term Taylor series, rather than as arbitrary functions of time. The task then becomes to try to determine the Taylor coefficients of each parameter. They begin by showing that were indeed the phase and amplitude of an actual sinusoid to be given by such a 3-term sum, then the spectrum could be well-approximated by a closed-form function of the 6 Taylor parameters. These in turn could be determined from the spectrum by successive differentiation of the spectral shape with respect to angular frequency. For actual signals which are not of the form of a 3-term sum, good results

can be obtained if the phase and amplitude functions can be approximated by this form within a time segment close to the point of analysis. If variation is too rapid, then tracking errors will occur. Additionally, the power of neighboring components must be negligible at the spectral peak of the component under analysis or else interference errors will occur. This limits the utility of the algorithm to lower order speech harmonics only, since for higher order harmonics spectral spread increases to the point at which energy of one harmonic may overlap with energy of a neighboring harmonic.

This again underscores the repeated difficulty of parameter estimation in the presence of competing signals which we have seen is so often specified as a caveat in the case of many of the conventional and newer algorithms, and which our own methods are designed to overcome.

Advantages of the Local Vector Transform methods over conventional methods are the ability to calculate parameters directly without iterations, and the incorporation of phase information which provides for more accurate reconstruction.

Results shown by the authors on synthesized speech comparing the LVT method to autocorrelation methods, cepstral methods and the reassigned spectrogram of the previous section seem to show that the best frequency determination accuracy is produced by the LVT method, with the reassigned spectrogram a close second. For amplitude determination, the LVT was an order of magnitude better than the reassigned spectrogram and conventional spectral peak picking methods, with the reassigned spectrogram actually faring a bit worse than the conventional peak picking method.

2.10 Conclusion

There is a wealth of further information on source and speech separation in the literature. We have reviewed only those aspects which relate most strongly to our work. There are numerous promising routes that are being explored in many institutions, but a common unifying theme is that related modulation patterns seem to be universally acknowledged as playing a strong role in the separation process.

Our own approach to source separation shares some similarities with the general CASA approach in the fact that it attempts to look for common features among spectral components.

But, as we will see in the following chapters, in certain aspects it is more mathematically rigorous in the method in which it handles overlapping or ambiguous cases.

Chapter 3

Aspects of Comodulation

3.1 Introduction

Having introduced comodulation in Chapters 1 and 2, we now examine the concept in more detail, and explore the applicability and limits of using it as a basis for auditory source separation. In the next few sections, we look at spectrograms of various musical instruments to gauge the extent to which comodulation applies. We will then discuss certain ways in which the property of comodulation may be traced to the mechanisms of sound generation in those instruments. We will focus on both amplitude and frequency comodulation. Our source separation algorithms in Chapter 4 will depend on amplitude comodulation. An abundance of psychophysical data emphasizes the importance of frequency comodulation in source separation, as well. We discuss a further condition, which we term phase comodulation, that leads to an interesting graphical description of signals which have this property, but which we have found is not an accurate model of most realistic sound sources. We discuss why this may be the case. After that we contrast two broad approaches in using comodulation for source separation, which we term *a priori* and *a posteriori* comodulation. We then demonstrate that an assumption of comodulation can reduce certain ambiguities in the characterization of signals.

3.2 Application to Music

In the next chapter we will develop a comodulation-based algorithm for separation of musical sounds. It is therefore useful to examine some properties of musical instruments and their effect on the extent of comodulation.

3.2.1 Timbre

Musical instruments differ from each other in two primary ways. The first is in the relative weighting of the spectral components, i.e., the various overtones produced by the instrument; and the second is in the time progression of the sound. Both affect the perception or color of the note, and this quality is usually termed timbre. It is not an easily defined concept, as it does not correspond to any one physical property. Different instruments emphasize different harmonics due either to the particular sound production mechanism of the instrument, or to filtering by the resonances of the instrument. In addition, a piano will never sound like a violin, even if the harmonic ratios were the same. Since a piano string is struck suddenly, the sound has a different time evolution than does the sound of a violin which is excited by means of the more gradual bowing process, or a horn which is produced by a vibrating column of air.

The time progression of a note is often described in terms of an attack, decay, sustain, and release (ADSR). The attack is the immediate buildup to a high level of amplitude as soon as the note begins. It then decays rapidly to a smaller value, and sustains close to that value for some period of time. Finally, as the note is released, it falls back to zero with some characteristic time progression. Within the sustained portion of the note, certain instruments are often varied in amplitude to produce a more pleasing effect. This is known as tremolo. The overall amplitude behavior can be captured by a time-dependent modulation function which we will describe and use in Chapter 4.

Figure 3 through Figure 6 taken from (Higgins, 2001) show the ADSR characteristics in graphical format for the general case, and for a number of specific instruments.

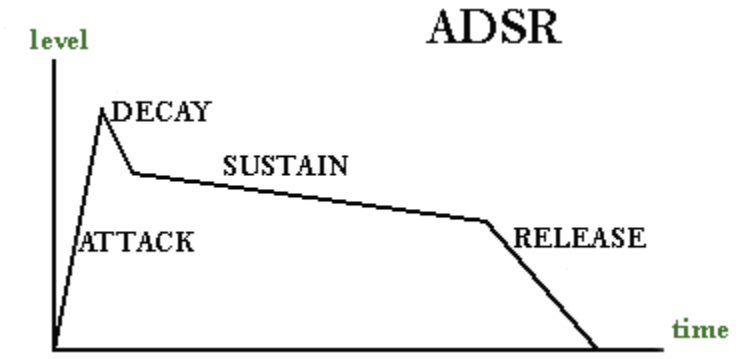

Figure 3. The four general time characteristics of a musical note: Attack, Decay, Sustain and Release. From (Higgins, 2001).

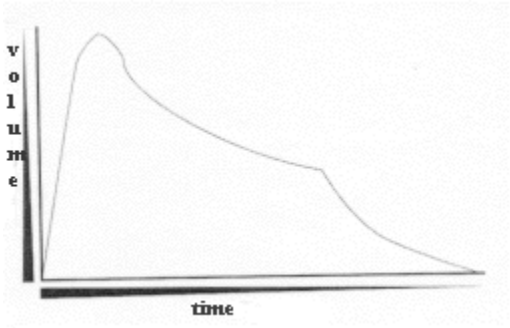

Figure 4. The envelope shape of a struck string such as in a piano. From (Higgins, 2001).

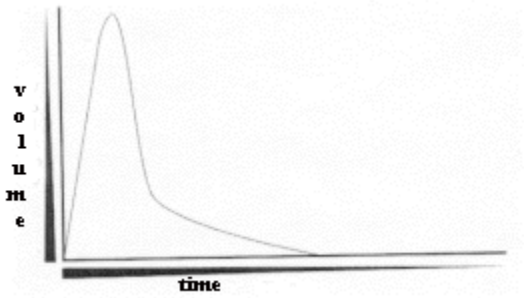

Figure 5. The envelope of a plucked string such as in a guitar. From (Higgins, 2001).]

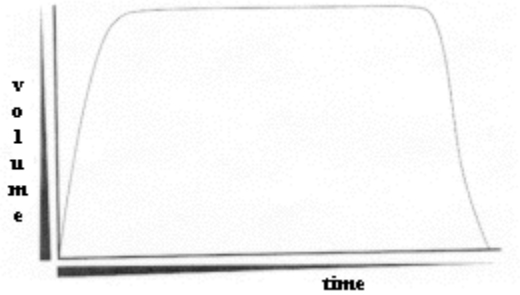

Figure 6. The envelope of a bowed string, such as in a violin. From (Higgins, 2001).

3.2.2 Sample instruments

We present a series of waveform plots and spectrograms of various instruments from the McGill University Master Samples collection. This database contains a catalogue of sounds produced by many different types of instruments with each note played by a world-class musician on the finest instrument available of that type. For each instrument, we first illustrate a waveform plot, a conventional spectrogram plot (from the top), and then two 3-D plots of the spectrograms at angles of elevation of 45 degrees. The first 3-D plot in a set shows the lower harmonics towards the front. Since these are often the strongest, and have a tendency to obscure

the weaker, higher harmonics, we then show a reverse plot with the higher harmonics towards the front. The conventional, top view generally best captures frequency modulation, while the 3-D views are more useful for viewing amplitude modulation. Spectrograms were computed using a Hann window of 30-45 ms depending on the instrument, with an overlap of 90%. Sampling rate was 11.025 KHz for the piano, and 22.050 KHz for the remaining instruments. FFT length was 1024. While these would be considered narrow-band spectrograms in speech analysis, in music the fundamental can vary over a wider range, and hence the choice of parameters that will give the best tradeoff in time and frequency resolution may need to be found with some trial and error.

We begin with the altoflute which strongly exhibits amplitude comodulation.

3.2.3 Altoflute

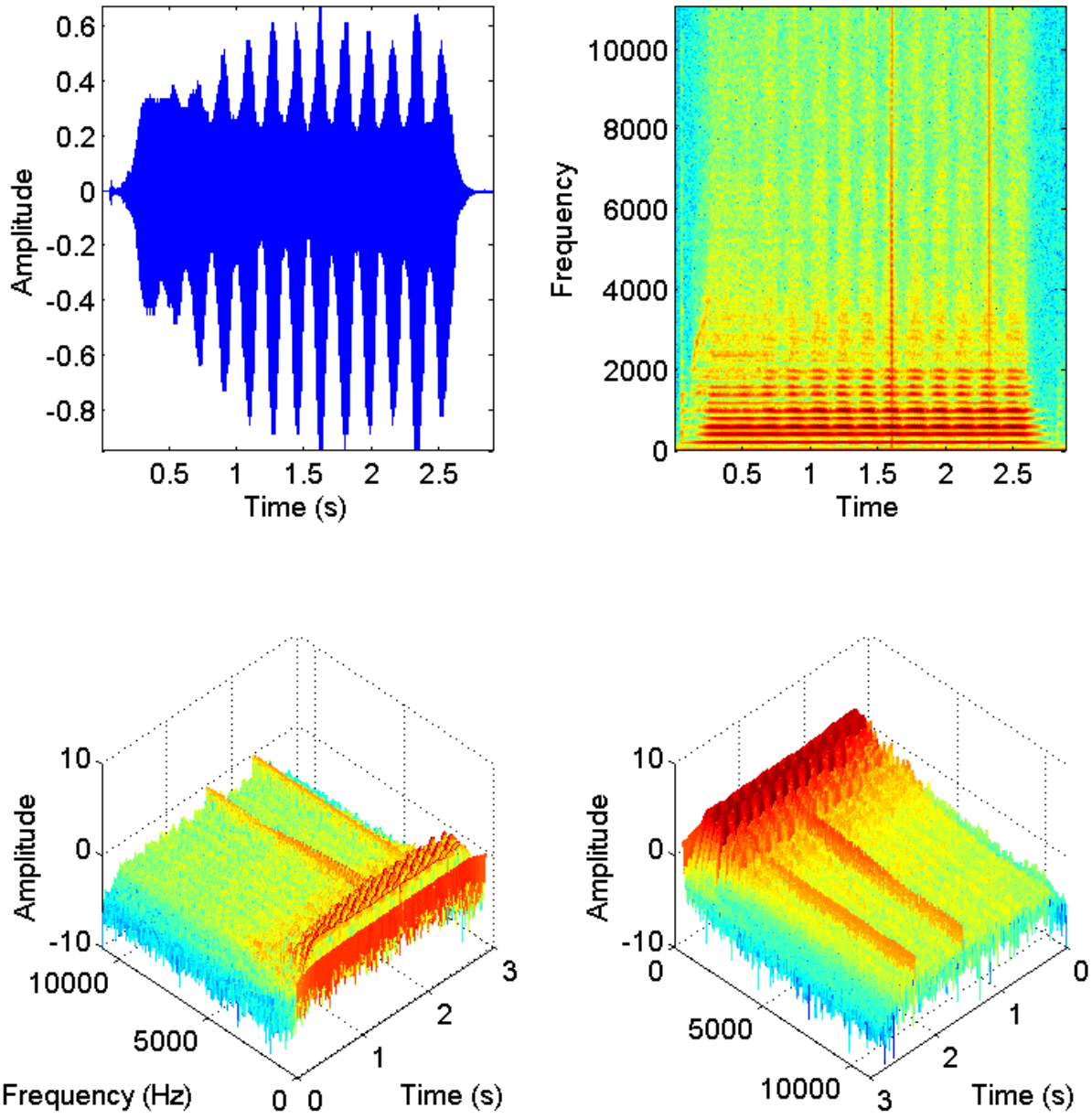

Figure 7. Top left: The waveform of an altoflute playing G3. Note the strong up and down modulation envelope which can be attributed to tremolo. Top right: Spectrogram of the altoflute. Harmonic traces are close to horizontal, indicating constant frequency. Strong evidence of amplitude comodulation in synchronous rise and fall of most harmonics (red color). Bottom left: 3-D mesh plot of spectrogram. Note rise and fall of most of the harmonic traces in unison. Bottom right: Reverse view with higher harmonics shown in front.

From the spectrogram plots of the altoflute in Figure 7, it is clear that the envelopes of the preponderance of the harmonics are strongly correlated with each other and with the envelope of the original time waveform. The possible exceptions of the first and second harmonics may

be due to the body of the instrument absorbing energy at its resonant frequencies during periods of maximal excitation, and re-emitting it at periods of minimal excitation, thus smoothing the amplitudes of those harmonics. In other words, the body of the instrument acts as an acoustic filter. Overall, amplitude comodulation well-describes the behavior of this instrument.

3.2.4 Violin

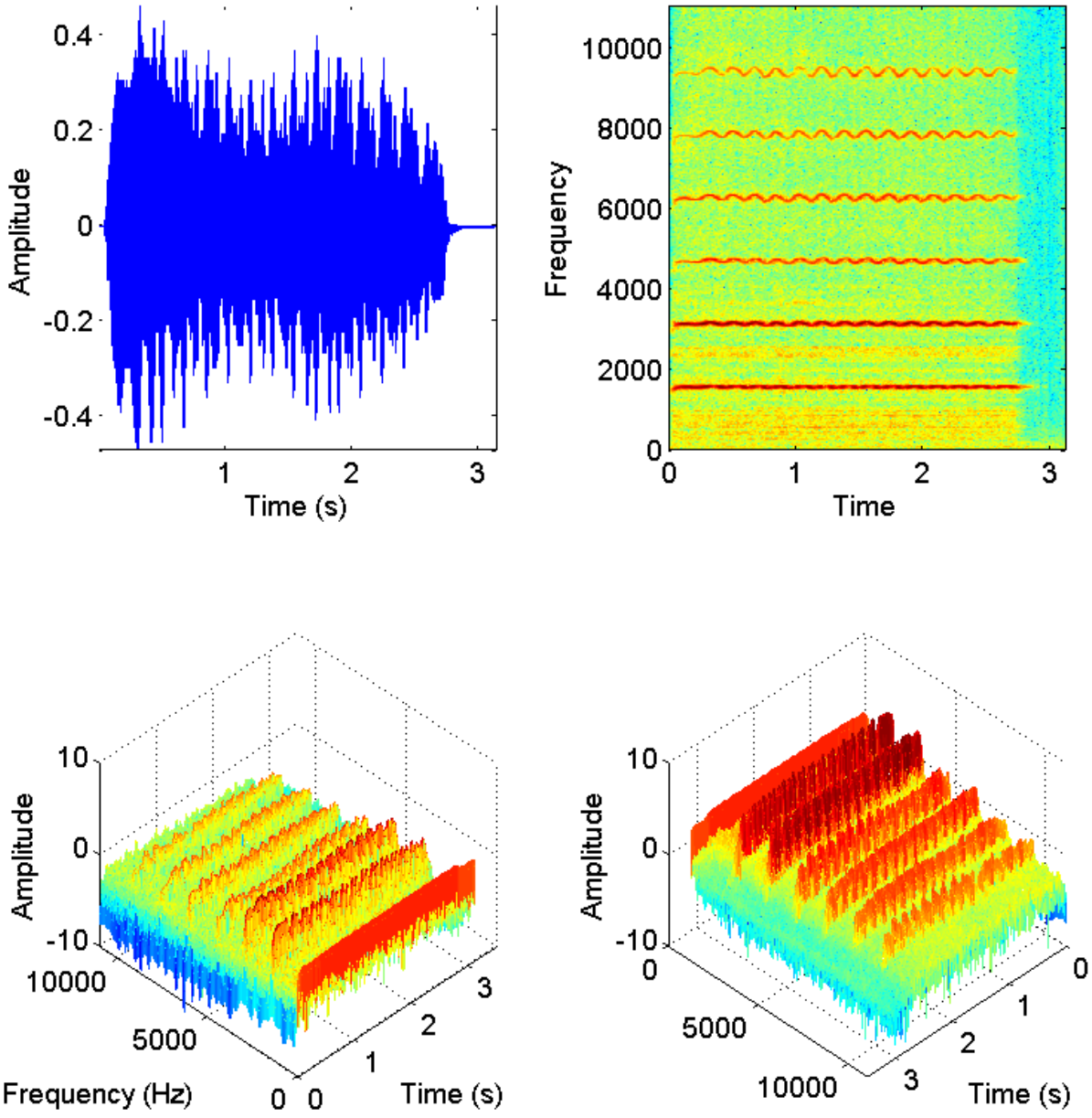

Figure 8. Top left: The time course of a violin playing G6. There are strong amplitude fluctuations in the envelope which may be related to the vibrato applied by the musician as discussed in the text. Top right: The spectrogram of this violin note clearly illustrates the effect of vibrato on the harmonics. The frequencies of all the harmonics vary almost sinusoidally in lock step. Strong frequency comodulation is apparent. Bottom left: 3-D mesh plot of spectrogram. The harmonics of the violin appear to exhibit some amplitude modulation related to the vibrato movement in addition to the more obvious frequency modulation. Bottom right: Reverse plot with higher harmonics shown in front.

The violin, shown in Figure 8, is usually played with a musical effect known as vibrato to give a richer and more pleasing sound. Vibrato is a rhythmic back and forth vibration of the finger on the keyboard at about 4-7 Hz depending on the musician's style. (In this example, we can

visually estimate the vibrato frequency as about 15 cycles in 2.8 seconds or 5.4 Hz.) This back and forth movement effectively lengthens and shortens the active part of the string, thus raising and lowering the frequency of the fundamental and all harmonics. Since, as can be seen from the spectrogram, the frequencies of all the harmonics move in lock step with the frequency of the fundamental, it is properly an example of frequency comodulation.

As an aside, we note that the percentage variation of the movement of all frequencies is the same. The higher frequencies move more than the lower frequencies. For this reason, if one desires to capture common frequency variation, one might want to select wider filters for the higher frequencies than for the lower frequencies, using a logarithmic pattern so that the variation within all filters will be similar. This type of an arrangement corresponds to a constant-Q filter bank, rather than a constant-frequency structure. It is thought that auditory filters may resemble this type of design in certain respects.

What is curious is the apparent variation in amplitude, as well, that can be seen in both the time waveform and in the 3-D mesh plots of Figure 8. The motion of the musician's finger may have a rhythmic damping effect, that may affect amplitude, as well as frequency. In addition, changes in frequency or phase may affect the relative addition and cancellation of spectral components whose magnitudes contribute to the overall time envelope.

3.2.5 Trumpet

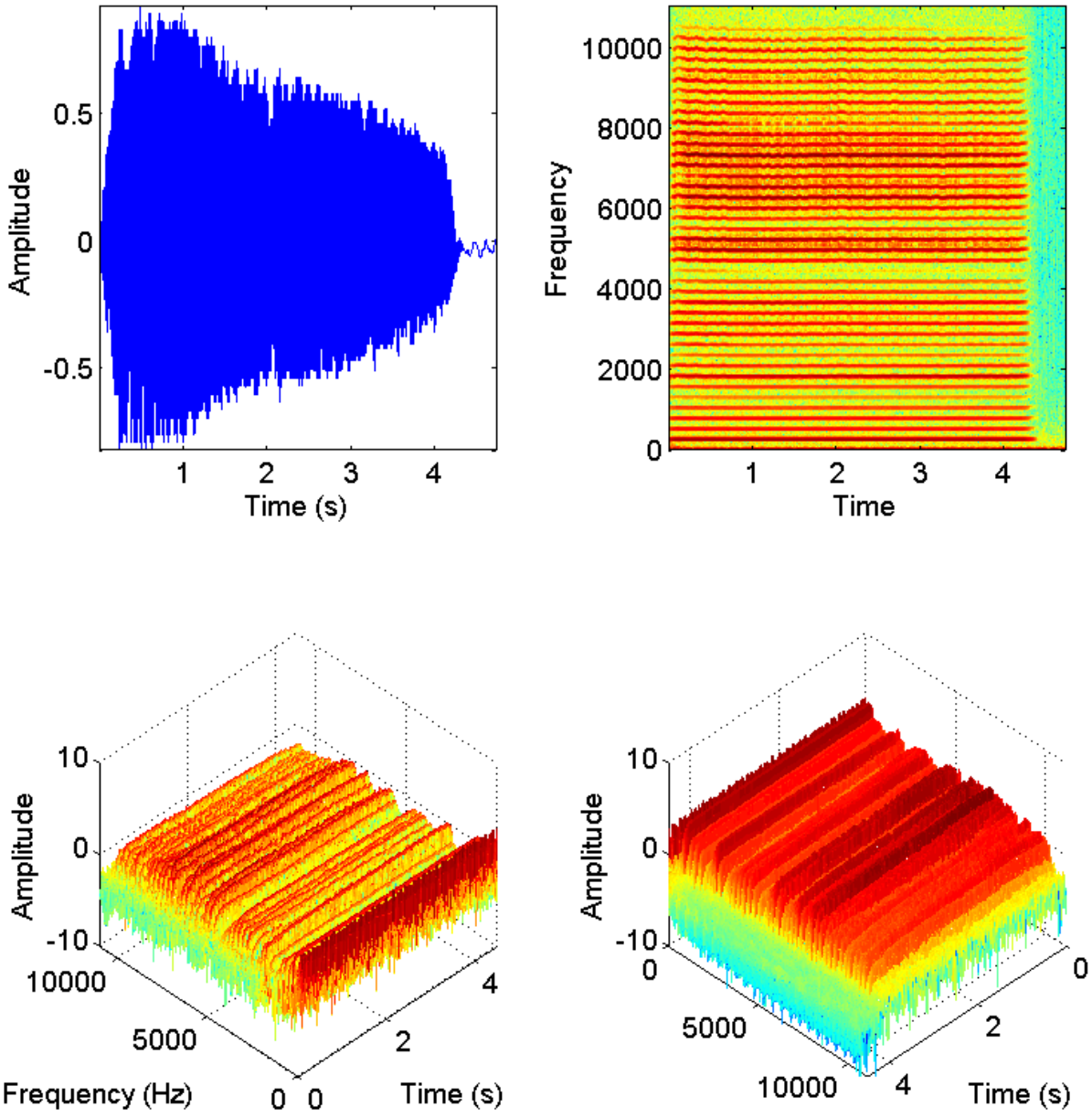

Figure 9. Top left: The time waveform of a muted trumpet playing C4. The envelope shows a gradual, steady decrease in amplitude. Top right: The spectrogram shows slight frequency modulation in the horizontal harmonic traces. The frequency variations of all harmonics are strongly correlated. Bottom left: The amplitude envelopes of all the harmonics of the trumpet exhibit the same gradual, steady decrease as does the overall waveform, and are strongly correlated with each other. The first trace corresponds to the excitation. Bottom right: Reverse view with higher harmonics shown in front.

The trumpet is noteworthy, as it exhibits both strong frequency and amplitude comodulation across all harmonics. There are some variations and fluctuations in the fine structure of the various harmonic envelopes which we will discuss further in Section 3.2.7 on the bass clarinet.

Note that the lowest frequency spectral component appears to be the excitation in the form of pulses of air from the point of contact between the mouth and the instrument. These produce a rough envelope shape. The effect of the body of the instrument in many cases is to smooth out these variations by acting as a filter. The instrument alternately absorbs and releases energy at those frequencies close to its natural body resonance modes. In many instruments the higher frequency harmonics appear more jagged than the lower harmonics, presumably because they are farther from the natural resonance frequencies, and do not benefit from this smoothing effect.

Bowed instruments similarly appear to have roughness in their excitation due to the movement of the bow against the strings which we will briefly describe in Section 3.6.5. The body of these instruments also acts as a filter, and helps to radiate the sound, as well. Here, too, the lower harmonics are smoother and their envelopes appear to track each other more closely, whereas the higher harmonics are more jagged and do not move in lock step.

3.2.6 Piano

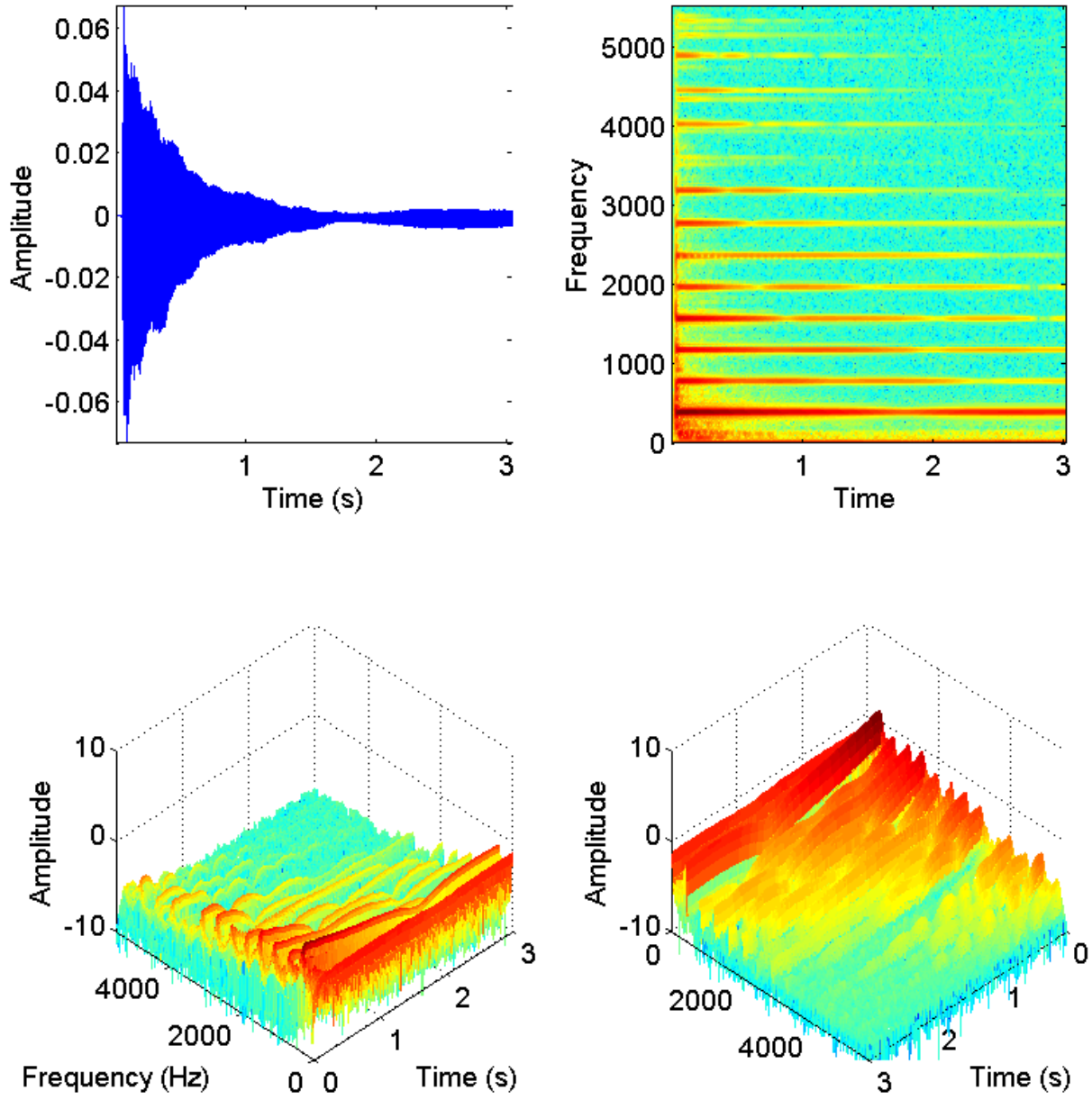

Figure 10. Top left: The time course of a piano playing the note G4. The envelope exhibits the characteristic attack, decay, and sustain described in the text. Top right: Spectrogram of the piano note. The harmonics do not exhibit any frequency variation, as there is no way to do so on a piano. The lower frequency noise at beginning of note is probably due to percussion noise of key strike. Bottom left: Mesh plot of piano spectrogram. Amplitudes are not correlated, thus comodulation is not an accurate model of the behavior of piano harmonics for reasons discussed in text. Bottom right: Reverse view with higher harmonics shown in front.

The piano is not capable of frequency variation, as the strings are fixed. As can be seen in Figure 10, the amplitudes of the harmonics do not appear to be correlated. A possible explanation for the difference between the piano and some of the other instruments we have examined is that

the piano is only struck at the beginning of the note. Hence, the entire response may be characterized as a transient response. Each harmonic may have its own characteristic mode of resonance which is independent of the others, and hence a different time course. In contrast, those instruments which are continuously driven such as the wind instruments and the bowed instruments all share a common excitation. When this excitation becomes stronger or weaker, the harmonics will tend to follow in a similar manner. Thus, those instruments are better candidates to fit an amplitude comodulation model.

What is unusual in the behavior of the individual harmonics is that, in general, transient responses tend to decay with time. An example is the decaying exponential response of an electric circuit with an RC time constant. In the case of the piano, it appears that some of the harmonics become stronger at times. This would seem to indicate that much complex coupling exists between the various resonant modes and possibly with the housing and other components of the piano, as well. Energy may be transferred from one mode to another, thus strengthening some harmonics at the expense of others in a back and forth manner. A general observation appears to be that the higher frequency harmonics have a shorter time constant than the lower frequency harmonics.

An additional observation is that at time 1.6 seconds, the time waveform appears to approach a minimum and then increases. Possibly this might be due to relative phase fluctuations among the various harmonics causing alternating constructive and destructive interference at various points in time.

Finally we note the existence of double harmonic lines at times 3500, 4000 and 4500 Hz and above. The explanation seems to be due to the fact that piano strings do not exhibit perfect harmonic behavior, but rather the exact overtone frequencies are non-integral multiples of the fundamental. At least one author attributes this to the fact that the wave speed in a string is not constant, but actually depends slightly on frequency. Because of this, dispersion occurs, and the overtones are actually a bit sharp, increasing with the number in the series. We cite from (Scavone, Abel and Berners, 2007) the following paragraph.

The actual relationship between the fundamental and the overtones is given by

$$f_n = nf_1\left[1+(n^2-1)J\right]$$

where f_n is the frequency of the n^{th} harmonic and f_1 is the frequency of the fundamental. For a solid wire without wrapping (of a second wire around it for extra mass, as is done on some of the piano keys):

$$J = \frac{\pi^3 d^4 Y}{128 T L^2}$$

where d is the diameter of the string, Y is Young's modulus (a measure of the stiffness of the string), T is the tension, and L is the length of the string.

The n^2 term gives rise to the nonintegral relationship among the frequencies of the overtones. Because of this, the lowest notes on pianos are usually tuned flat, so that their overtones will be in tune with the higher fundamentals, and the higher strings are tuned sharper so they will be in tune with the overtones of the lower strings.

The double lines, therefore, come about because the fundamental or overtones of a given string may be close enough to the fundamental or overtones of another string which corresponds to the same note, but located an octave higher or lower. This might set in motion sympathetic vibration that will excite the second string, as well. However, the higher overtone frequencies of the first string will not exactly correspond to the higher overtone frequencies of the second string due to the nonintegral relationship of overtones to fundamental in a piano. The net result will be lines of higher overtones that do not exactly match, since they originate from different strings, and correspond to a different harmonic number *n* in each string.

3.2.7 Bass Clarinet

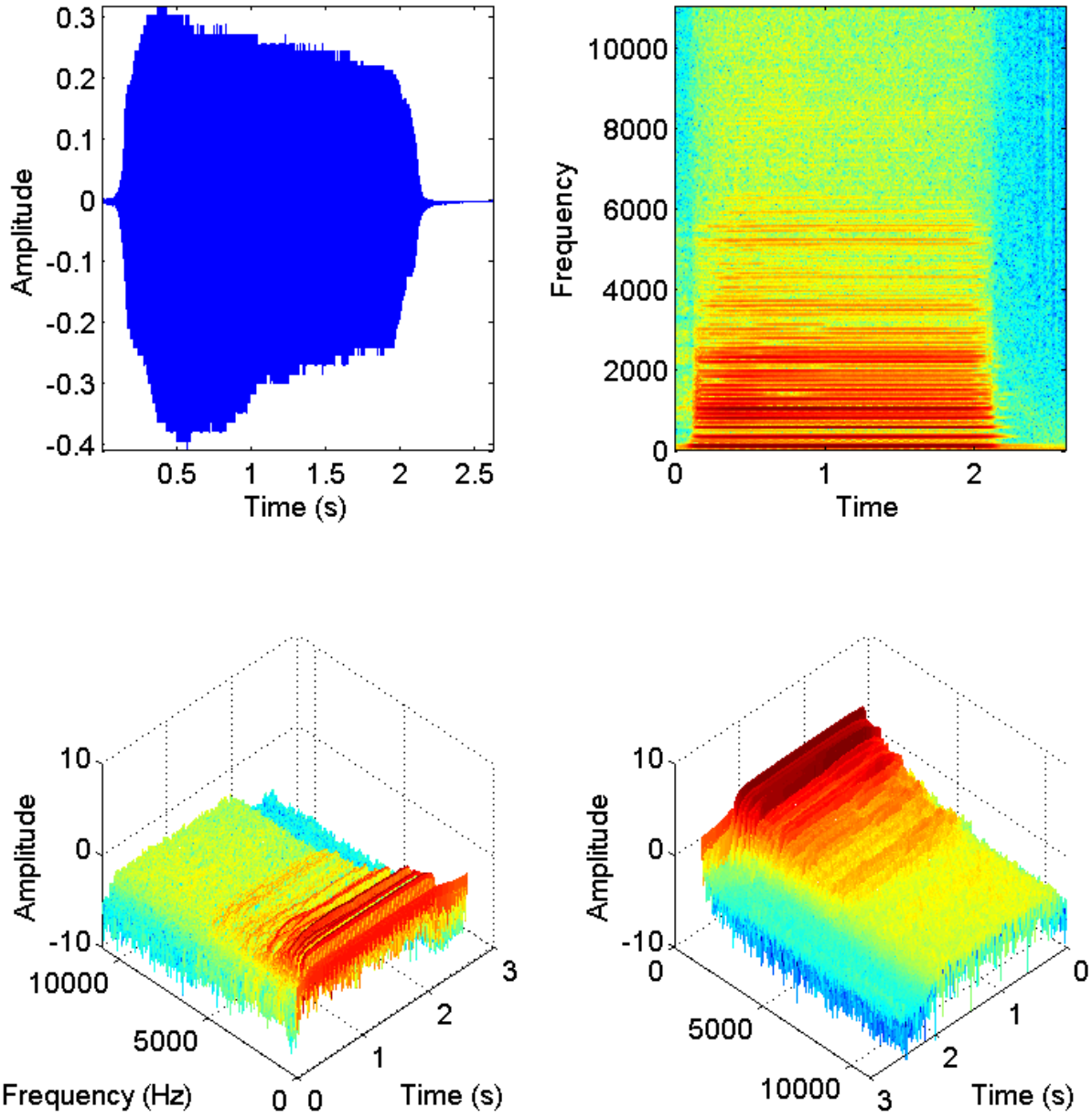

Figure 11. Top left: Time waveform of bass clarinet playing A#2. Top right: Spectrogram of bass clarinet. Harmonics do not exhibit any frequency variation. The richness of the sound may be due to the very high number and density of strong harmonics. Bottom left: 3-D mesh plot of spectrogram of bass clarinet. The excitation is smoother than for other wind instruments, but the higher harmonics become more and more jagged with increasing harmonic number. See text. Bottom right: Reverse plot wih higher harmonics shown in front.

We note that, as can be seen in Figure 11, the excitation of the bass clarinet seems smoother than we have seen in other wind instruments so far, but the higher harmonics still become quite jagged. This may be due to the types of interactions we noted in the case of the piano, where

energy appears to be transferred from one resonant mode to another in a back and forth manner. The mechanisms by which this may occur are beyond the scope of this thesis. Alternately, the large, heavy body may smooth low frequency components of the excitation, while the higher components are unfiltered.

3.2.8 Oboe

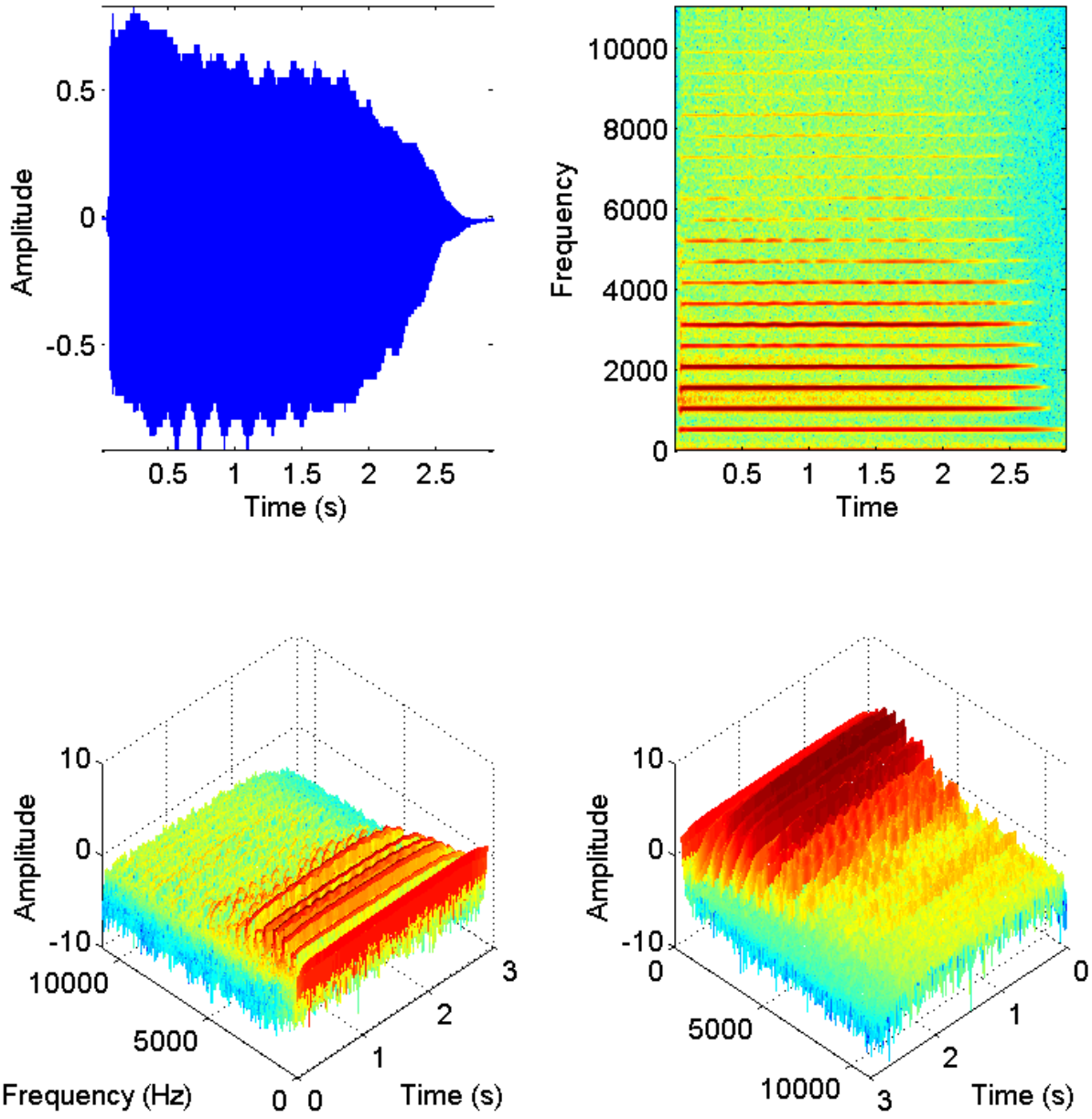

Figure 12. Top left: Time waveform of an oboe playing C5. The up and down modulation is due to musician's tremolo. Top right: Spectrogram of oboe. No frequency variation is observed, but amplitude variation can be seen in the intensity of traces. Bottom left: 3-D mesh plot of spectrogram of oboe. Amplitude variation can be observed due to tremolo. Although most harmonics show clear amplitude variation, the peaks and troughs of the various harmonic envelopes do not coincide with perfect synchrony. Bottom right: Reverse plot with higher harmonics shown in front.

The tremolo of the oboe in Figure 12 is apparent in both the time waveform and in the mesh plot of the harmonic envelopes. However, the harmonics do not move exactly in step as they

did in the case of the altoflute. Various factors and interactions, some of which we have examined in earlier sections, may prevent perfect synchrony among spectral components.

3.3 Relevance to Speech

3.3.1 Source-Filter Model

While we might have expected musical instruments to be candidates for displaying high levels of comodulation, in speech we should be more skeptical from the outset. The main reason is that speech is thought to follow a time-varying source-filter model. This model divides the speech production system into an acoustic source entity which models the vibration of the vocal folds of the glottis, and a filter entity which models the effect of the vocal tract. The vocal tract consists of the pharynx, and the oral and nasal cavities. These may be thought of as tubes. The pharynx and oral cavity are considered to be the main pathway, while the nasal cavity is considered to be a side branch. By modifying the shape (cross-section) of particular locations along the vocal tract, the resonant frequencies of the chamber are changed. These modifications are effected by motion of various muscles. Among them are muscles which extend and retract certain cartilaginous structures in the lower part of the vocal tract. The tongue muscle also does a major part of the work, elevating and lowering different regions to constrict or enlarge the effective cross sectional area of the significant portion of the vocal tract that is within its reach. This extends from way back at the base of the tongue all the way to its tip where it may make contact with various surfaces of the palate and/or teeth. The jaw and the lips also move in controlled manners to modify the resonances of the system.

3.3.2 Formants

Because of their importance in speech analysis, there is a specific term associated with the resonances of the vocal tract. These regions of the frequency spectrum are known as formants. The first four are thought to be significant for distinguishing the different sounds that comprise spoken speech. The frequencies of the set of formants change for each sound in accordance with the configuration of the vocal tract. The width and shape of these resonances have been characterized by a number of authors (Fant, 1960), (Stevens, 1999). Good agreement with acoustic theory has been achieved via the use of models of the vocal tract which attempt to approximate the system as a set of concatenated tubes of differing cross sections. The number,

length and width of the tubes is chosen for the particular sound and for the particular speaker. Men have longer vocal tracts than women, and women have longer vocal tracts than children.

3.3.3 Vowels

In the case of vowels, the glottal source emits a series of pulses which are rich in harmonic content. The fundamental frequency for men ranges from about 80-160 Hz, while for women it ranges from about 160-320 Hz. Those harmonics which fall within the region of a formant are emphasized while those which do not fall near formants are attenuated. It is the effect of the formants that alters the sound so that we recognize the particular vowel being spoken. Note that the situation here is quite different than music. In music, the distinguishing characteristic that differentiates one note from another is the fundamental frequency or pitch of the note. The relative amplitude of the harmonics affects only the timbre, as we discussed previously. In vowels, it is the relative amplitude of the harmonics that determines which vowel is spoken. In speech, except in tonal languages like Mandarin Chinese, pitch does not affect the determination of which phoneme is heard. Rather, pitch is used to convey context and intent, and to emphasize particular words or phrases. For example, in a question, the pitch rises towards the end of a sentence, whereas in a statement, it tails off. An exclamation has still a different form. The study of the use of pitch or intonation in speech is known as prosody.

Pitch also differs from speaker to speaker, and this is one of a number of distinguishing characteristic of a particular voice.

3.3.4 Consonants

The speech production system for consonants is different than for vowels. Whereas for vowels there is a single source in the glottis, for consonants many alternate means of sound production are employed. Various clicks or noise bursts formed by the sudden opening or closing under pressure of particular landmarks in the vocal tract generate plosive sounds such as the stop consonants /p/, /t/, /k/. The forcing of air through or against various surfaces creates turbulent swooshing sounds such as the fricatives /s/, /ʃ/, /f/.

In addition to the unvoiced consonants, above, there are voiced analogs of the stop consonants, such as /b/, /d/, /g/; and voiced analogs of the fricatives such as /z/, /ʒ/, /v/. Voiced

consonants also include the nasals, /m/ and /n/. In these consonants, a hybrid of voicing and plosive or frication sounds are generated.

While formants play a role in consonants as well as vowels, however, because the analysis of consonants is very complex, involving noiselike, aperiodic waveforms, we will not focus on those sounds in this thesis. We will note that amplitude comodulation may be applicable to aperiodic sources, as well, in the sense that one can control the overall level of a fricative without perceptibly altering its frequency content. However, one would need to define just what is being comodulated, as the frequency spectrum becomes continuous, rather than a set of discrete harmonics.

3.3.5 Continuity in Speech

A point which must be mentioned when analyzing speech is the fact that formants and pitch do not change abruptly when transitioning between phonemes. Rather, parameters change gradually and flow from one into the other. This is known as anticipation, in which one begins to set up the vocal tract for the next sound before the previous one is concluded.

3.3.6 Radiation Resistance

In addition to the source and filter characteristics, there is a final frequency-dependent factor that is used in vocal-tract models to account for the radiation resistance of the termination (at the lips). This attenuation is due to the fact that the sound must leave the 1-dimensional vocal tract and begin to propagate 3-dimensionally in open space, causing a loss of pressure due to the change in impedance.

3.3.7 Singing

We note a few observations on the acoustics of singing from (Scavone, Abel and Berners, 2007). Singing combines aspects of speech with aspects of music. On one hand, the vowels must retain some of the features of spoken vowels such as emphasis of specific harmonics to be recognizable. On the other hand, pitch now becomes important for conveying melody, and the pleasantness of the timbre must be maximized. The vowels and voiced consonants dominate, as music is primarily concerned with periodic vibrations.

To achieve these somewhat contradictory ends, the following methods seem to be employed. Singers are sometimes able to tune their formants to match one or more harmonics of the sung pitch. The first formant usually contributes most to the timbre, due to its lower frequency and higher amplitude, being closer to the fundamental. Singers sometimes modify the sound of a vowel to improve musical tone. It has been observed that singers appear to make use of an additional formant located at about 2500-3000 Hz known as the "singer's formant" which is independent of the particular vowel and pitch, and which adds brilliance and carrying power to the voice. It is attributed to a lowered larynx and widened pharynx, which forms an additional resonance cavity within the vocal tract.

The louder amplitude levels desired by the singer are produced via increased airflow to the glottis, and this causes more rapid glottal closures. The consequent sharper contours of the injected airflow yields higher frequency energy to the voice.

Table 1 summarizes some characteristics of speech, music and singing.

	Speech	Music	Singing
Pitch	• Speaker characteristics. • Prosodic effects. • (Aesthetic)	• Determines note heard. • Follows melody • (Essential)	• Determines note heard. • Follows melody • (Essential)
Harmonic Weighting	• Determines vowel heard. • Governed by vocal-tract resonances (formants). • Some variation among speakers. • (Essential)	• Timbre or quality of note. • Determines instrument heard in combination with shape of modulation envelope. • (Aesthetic)	• Determines vowel heard. • Governed by formants • Presence of additional singing formant. • Modification of vowels to enhance timbre. • (Essential)

Table 1. Contrasts between the use of pitch and the relative weighting of harmonics in speech and music.

From the table it is apparent that what is essential in normal speech is aesthetic in instrumental music, and what is essential in instrumental music is aesthetic in normal speech. A vowel will completely change if the relative harmonic amplitudes are altered. The words *hit, height, hate, heat, hot, hat, hut* would all be indistinguishable from each other. However pitch will not affect

the word that is heard. In instrumental music, on the other hand, if we change the harmonic weightings, we may change the perceived instrument or the quality of the sound, but the song would still be recognizable. Yet, if pitch is altered, then the note becomes completely different, and the song would no longer be the same.

In the case of singing, a hybrid of normal speech and instrumental music is employed. Pitch is important for melody which takes the place of the prosody of normal speech. For this reason, it is difficult to distinguish between a statement and a question in singing (without understanding the words), as the use of pitch for melody displaces its normal usage for prosody. Vowels are still determined by formants, but modifications are made to enhance timbre.

3.3.8 Amplitude Comodulation and Speech

From all of the above, it should be apparent that it will be somewhat difficult to fit an amplitude-comodulation model to actual speech data. In the case of speech that includes multiple speech sounds or units (phonemes), there cannot be uniform amplitude modulation of all harmonics in the speech sample, since in order to produce different phonemes, it is necessary to emphasize and deemphasize the amplitude of different harmonics for each phoneme. The time-varying vocal-tract filter selectively, not collectively, adjusts the amplitude to the correct relative value for the desired phoneme. In addition, as we have seen, at the beginning and end of a phoneme the vocal tract already begins to modify itself in order to segue into the next phoneme in a seamless manner.

If there is a way to accommodate an amplitude comodulation model in speech, it is more likely to work within a single phoneme, away from the boundaries of the preceding or following phoneme. In a steady-state protracted vowel, one might expect that raising or lowering the volume of one's voice should affect all harmonics equally. However, even in this region there are complicating factors, because increasing the driving force, i.e., the subglottal pressure, causes pitch to rise (Catford, 1988). That will introduce a degree of frequency modulation into the signal, as well. In addition, the naturalness of the human voice depends on many phenomena that might be termed irregularities, but which actually give the voice a pleasant quality that distinguishes it from a mechanical buzz. These include frequency fluctuations (jitter), amplitude fluctuations (shimmer), aspiration noise, pitch doubling, and many other effects that may be amplitude-dependent, and cannot be captured by a single modulation

coefficient. In passing, we note that, carried to the extreme, irregularities are detrimental to the sound of the voice and are associated with pathological conditions like creakiness, hoarseness, or roughness, etc.

For the above reasons, amplitude comodulation would be expected to be of more limited use in speech then it is in music. However, we note that onsets, which are a subset of amplitude comodulation, can be applicable in speech, as well, as all harmonics across the board are likely to start and stop together at the onset or offset of voicing. This may occur fairly often within a sentence, as consonants, especially the stop consonants, punctuate the flow of sound, and periodically interrupt and restart it. In Chapter 4, we will describe an algorithm which depends on the presence of onsets and offsets for source separation.

3.3.9 Frequency Comodulation and Speech

While amplitude comodulation may be more difficult to utilize in speech for the reasons described, frequency comodulation would be expected to be more useful, assuming one could develop an algorithm to harness it effectively. The reason is that the filtering effect of the vocal tract is assumed to be linear for all practical purposes. Linear filters cannot generate or alter frequencies, but can only modify the amplitudes and/or phases of the input signal. Because frequencies will, for the most part, be unchanged by the vocal-tract filters, common frequency variation should still be observed after the filtering process. The amplitudes of harmonics may change relative to each other, but frequency tracks will remain multiples of the fundamental. The periodic attribute of the glottal source we noted earlier mandates this behavior. In Chapter 5, we will analyze actual speech recordings, and verify that this is indeed the case. We will then describe work on algorithms which attempt to track frequency variations.

3.4 Comodulation: A Priori or a Posteriori

In the next chapter, comodulation will be used in an algorithm to dissect a mixture of sources into its constituents. The individual harmonics of each source will be found by the algorithm as it attempts to find the best fitting set of source signatures and modulation vectors (to be defined) that could explain the data. The parameters (amplitude and frequency) of the harmonics of each source are determined by the algorithm. Source segregation is performed on the data as a whole. Comodulation serves as a constraint in limiting the universe of possible

solutions. We refer to this type of scheme as *a priori* comodulation. We do not provide to the algorithm in advance the parameters of the waveforms. Rather, the algorithm calculates how much energy to allocate to each source in those frequency regions where there is overlap between harmonics of both sources, based on the behavior of other frequency regions. That is actually the ideal method of applying comodulation to the separation problem. Unfortunately, it is extremely difficult to design such an algorithm for frequency-varying cases. The immediate reason is that the curving of the harmonic lines makes the trajectories difficult to capture in conventional matrix form. But a more basic reason is the difficulty of constraining the now much larger set of possibilities to a single, unique solution.

In Chapter 5, we will look at alternate approaches for using comodulation for source separation. We will first use other methods which we will discuss later to independently determine the parameters of each spectral component in the mixture. We can then group them together on the basis of common modulation characteristics. We refer to this as *a posteriori* application of comodulation.

Although this is not as elegant an approach as is the *a priori* application of comodulation, it has an advantage in that it allows one to handle deviations from completely perfect comodulation which are expected, based on our observations of real world signals.

3.5 Reduction of Ambiguity

We would like to examine a few conceptual difficulties in the definition of modulation. (A. L.-C. Wang, 1994) in his thesis raises an interesting question. If we define a signal as

$$x(t) = a(t)\sin(\omega t + \phi) \tag{3.1}$$

can we separately define a modulation term, and an instantaneous frequency term? He says that we cannot, since we can always find some function $\omega(t)$ which satisfies

$$\omega(t) = \arcsin\left[\frac{x(t)}{a(t)}\right] - \phi \tag{3.2}$$

In other words, we can't unambiguously attribute amplitude changes to a modulation or envelope term, when they could in fact be due to frequency changes. (This argument requires

some care, as the *arcsin* function will become complex if the absolute value of its argument exceeds 1.0.)

We furthermore raise an additional question: We know that beating (constructive and destructive interference) can produce changes in the amplitude of the resultant.

For example, if we have

$$s_1 = \sin(\omega_1 t + \phi_1) \qquad (3.3)$$
$$s_2 = \sin(\omega_2 t + \phi_2)$$

using trigonometric identities, the sum can be written as

$$x = s_1 + s_2 = 2\cos\left[\frac{(\omega_1 - \omega_2)t}{2} + \frac{(\phi_1 - \phi_2)}{2}\right]\sin\left[\frac{(\omega_1 + \omega_2)t}{2} + \frac{(\phi_1 + \phi_2)}{2}\right] \qquad (3.4)$$

where we have a slowly varying cosine envelope modulating a more rapidly varying sine carrier. In fact, this has the mathematical form of double-sideband suppressed-carrier (DBSC) modulation. If we look over a short time region, we might wonder whether we have beating of two signals, or a single signal which is being modulated by a cosine. We specify a short region, since over longer regions, the waveform doesn't look like what we would ordinarily refer to as a regular amplitude-modulated waveform, i.e., with a full carrier, since in the beating case, the waveform undergoes certain phase reversals which do not happen in a regular AM waveform.

Let us look at an example consisting of plots of two waveforms shown in Figure 13 and Figure 14, respectively, the first produced by beating, and the second produced by 100% modulation. In each case, the phases were 0. We seek to understand the differences between the resultant waveform of a pair of interfering sines at 20 and 22 Hz, and a single sine of frequency 21 Hz with appropriately applied modulation. From Equation 3.4 with phases 0, the equation of the first plot is:

$$x = s_1 + s_2 = 2\cos\left[\frac{(\omega_1 - \omega_2)t}{2}\right]\sin\left[\frac{(\omega_1 + \omega_2)t}{2}\right]$$

Substituting the values in the text we have:

$$
\begin{aligned}
x &= \sin[2\pi(20)t] + \sin[2\pi(22)t] \\
&= 2\cos\left[2\pi\frac{(22-20)t}{2}\right]\sin\left[2\pi\frac{(22+20)t}{2}\right] \\
&= 2\cos[2\pi(1)t]\sin[2\pi(21)t]
\end{aligned} \tag{3.5}
$$

This is shown in Figure 13.

In Figure 14, we added a carrier, and doubled the modulation rate to match the first figure. (The need for this doubling is that when we add a carrier, we prevent the modulation term from going negative. This prevents the phase shifting that occurs with suppressed carrier modulation, and appears to reduce the number of lobes by one half. As long as we are consistent, we can double the modulation frequency during the following discussion in each case without affecting the thrust of our argument.)

It was thus generated by using

$$
\begin{aligned}
x &= [1+\cos(\omega_1-\omega_2)t]\sin\left[\frac{(\omega_1+\omega_2)t}{2}\right] \\
&= [1+\cos 2\pi(22-20)t]\sin\left[2\pi\frac{(22+20)t}{2}\right] \\
&= [1+\cos 2\pi(2)t]\sin[2\pi(21)t]
\end{aligned} \tag{3.6}
$$

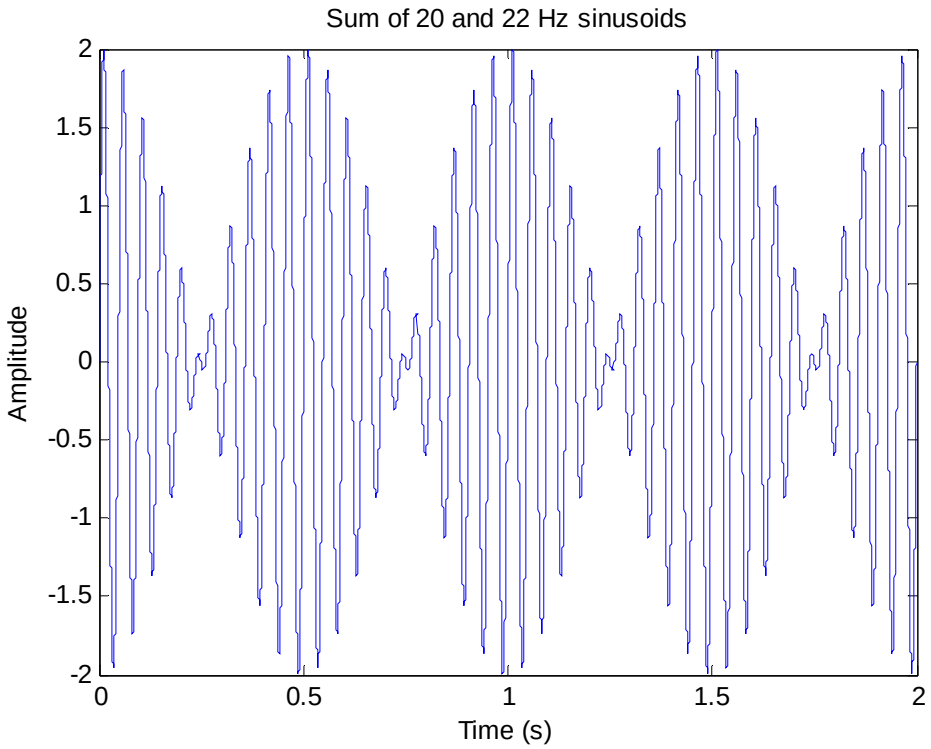

Figure 13. The sum of 20 and 22 Hz sinusoids. Envelope exhibits beating pattern from constructive and destructive interference at the 2 Hz difference frequency.

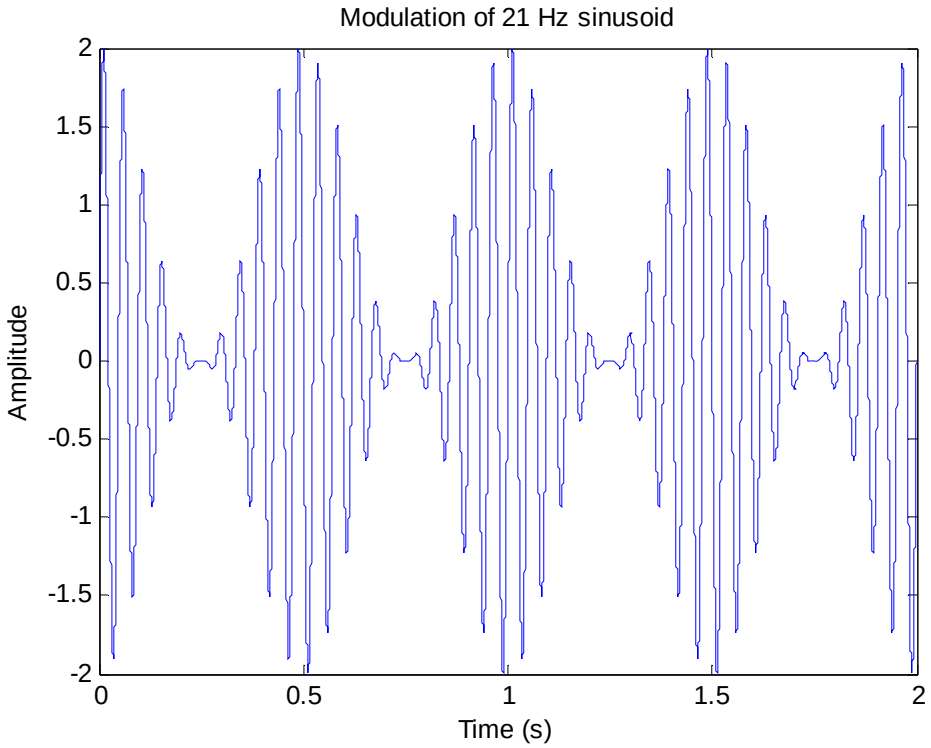

Figure 14. Cosinusoidal amplitude modulation of 2 Hz applied to 21 Hz sinusoid. Note similarity to beating pattern of sum in previous figure. Envelopes of both cases exhibit 2 Hz variation.

A comparison of these two plots indicates that they are quite similar. We maintain that, for example, looking at the region from 0.45 to 0.55 seconds in both plots that there is really no obvious way to tell whether there are one or two signals.[2] As a matter of fact, Fourier analysis is based on the principle that all waveforms (satisfying certain fairly common conditions) can actually be produced by beating many sinusoids of constant amplitude and infinite length against each other. This includes AM and FM signals. So we are left with a disturbing ambiguity in our description of signals. It is especially troubling if we are trying to do source separation, and can't determine whether we have one or two sources to begin with.

However, the assumption of comodulation may clarify matters. Granted, that by looking at the first harmonic alone, the situation is indeterminate, however, by looking at additional harmonics, we will show that the situation becomes clearer.

As an example, let us examine the second harmonic of the previous two cases, the case of beating and the case of AM modulation.

In the case of beating, we now have

$$x = \sin\left[2\pi(40)t\right] + \sin\left[2\pi(44)t\right] \tag{3.7}$$

In the case of modulation, we now have

$$x = \left[1 + \cos 2\pi(2)t\right]\sin\left[2\pi(42)t\right] \tag{3.8}$$

Figure 15 and Figure 16 show the situation in each case.

[2] (Terhardt, 1974) and others cited in (Hartmann, 1998) actually conducted tests on listeners to determine what differences in percept, if any, distinguish between beating and modulation. Both are perceived as having an element of roughness due to the fluctuating envelope. Our work in this section suggests that in the absence of additional harmonics, it would be difficult to distinguish the two. This author is not aware if similar tests using multi-harmonic signals have been performed that might measure any advantage in the latter situation.

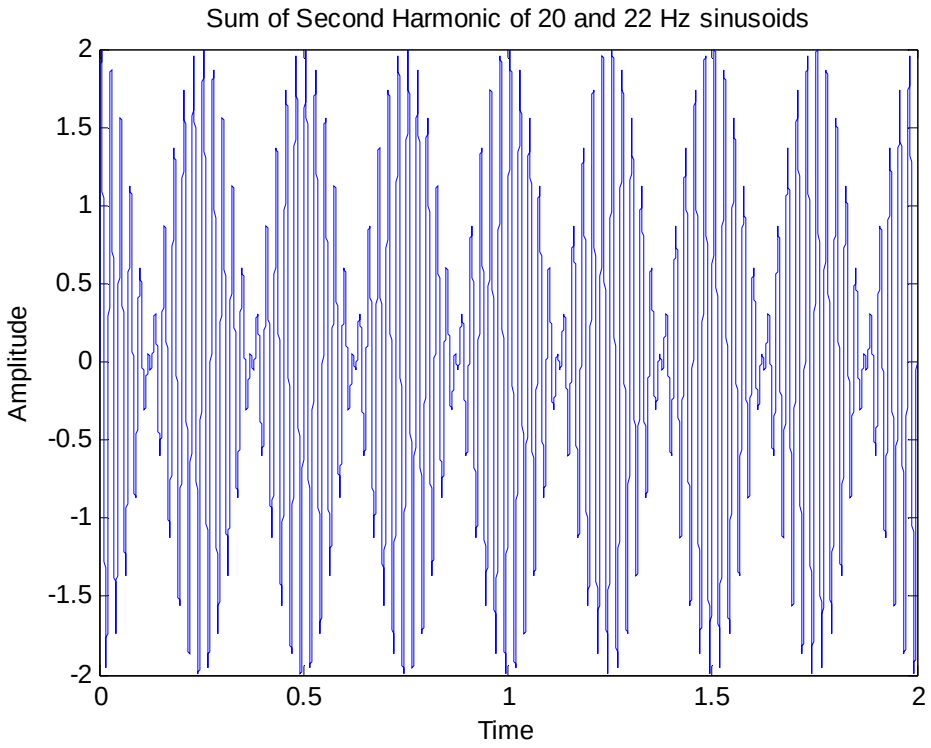

Figure 15. The sum of the second harmonics of the 20 and 22 Hz signals at 40 and 44 Hz, respectively. Note that the beat frequency is now at the 4-Hz difference frequency of the second harmonics.

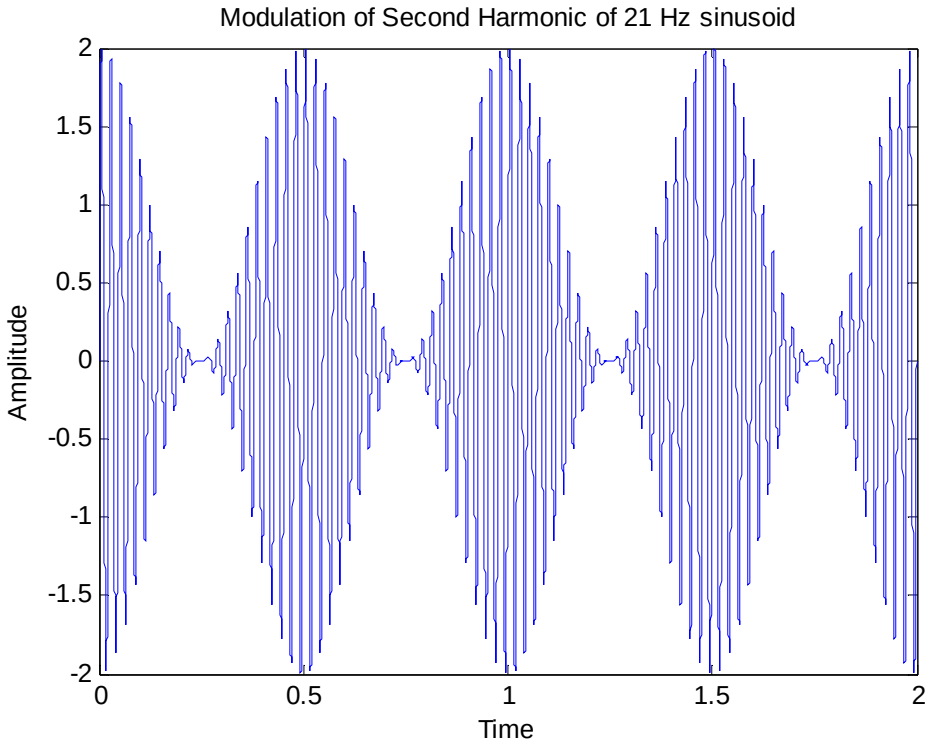

Figure 16. 2-Hz amplitude modulation of Figure 14 now applied to the 2nd harmonic of the 21-Hz carrier. Envelope still exhibits 2-Hz variation. Note difference from Figure 15, where 4-Hz variation was observed.

It is apparent that the summed waveform of the 2nd harmonics will beat at twice the frequency as that of the fundamentals, while the modulation envelope of the second harmonic will remain at the same frequency as the envelope of the fundamental, assuming comodulation.

This is one example of how the redundant structure of speech and music may help in the interpretation and separation of an audio scene. The use of multiple harmonics might serve to reduce ambiguity by distinguishing interference due to multiple sources from modulation of a single source.

In Chapter 5 we will look at another major redundancy in auditory signal processing. That is the use by the auditory system of multiple overlapping filters, so that each frequency component appears in multiple channels.

3.6 Phase Comodulation: Scaling

It turns out that if we impose an additional condition on an amplitude- or frequency-modulated source, an interesting time-domain interpretation arises. This requirement is that the relative phase among the spectral components remains fixed. We term this phase comodulation.

3.6.1 Amplitude Comodulation: Vertical Scaling

A useful way to think about comodulation is as a scaling, either an expansion or contraction of the signal. Consider a signal which can be written as the sum of a set of spectral components.

$$x(t) = \sum_n a_n e^{j(\omega_n t + \phi_n)} \tag{3.9}$$

where a_n, ω_n and ϕ_n are the amplitudes of the frequency and phase of the nth spectral component, respectively. If this signal is amplitude-comodulated, then we can write it as

$$x'(t) = \sum_n a'_n(t) e^{j(\omega_n t + \phi_n)} \tag{3.10}$$

where

$$a'_n(t) = a_n m(t) \tag{3.11}$$

and $m(t)$ is the shared modulation function of the entire source, i.e., all the components. Note that we have considered $a'_n(t)$ to be a function of time. This is a different perspective than the

Fourier concept of constant-amplitude sources which add in such a manner so as to produce a function of time.

Making this substitution we get

$$x'(t) = \sum_n m(t) a_n e^{j(\omega_n t + \phi_n)} \tag{3.12}$$

Since $m(t)$ does not depend on n, as all bands are modulated by the same factor, then we can factor out $m(t)$ and write as

$$x'(t) = m(t) \sum_n a_n e^{j(\omega_n t + \phi_n)} \tag{3.13}$$

We see that

$$x'(t) = m(t) x(t) \tag{3.14}$$

This shows us that amplitude comodulation can be viewed as a simple amplitude-scaling of the original signal. If $m(t)$ is less than 1, then we have a contraction in the vertical direction. If $m(t)$ is greater than 1, then we have a dilation in the vertical direction.

3.6.2 FM Comodulation: Horizontal Scaling

For the following discussion we use a similar set of spectral components, but with one important addition. We require all components to be harmonically related. As before, we begin with

$$x(t) = \sum_n a_n e^{j(\omega_n t + \phi_n)} \tag{3.15}$$

But now we further specify that

$$\omega_n = n\omega_0 \tag{3.16}$$

We therefore have

$$x(t) = \sum_n a_n e^{j(n\omega_0 t + \phi_n)} \tag{3.17}$$

where ω_0 is the fundamental frequency.

We recall from study of elementary signals and systems that such a signal will be periodic. An integral number of periods of all harmonics will exactly fit in the time interval of the fundamental period, and as a result, the entire pattern will repeat in the next cycle.

Let us now assume that the source is frequency-comodulated, with the additional requirement of phase comodulation. We then have

(3.18) $$\omega_0' = k\omega_0$$

and

(3.19) $$\phi' = \phi$$

We can then write the new signal as

(3.20) $$x'(t) = \sum_n a_n e^{j(n\omega_0' t + \phi_n')}$$

If we make the substitutions above, we have

(3.21) $$x'(t) = \sum_n a_n e^{j(nk\omega_0 t + \phi_n)}$$

This can be rearranged as

(3.22) $$x'(t) = \sum_n a_n e^{j(n\omega_0 kt + \phi_n)}$$

But this can be rewritten as

(3.23) $$x'(t) = x(kt)$$

which is simply a time-scaling of the original signal. That is equivalent to a horizontal dilation or contraction. Since we scale each component by the same amount, we maintain the overall shape of the signal.

We emphasize again that this will only hold if the relative phases remain unchanged from their original values, i.e., under the condition of phase comodulation. If the physical mechanisms that vary the frequency or amplitude of the waveform do not preserve phase, then scaling will not occur.

3.6.3 Phase Comodulation: Time-Domain Perspective

The physical picture that results from the above discussion is that to build a perfectly phase-comodulated signal, one begins with a particular waveshape corresponding to a single cycle of a periodic[3] signal. We refer to that particular shape as a motif, and it is characteristic of the particular sound source. A series of these motifs is then concatenated to form a chain of similar shapes. The chain may be stretched and compressed both vertically and horizontally, as desired. A mixture of sources is formed by adding this chain to another chain which has some other motif and a different pattern of horizontal and vertical scaling. The object of comodulation-based source separation is to separate the signals and recognize that the very complex pattern that emerges in the sum is composed of very simple repeating building blocks which have merely been stretched or compressed. Figure 17 shows an example of 3 different types of signals which were all derived from repeating motifs in the preceding manner, but whose sum in the bottom plot is quite difficult to decipher. The first plot is a sinusoidally FM-comodulated set of 5 unity-amplitude, zero-phase harmonics. The motif is a period of a 5-harmonic set. The modulation is rhythmic contraction and dilation in the horizontal direction. The second plot is a ramped amplitude-comodulated triangle wave. The motif is a single period of a triangle wave, and the modulation is a vertical stretching. The third is a sinusoidally amplitude-comodulated square wave. The motif is a single period of a square wave, and the modulation is a vertical stretching and compression.

The resultant bears very little resemblance to any of the others. Although comodulation is an important cue for separation of signals, it is still extremely difficult to reverse-engineer the mixture. The simple, orderly patterns obliterate each other when added together. This is true even if perfect phase comodulation holds, which as we will see in the following discussion, is not an accurate model for realistic sources. Correctly partitioning a sum of signals into parts that preserve particular relationships among components is the central problem in this thesis.

[3] Strictly speaking, once we modulate the signal, it won't necessarily be periodic anymore, but we use the term loosely.

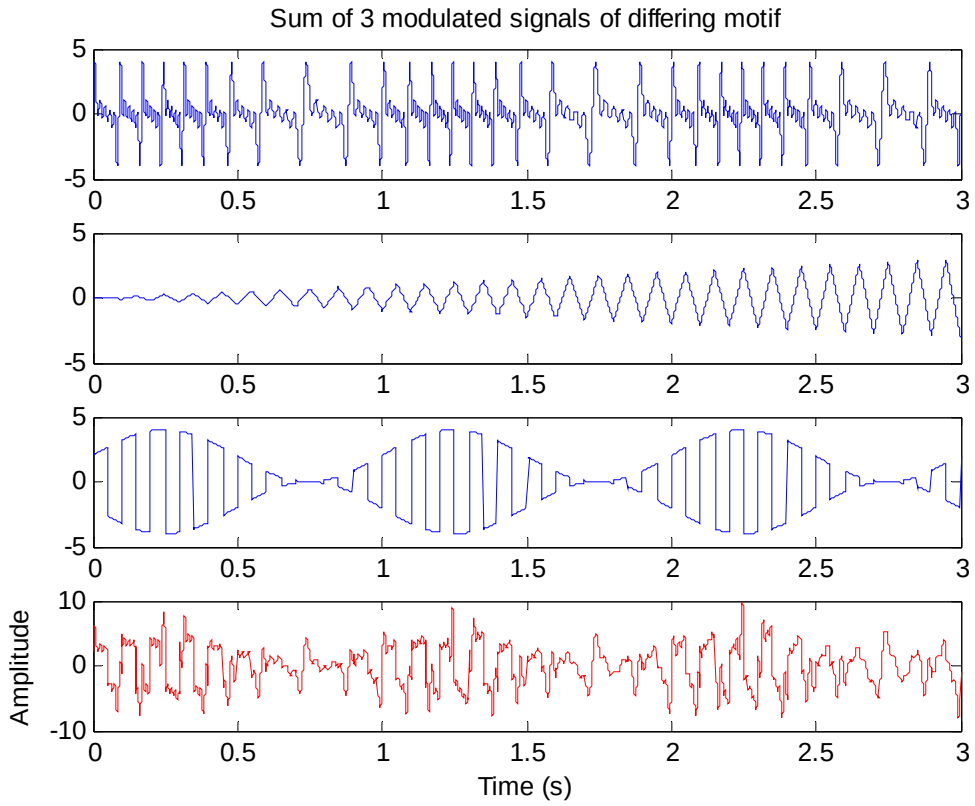

Figure 17. Top plot: a set of 5 unity-amplitude, zero-phase harmonic sinusoids frequency-comodulated by a sinusoid. The carrier fundamental is 10 Hz with frequency deviation of +/- 4 Hz. The modulating frequency is 1 Hz. 2nd plot: A square wave of frequency 10 Hz amplitude-comodulated by a ramp of amplitude *t*. 3rd plot: A square wave of frequency 10 Hz amplitude-comodulated by a sinusoid of frequency 1 Hz. Bottom plot: The sum of the first 3 signals. Very difficult to discern any of the motifs or modulating patterns.

3.6.4 Waveforms: Cycle to Cycle Comparison

We now look closer at the waveforms of some of the instruments we have studied in Section 3.2 to see whether they conform to what we would expect for a perfectly phase-comodulated instrument. Each pair of plots in the following three figures (Figure 18, Figure 19, and Figure 20) is taken from the same recording of the sustained note, but from two different regions in time.

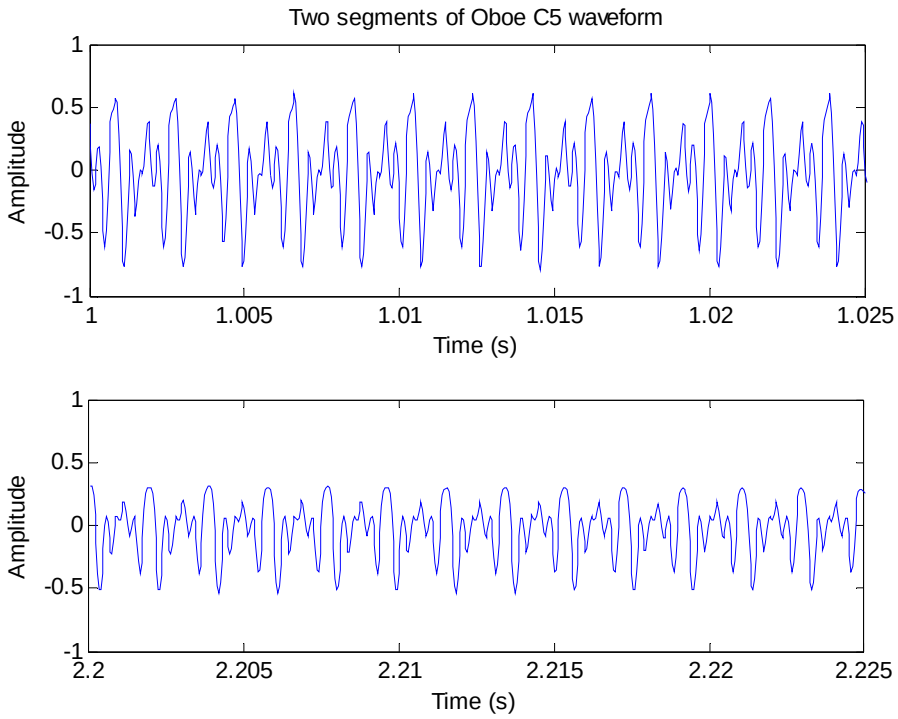

Figure 18. Two time segments from same oboe note are shown which illustrate the greater cycle-to-cycle similarity seen in the shorter term over that seen in the longer term. Perfect scaling is not observed due to factors discussed in text.

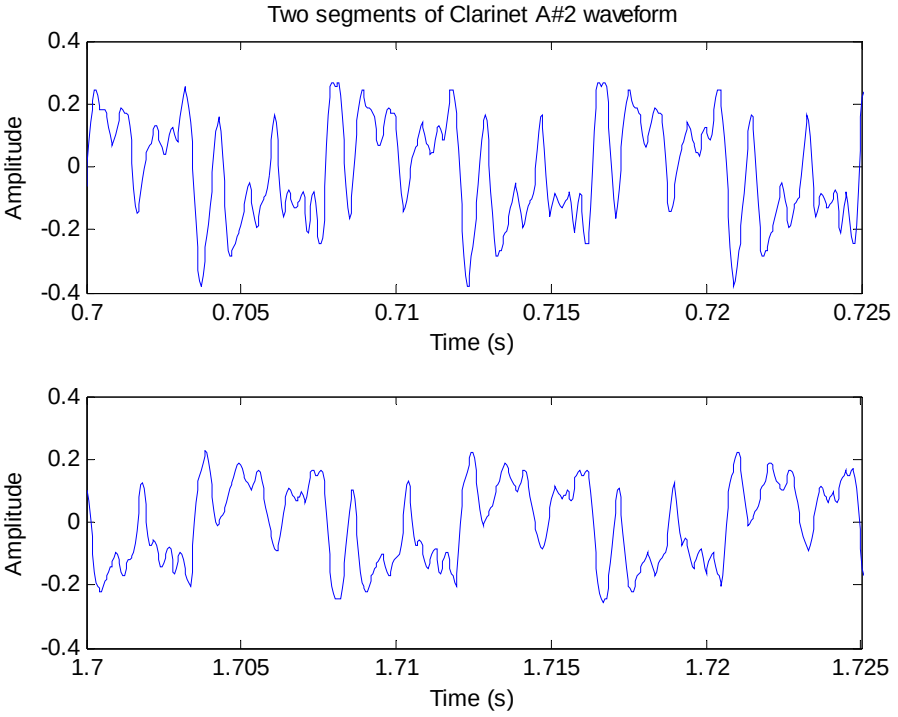

Figure 19. Two time segments from same clarinet note are shown which again illustrate the greater cycle-to-cycle similarity seen in the shorter term over that seen in the longer term. Perfect scaling is not observed due to factors discussed in text.

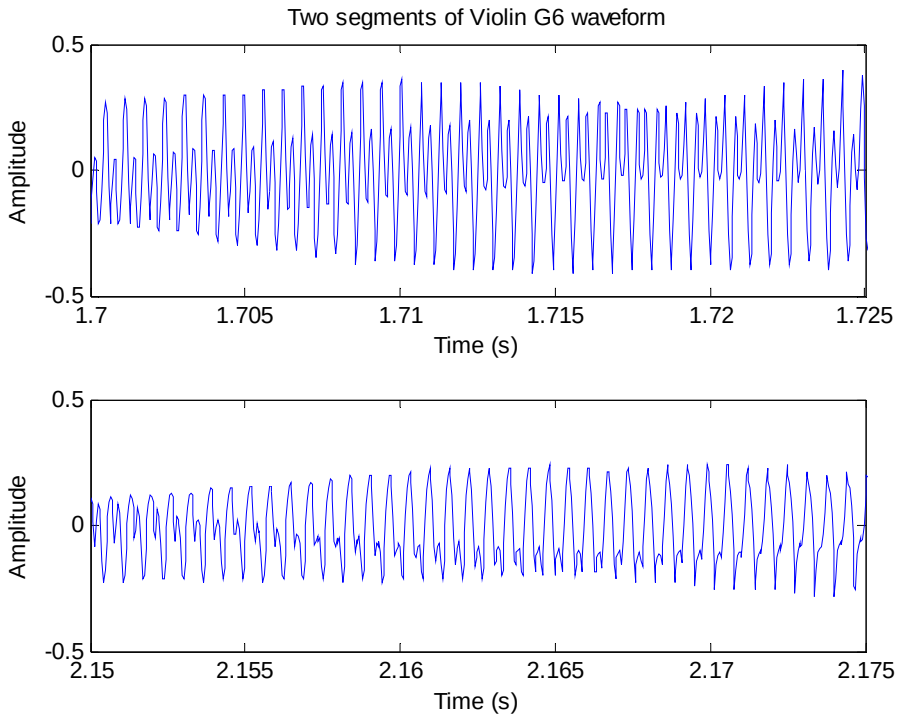

Figure 20. Two time segments of same violin note are shown. In this case even in the short term one does not see great cycle-to-cycle similarity. Perfect scaling is not observed due to factors discussed in text.

In all 3 cases, the nearby waveshapes are fairly similar, as can be seen by examining two cycles in close proximity to each other. However, as we move apart in time, the waveshapes become different, and do not scale, as can be seen by comparing a few cycles from the top plot of a pair with a few cycles from the bottom plot of a pair. From a visual estimation it is apparent that perfect phase comodulation is not a completely accurate model for any of these three instruments. Nevertheless, as we saw earlier, amplitude and/or frequency comodulation models alone can still be useful for describing the behavior of these instruments.

3.6.5 Why Imperfect Scaling Occurs

A possible reason why perfect scaling is not observed, in the case of constant-frequency instruments like the oboe and clarinet, is the fact that the resonant modes may not all be excited in perfect unison. Sound generation in instruments which have a reed is quite complex. There may also be variations in coupling between the mouth and instrument which affect the amplitude of some modes more than others. Movement of the instrument may affect this coupling. This may explain why, within short time segments, the waveshapes seem to scale, whereas over longer time intervals, the scaling becomes less accurate. As the musician changes

his position, the coupling will change. In addition, when playing louder, coupling may change, as well, due to the different configuration of the mouth needed to blow louder sounds.

In the case of the piano, we have seen in Section 3.2.6 that the overtones are nonintegral multiples of the fundamental, and this violates the assumption of Section 3.6.2, thus we cannot expect scaling for that reason. The explanation is that because the overtones are nonintegral multiples, the composite waveform cannot be periodic as we explained before, since an integral number of overtone periods will not fit within the period of the fundamental. At the next cycle (of the fundamental) the overtones will therefore not be in the same phase relationship as in the previous cycle, thus causing a difference in shape from one cycle to the next.

For frequency-varying instruments, if scaling is to hold, in addition to the requirement of frequency comodulation, the relative amplitudes of the harmonics must remain unchanged, as well. In the case of the violin, an important consideration which works against perfect amplitude comodulation is the fact that most of the sound is radiated from the body, and not from the strings. The sound is coupled from the string to the body through the bridge. Near the bridge, between the lower and upper surfaces of the hollow body is wedged a sound post that transfers energy from the upper to the bottom surface. As the body vibrates, the sound escapes through the "*f*" holes in the upper surface of the body. However, since the body has its own particular resonances, it acts like a filter. Because of the characteristic frequency response of the body, as frequency varies, some harmonics will be boosted more strongly than others. This will upset the relative amplitude relationships among the harmonics, as some will move closer and some farther from frequencies of resonance. This will upset the amplitude comodulation requirement.

Other considerations that affect the relative amplitude of harmonics are the position at which the bow meets the string. For a plucked string, it is well known that the position at which the string is plucked will affect the amplitudes of certain harmonics. For example if one plucks a string in the middle, it will silence the second harmonic. The reason is that the second harmonic requires a node in the middle of the string. This allows the two halves of the string to vibrate in the characteristic double loop standing wave pattern. But plucking a string in the center produces an antinode at that point. Since a node is not present at that location, the second harmonic will be absent.

In a violin, the motion of the bow is often analyzed as a pulling and sliding motion. Friction between the string and the bow pulls the string in the direction of the bow. (Violinists use rosin to increase this friction.) Once the string reaches a tension that is greater than the frictional force, it slips backwards, until it is pulled forward again by the moving bow. Therefore, bowing can be thought of, in a sense, as a plucking excitation. Since the bow is liable to move slightly in a lateral direction, as no violinist can pull the bow completely perpendicularly to the strings, this will slightly vary the harmonic content of the note, as well.

A final factor limiting scaling in musical instruments is simple phase incoherence or jitter that disrupts phase relationships among harmonics.

In speech, among factors which would prevent perfect scaling are possible dissimilarities in glottal pulses from one to the next, especially under conditions of changing amplitude and frequency which are normal in the course of conversation and prevent a speaker from sounding monotonous by giving expression to the voice. Another factor was discussed in Section 3.3.1, the action of the vocal-tract filter which selectively boosts certain harmonics over others. Time-varying changes in its configuration differentiate vowels from one another.

3.7 Summary

In summary, while perfect phase comodulation appears to be rare in real sound sources and hence perfect scaling is not observed, nevertheless, there are still fairly strong relationships among the amplitudes and frequencies of many of the spectral components of these sources that can potentially be exploited in source separation. In Chapter 4, we begin work on using amplitude comodulation as a basis for separation of constant-frequency sources.

Chapter 4

A Mathematical Analysis of Amplitude Comodulation

4.1 Introduction

As we have discussed in earlier chapters, amplitude comodulation refers to common amplitude variation among spectral components that is observed in many types of sound sources. In Chapter 2 we reviewed a number of psychophysical studies that support the idea that comodulation may play an important role in perceptual grouping of sound components. In this chapter, we develop an approach that uses amplitude comodulation as a basis for auditory source separation. We present a derivation of the equations governing a comodulated system for the case of constant-frequency sources. We prove a theorem giving necessary and sufficient conditions for uniquely decomposing a sound mixture into its constituent sources. A consequence of these conditions is an understanding of the importance of common onsets and offsets in source separation. Finally we present an algorithm for computing the solution in those cases where one exists.

4.2 Initial Intuition

As mentioned in Chapter 1, a simple separation scheme based on comparing channel outputs from a filter bank and grouping those bands with similar envelopes together, will likely lead to unreliable results due to the fact that bands containing energy from multiple sources may well have different envelope shapes than either source alone. This will likely cause confusion as to the true number of sources. Figure 21 illustrates such an occurrence. Looking at the right hand column (red waveforms), on the basis of shape alone one might conclude that there are three

sources, a rising, a falling and a stationary source. In truth there is only a rising source and a falling source, with the middle band representing an overlap or summation of the other two.

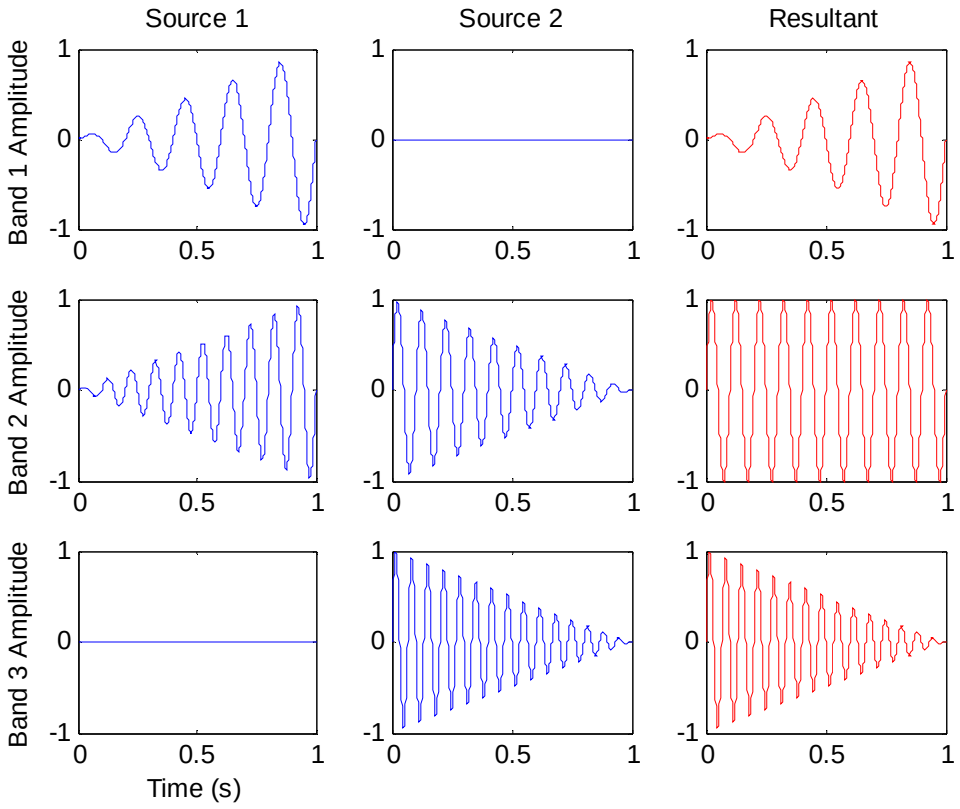

Figure 21. Example of 2 in-phase sources overlapping in middle band of a 3-band system. First source has components in bands 1 and 2, and is modulated by an upward sloping ramp. Second source has components in bands 2 and 3, and is modulated by a downward sloping ramp. Resultant waveform for each band is shown in right hand column. The middle band has a flat envelope, although it does not represent a new source.

The problem then is three-fold. Given a sound sample, (1) How do we correctly determine the number of sources? (2) How do we accurately determine the contribution of each source to the total energy in each band? And (3), if we find a solution, is it unique, or can another set of sources be found which will yield the same result?

In the following discussion all sources are assumed to be constant in frequency.

4.3 Source Representation

We represent each source i of the set of r sources by a column vector $\mathbf{s}_i = [s_i(f_1), s_i(f_2), \ldots, s_i(f_n)]$ which we call a spectral signature. The elements $s_i(f_j)$ of this vector represent the relative amplitudes that would be output from filter j in a filter bank of n filters, were source i to be present alone. At this point we make no specifications on the nature or number of filters in the filter bank. The name source signature is appropriate since it reflects the fact that the distribution of the amplitudes of the various spectral components is unique to that type of source, i.e., each musical instrument will emphasize certain overtones and attenuate others depending on the acoustic properties of the instrument. These include the type of instrument such as wind, string or percussion; the method of excitation such as whether produced by bowing, plucking or striking; and the physical dimensions and material properties of the instrument. It should be noted that in many cases, a complete mathematical description of all the complex interactions between the various parts of the instrument and its support structures, and the resultant effect on the natural modes of the instrument is very difficult to achieve. Whole books have been written just on the physics of the violin. As we noted earlier, in some cases the frequencies of the overtones will not be harmonically related. It is this wide variation that gives each instrument its unique tonal qualities or timbre. We discussed most of these points in greater detail in Chapter 3. We operate under the assumption that the source signature remains constant for the duration of the note.

We assume that at every instant of time each source vector is modulated by some multiplicative factor. Since we have made the previous assumptions that all sources are constant in frequency, and that all frequency components are amplitude-comodulated, this factor scales the amplitude of all frequency components produced by that particular source by the same amount. No individual adjustments of the amplitude of an isolated frequency component are assumed to occur. In other words, one expects that within the duration of a given note the instrument as a whole may get louder or softer, but its properties will not change.

For each source i of the set of r sources, we represent the time series of instantaneous amplitude values by a row vector $\mathbf{m}_i = [m_i(t_1), m_i(t_2), \ldots, m_i(t_m)]$ which we call the modulation vector. Each element of the modulation vector $m_i(t_k)$ represents the amplitude of source i at a

particular instant k of the set of m time points. The product of a source vector by its modulation vector gives the contribution to the auditory scene due to one particular source.

The resultant of all sources is represented by a matrix $\mathbf{B}$ whose rows are frequency bands and whose columns are time points. This matrix can be thought of as being similar to a spectrogram. We have:

(4.1) $$\mathbf{B} = \mathbf{s}_1\mathbf{m}_1 + \mathbf{s}_2\mathbf{m}_2 + \ldots + \mathbf{s}_r\mathbf{m}_r$$

We can consolidate all the r column vectors $\mathbf{s}_i$ into a matrix $\mathbf{S}$, and all the r row vectors $\mathbf{m}_i$ into a matrix $\mathbf{M}$. We then have compactly $\mathbf{B} = \mathbf{SM}$ as shown below for a 2 source case.

(4.2) $$\begin{bmatrix} \\ \mathbf{B} \\ \\ \end{bmatrix} = \begin{bmatrix} | & | \\ \mathbf{s}_1 & \mathbf{s}_2 \\ | & | \end{bmatrix} \times \begin{bmatrix} - & \mathbf{m}_1 & - \\ - & \mathbf{m}_2 & - \end{bmatrix}$$

4.4 Determination of Number of Sources

We maintain that for a system of constant-frequency, amplitude-modulated sources with no noise present, the number of sources is given by the rank of $\mathbf{B}$. The reason is that since each column of $\mathbf{B}$ is composed of a linear combination of the source vectors (columns of $\mathbf{S}$), each must lie in the space spanned by those vectors. The rank of a matrix gives the dimension of the column or row space.

4.5 Source Identification

Our goal, then, is given $\mathbf{B}$ try to find $\mathbf{S}$ and $\mathbf{M}$. One can immediately see that $\mathbf{S}$ and $\mathbf{M}$ cannot be uniquely determined to better than a multiplicative factor and a permutation of the columns of $\mathbf{S}$ and rows of $\mathbf{M}$. I.e., if one multiplies $\mathbf{S}$ by an arbitrary scalar, and divides $\mathbf{M}$ by the same scalar the result will be the same. Similarly, if one makes a permutation of the columns of $\mathbf{S}$ (by postmultiplying by a permutation matrix $\mathbf{P}$), and makes a corresponding inverse permutation on the rows of $\mathbf{M}$ (by premultiplication by $\mathbf{P}^{-1}$), the product will be the same. However, neither of these is significant. To eliminate the multiplicative ambiguity one need only normalize the columns of $\mathbf{S}$ to any suitable factor, and adjust the rows of $\mathbf{M}$ by the inverse of that factor. And on closer examination, the permutation ambiguity is not really bothersome

either, as it is merely an ordering issue, i.e., a question of which should be labeled source *1* and which source *2*. The source and modulation characteristics of each source are still the same as they were originally, since even after the permutation operations each column of $\mathbf{S}$ is still multiplied by the same row of $\mathbf{M}$; they both get reordered in the same manner, as we explain later.

The major problem is in the fact that if one finds any valid solution $\mathbf{S}$ for the set of spectral signatures, then any linearly-independent combination of the columns of $\mathbf{S}$ is also a solution. I.e., let $\mathbf{S}$ and $\mathbf{M}$ be solutions to the equation $\mathbf{B} = \mathbf{SM}$. Let $\mathbf{A}$ be any invertible matrix. Form the matrix $\mathbf{S}' = \mathbf{SA}$ in which each column is a linear combination of the columns of the original $\mathbf{S}$. If we now compute $\mathbf{M}' = \mathbf{A}^{-1}\mathbf{M}$, then we have:

$$\mathbf{S}'\mathbf{M}' = \mathbf{SAA}^{-1}\mathbf{M} = \mathbf{SM} = \mathbf{B} \tag{4.3}$$

So the set of solutions must be narrowed if comodulation is to give useful results. We seek a constraint that sufficiently limits the solution set, but which is physically meaningful. We propose the following: All elements of matrices $\mathbf{S}$ and $\mathbf{M}$ must be nonnegative. Since the columns of $\mathbf{S}$ represents the source signatures, it is reasonable that all elements be nonnegative. Physically that corresponds to the fact that $\mathbf{S}$ can be thought of as a measure of the power spectral density of the sources which is a nonnegative quantity. Similarly, we lose no physically meaningful information by requiring $\mathbf{M}$ to be nonnegative, since it represents the source strength at each instant which is an unsigned quantity. We are doing nothing more than discarding the negative solution of a quadratic when it is not physically meaningful.

4.6 Uniqueness Theorem for Non-Negative Matrix Factorization

We now prove the following theorem. If a matrix $\mathbf{B}$ is formed as the product of two nonnegative matrices $\mathbf{S}$ and $\mathbf{M}$ so that $\mathbf{B} = \mathbf{SM}$, there exists no other pair of nonnegative matrices into which $\mathbf{B}$ can be factored, except for those matrices that are simply permutations and scalings of the rows and columns of $\mathbf{S}$ and $\mathbf{M}$, provided the following conditions hold: Every column of $\mathbf{S}$ must contain at least one nonvanishing element for which the elements in the corresponding position of all other columns vanish. Similarly, each row of $\mathbf{M}$ must contain at least one nonvanishing element for which the elements in the corresponding position of all other rows vanish.

We start by proving the case where $\mathbf{S}$ has 2 columns and $\mathbf{M}$ has 2 rows. We then extend to the r dimensional case.

4.7 Proof

Let $\mathbf{S}$ be a matrix containing 2 columns all of whose elements are nonnegative and let $\mathbf{M}$ be a matrix containing 2 rows all of whose elements are nonnegative. If we can find an invertible matrix $\mathbf{A}$ such that the matrices $\mathbf{S}' = \mathbf{SA}$ and $\mathbf{M}' = \mathbf{A}^{-1}\mathbf{M}$ both satisfy the nonnegativity constraint, then the decomposition is nonunique. If no such matrix $\mathbf{A}$ can be found, then the decomposition is unique. Note that one can easily show that any other solution $\mathbf{S}'$ to the equation $\mathbf{B} = \mathbf{S}'\mathbf{M}'$ can always be reached from a starting solution $\mathbf{S}$ by multiplication by a square matrix $\mathbf{A}$, so there is no loss of generality in considering $\mathbf{S}'$ to be of the form $\mathbf{S}' = \mathbf{SA}$. To see this we begin with

$$\mathbf{S}'\mathbf{M}' = \mathbf{SM} = \mathbf{B} \tag{4.4}$$

Multiply both sides by $\mathbf{M}'^{\mathbf{T}}$ to get

$$\mathbf{S}'\mathbf{M}'\mathbf{M}'^{\mathbf{T}} = \mathbf{SMM}'^{\mathbf{T}} \tag{4.5}$$

$\mathbf{M}'\mathbf{M}'^{\mathbf{T}}$ is now square and invertible since the rows of $\mathbf{M}'$ are linearly independent by assumption, if they are to represent distinct sources. We can therefore multiply both sides by $(\mathbf{M}'\mathbf{M}'^{\mathbf{T}})^{-1}$ to get

$$\mathbf{S}' = \mathbf{SMM}'^{\mathbf{T}}(\mathbf{M}'\mathbf{M}'^{\mathbf{T}})^{-1} \tag{4.6}$$

which is of the form $\mathbf{S}' = \mathbf{SA}$ where

$$\mathbf{A} = \mathbf{MM}'^{\mathbf{T}}(\mathbf{M}'\mathbf{M}'^{\mathbf{T}})^{-1} \tag{4.7}$$

We now turn to the question of whether the pair $\mathbf{S}' = \mathbf{SA}$ and $\mathbf{M}' = \mathbf{A}^{-1}\mathbf{M}$ satisfy the nonnegativity constraint. We proceed by examining the effect of $\mathbf{A}$ and $\mathbf{A}'$ on the rows and columns of $\mathbf{S}$ and $\mathbf{M}$, respectively. It is well known in linear algebra that any invertible square matrix $\mathbf{A}$ can be written as a product of elementary matrices

$$\mathbf{A} = \mathbf{E}_n\mathbf{E}_{n-1}\ldots\mathbf{E}_1 \tag{4.8}$$

where each elementary matrix $\mathbf{E}_i$ produces a single elementary row operation on the matrix immediately to its right ($\mathbf{E}_1$ is assumed to operate on the identity matrix $\mathbf{I}$). The possible operations are:

1) Swapping of one row with another (permutation).

2) Multiplication of a row by a scalar.

3) Addition of a multiple of one row to another.

We have then:

(4.9) $$\mathbf{S}' = \mathbf{S}\mathbf{E}_n\mathbf{E}_{n-1}\ldots\mathbf{E}_1$$

and

(4.10) $$\mathbf{M}' = \mathbf{E}_1^{-1}\mathbf{E}_2^{-1}\ldots\mathbf{E}_n^{-1}\mathbf{M}$$

We note that in most texts the elementary matrices are discussed in terms of row operations only, and multiply on the left. However they produce the same effect on columns when multiplying on the right, as can be seen by taking transposes. We examine each operation one at a time, considering the form of the corresponding elementary matrix and its effect on the columns of $\mathbf{S}$ when multiplied on the right; and in addition, we consider the form of the inverse of each elementary matrix and its effect on the rows of $\mathbf{M}$ when multiplied on the left. In the case of column permutations, the form of the elementary matrix is simply the corresponding permutation operation performed on $\mathbf{I}$. The inverse is simply the reverse permutation (i.e., the one that will give back the original column order). Since both the elementary matrix and its inverse contain no negative elements, no negative numbers are generated by either one when right multiplying on $\mathbf{S}$ to get $\mathbf{S}'$, or when left multiplying on $\mathbf{M}$ to get $\mathbf{M}'$. Hence a matrix $\mathbf{S}'$ formed as a permutation of the columns of $\mathbf{S}$ will yield another nonnegative solution when multiplied by a matrix $\mathbf{M}'$ formed by the inverse permutation of the rows of $\mathbf{M}$. However, as noted above, a solution found in this manner is a trivial reordering of the same sources as in the original solution, since a particular column of $\mathbf{S}$ is still multiplied by the same row of $\mathbf{M}$ as before, only they are now row *2* and column *2* instead of row *1* and column *1*. This is true due to the fact that every permutation matrix $\mathbf{P}$ has the property that its inverse equals its transpose

(4.11) $$\mathbf{P}^{-1} = \mathbf{P}^{\mathbf{T}}$$

So starting with the equations

(4.12) $$\mathbf{S}' = \mathbf{SP}$$

(4.13) $$\mathbf{M}' = \mathbf{P}^{-1}\mathbf{M}$$

and taking transposes of the last equation we have

(4.14) $$\mathbf{M}'^{\mathbf{T}} = \mathbf{M}^{\mathbf{T}}(\mathbf{P}^{-1})^{\mathbf{T}} = \mathbf{M}^{\mathbf{T}}\mathbf{P}$$

and we see that $\mathbf{P}$ has the same effect on the rows of $\mathbf{M}$ as on the columns of $\mathbf{S}$, i.e., they are shuffled in the same sequence, and the pairings do not change. We may therefore ignore these solutions in a discussion of uniqueness, since although they are nonnegative, they contain no new possibilities for describing the sources.

For the next category of elementary column operation, the case of multiplying a column by a scalar, the appropriate $\mathbf{E}$ matrix is formed by multiplying the corresponding column of $\mathbf{I}$ by that scalar. To form the inverse $\mathbf{E}$ matrix, the same column of $\mathbf{I}$ is multiplied by the reciprocal of that scalar. If the scalar is negative we have no threat to uniqueness. The corresponding column will be negative in both $\mathbf{E}$ and $\mathbf{E}^{-1}$ and the effect on $\mathbf{S}'$ and $\mathbf{M}'$ will be to create negative values in both, since the original $\mathbf{S}$ and $\mathbf{M}$ were assumed nonnegative. Hence no new solutions can be generated that satisfy the nonnegativity constraint. However, in the case of a positive scalar, both $\mathbf{E}$ and $\mathbf{E}^{-1}$ will be nonnegative, so the products $\mathbf{S}'$ and $\mathbf{M}'$ will perforce also be nonnegative, seemingly indicating the existence of additional solutions meeting the nonnegativity constraint, and dealing a blow to uniqueness. However, closer examination reveals the following. When an entire column of $\mathbf{S}$ is multiplied by a scalar, that is equivalent to boosting the energy in all bands of that source by that scalar factor. But when $\mathbf{M}'$ is formed by multiplying by $\mathbf{E}^{-1}$ on the left, the corresponding modulation vector for that source (row of $\mathbf{M}$) will be multiplied by the reciprocal of the same factor. This simply means that one can alternately describe the same source as having twice the band energies, but half the modulation strength which yields the same physical output. The ambiguity can be made to disappear by simply normalizing the source signatures (columns of $\mathbf{S}$) one at a time to a suitable value such as by setting each ℓ^2 column norm to one, and making the inverse adjustment (reciprocal of the

normalization factor for that source) to the corresponding modulation vector, which weeds out these uninformative solutions and thereby preserves uniqueness.

The final possible column operation that needs to be considered is the addition of a multiple of one column to another. The corresponding $\mathbf{E}$ matrix for this operation is, as before, the matrix which results when this operation is performed on $\mathbf{I}$. If for example, ε times column *1* is to be added to column 2, we would have

$$\mathbf{E} = \begin{bmatrix} 1 & \varepsilon \\ 0 & 1 \end{bmatrix} \tag{4.15}$$

The inverse of this operation is the addition of $-\varepsilon$ times column *1* to column *2*. So the form of $\mathbf{E}^{-1}$ is

$$\mathbf{E}^{-1} = \begin{bmatrix} 1 & -\varepsilon \\ 0 & 1 \end{bmatrix} \tag{4.16}$$

Let us now consider the effect upon $\mathbf{S}$ from being multiplied on the right by $\mathbf{E}$ to give $\mathbf{S}'$. Clearly, since $\mathbf{E}$ has no negative elements (with no loss of generality we assume that ε is positive), $\mathbf{S}'$ will not have any, either. However, in this case, $\mathbf{E}^{-1}$ will have a negative element, $-\varepsilon$. We need to study the effect of this negative element on the product $\mathbf{M}' = \mathbf{E}^{-1}\mathbf{M}$. Writing the equations for the individual rows $\mathbf{m}_i$ and $\mathbf{m}_i'$ of $\mathbf{M}$ and $\mathbf{M}'$, respectively, we have:

$$\begin{aligned} \mathbf{m}_1' &= 1\mathbf{m}_1 - \varepsilon\mathbf{m}_2 \\ \mathbf{m}_2' &= 0\mathbf{m}_1 + 1\mathbf{m}_2 \end{aligned} \tag{4.17}$$

If there is no time point where $\mathbf{m}_1$ vanishes while the corresponding element of $\mathbf{m}_2$ is nonvanishing, then one can always find a nonzero ε small enough so $\mathbf{m}_1'$ is positive everywhere, and the solution $\mathbf{B} = \mathbf{SM}$ is nonunique, since $\mathbf{M}'$ also satisfies the nonnegativity constraint. However, if $\mathbf{m}_1$ is zero at some point where $\mathbf{m}_2$ is nonzero, and vice versa (by symmetry), then there is no way to make all elements of $\mathbf{M}'$ positive unless ε is zero, in which case $\mathbf{E} = \mathbf{E}^{-1} = \mathbf{I}$, and there exists only one solution. By duality this holds for the columns of S. (Start by assuming that ε is negative, and $-\varepsilon$ is positive, and compute $\mathbf{S}' = \mathbf{SE}$ where, as before,

(4.18) $$\mathbf{E} = \begin{bmatrix} 1 & \varepsilon \\ 0 & 1 \end{bmatrix}$$

and use similar argument.)

4.8 Interpretation

Before continuing with the proof for the multidimensional case, we pause to consider the implications of the theorem. The impact of the theorem on source separation is as follows: In order for a set of amplitude-modulated constant-frequency sources to be uniquely separable on the basis of amplitude comodulation, each source must have at least one point in time at which it is a soloist, all other sources being silent. In addition, each source must have at least one frequency point in its spectral signature which it shares with no other sources, meaning that the frequency is unique to that source, and at no time does any other source emit that frequency.

4.9 Multidimensional Case

The extension to the multidimensional case is straightforward. Every elementary matrix $\mathbf{E}_k$ which has a single off diagonal element ε_{ij} in row i and column j will have an inverse $\mathbf{E}_k^{-1}$ which has a single off diagonal element $-\varepsilon_{ij}$ in the same position. This will force $\mathbf{M}'$ to contain a negative element if at some time instant an element of row j of $\mathbf{M}$ vanishes while the corresponding element in row i does not, unless ε_{ij} is zero. If for each pair of rows there is such a point in which an element of one row of the pair vanishes while the corresponding element in the other does not, then the only way to prevent the occurrence of any negative elements in $\mathbf{M}'$ is for the off diagonal element ε_{ij} in each matrix $\mathbf{E}_k$ to be zero. In that case

(4.19) $$\mathbf{E}_k = \mathbf{E}_k^{-1} = \mathbf{I} \qquad \forall k,$$

so

(4.20) $$\mathbf{S}' = \mathbf{SA} = \mathbf{SE}_n\mathbf{E}_{n\text{-}1}\ldots\mathbf{E}_1 = \mathbf{S},$$

and

(4.21) $$\mathbf{M}' = \mathbf{A}^{-1}\mathbf{M} = \mathbf{E}_1^{-1}\mathbf{E}_2^{-1}\ldots\mathbf{E}_n^{-1}\mathbf{M} = \mathbf{M},$$

and the solution is unique.

One final caveat which must be considered in the multidimensional case is whether a negative element in $\mathbf{M}'$ or $\mathbf{S}'$ introduced by one of the earlier elementary matrices in the series $\mathbf{E}_n\mathbf{E}_{n-1}\ldots\mathbf{E}_1$ could be canceled out by one of the later matrices, such that the end result would be a pair of matrices $\mathbf{S}'$ and $\mathbf{M}'$ which are materially different from $\mathbf{S}$ and $\mathbf{M}$ but yet are still able to pass the nonnegativity test. In actuality, that could never happen, because under the conditions of the theorem each source must have a frequency point and a time point at which all other sources vanish. And since the only effect of this third category of elementary matrices when multiplying on the right of $\mathbf{S}$ is to add a multiple of one column to another, (or a multiple of one row to another, in the case of the inverses multiplying on the left of $\mathbf{M}$) such a negative value cannot be removed by addition of a multiple of any other column of $\mathbf{S}$ (or row of $\mathbf{M}$, in the latter case) since all of them will have a zero in the corresponding position.

4.10 Simple Proof for Square Case

In order to make the theorem a bit more plausible, we offer a very simple proof of uniqueness for the case of the nonnegative decomposition of a square matrix, which is a special case of the general theorem which holds for any size matrices. Consider an arbitrary $n \times n$ source matrix $\mathbf{S}$, and a modulation matrix $\mathbf{M}$ equal to the identity matrix $\mathbf{I}$ of the same size.

(4.22) $$\mathbf{M} = \mathbf{I}$$

Then for this case,

(4.23) $$\mathbf{B} = \mathbf{SM} = \mathbf{SI} = \mathbf{S}.$$

We now want to decompose $\mathbf{B}$ into $\mathbf{S}$ and $\mathbf{M}$ and want to know if the solution is unique. We have clearly at least 2 choices.

(4.24) $$\mathbf{B} = \mathbf{SI},$$

or

(4.25) $$\mathbf{B} = \mathbf{IS}.$$

These two solutions are different. In the first, the source is $\mathbf{S}$ while the modulation is $\mathbf{I}$, while in the second the source is $\mathbf{I}$ while the modulation is $\mathbf{S}$. The only way in which the two solutions can be the same, is if

$$\mathbf{S} = \mathbf{I} \tag{4.26}$$

(up to a scaling and permutation which we ignore because of their triviality, as explained earlier). This means that to guarantee uniqueness, the form of $\mathbf{S}$ must be similar to the form of $\mathbf{I}$ meaning that it contains only one nonzero element in every row and every column. This would then match the conditions of the theorem as stated above, that for uniqueness each column of $\mathbf{S}$ must have at least one location at which it is nonzero and at which all remaining columns are zero at the corresponding location; and that each row of $\mathbf{M}$ must similarly have at least one location at which it is nonzero and at which all remaining rows are zero at the corresponding location.

4.11 Examples

In order to more clearly illustrate the implications of the theorem, we first provide an example of a source/modulation matrix pair which does not meet the criteria of the theorem, and whose product, therefore, has multiple sets of nonnegative factors. We then provide an example of a source/modulation pair which does meet the criteria of the theorem whose product has no other sets of nonnegative factors.

4.11.1 Example 1: Non-Unique Non-Negative Decomposition

Start with the source vectors $\mathbf{s}_1$ and $\mathbf{s}_2$.

$$\mathbf{s}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \tag{4.27}$$

$$\mathbf{s}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{4.28}$$

Form S from the column vectors $\mathbf{s}_1$ and $\mathbf{s}_2$.

(4.29) $$\mathbf{S} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Let the modulation vectors $\mathbf{m}_1$ and $\mathbf{m}_2$ be as follows:

(4.30) $$\mathbf{m}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

(4.31) $$\mathbf{m}_2 = \begin{bmatrix} 1 & .9 & .8 & .7 & .6 & .5 & .4 & .3 & .2 & .1 & 0 \end{bmatrix}$$

Form M from the row vectors $\mathbf{m}_1$ and $\mathbf{m}_2$.

(4.32) $$\mathbf{M} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & .9 & .8 & .7 & .6 & .5 & .4 & .3 & .2 & .1 & 0 \end{bmatrix}$$

Compute $\mathbf{B} = \mathbf{SM}$ giving

(4.33) $$\mathbf{B} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & .9 & .8 & .7 & .6 & .5 & .4 & .3 & .2 & .1 & 0 \end{bmatrix}$$

However, the product of the pair $\mathbf{S}'$ and $\mathbf{M}'$ where

(4.34) $$\mathbf{S}' = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$$

and

(4.35) $$\mathbf{M}' = \begin{bmatrix} 0 & .1 & .2 & .3 & .4 & .5 & .6 & .7 & .8 & .9 & 1 \\ 1 & .9 & .8 & .7 & .6 & .5 & .4 & .3 & .2 & .1 & 0 \end{bmatrix}$$

gives the same matrix $\mathbf{B}$.

The reason for the nonuniqueness is because in the first source/modulation pair, the second row of $\mathbf{M}$ has no time point at which it is nonvanishing while the first row is vanishing. In other words, at no time is source *2* a soloist, because source *1* is constantly playing. In the second pair, although each of the sources in $\mathbf{M}'$ is now a soloist at one instant, however, there is no band in $\mathbf{S}'$ where source *1* has a frequency component and source *2* doesn't, since source *2* has a component in every band.

As an aside, note that as explained earlier, the second set of sources, $\mathbf{S}'$, was formed from the first set by the following simple linear combination:

$$\begin{aligned} \mathbf{s}_1' &= 1\mathbf{s}_1 + 0\mathbf{s}_2 \\ \mathbf{s}_2' &= 1\mathbf{s}_1 + 1\mathbf{s}_2 \end{aligned} \tag{4.36}$$

4.11.2 Example 2: Unique Non-Negative Decomposition

We now show a source/modulation pair which does meet the requirements of the theorem, and has no other nonnegative factorization.

$$\mathbf{S} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \tag{4.37}$$

and

$$\mathbf{M} = \begin{bmatrix} 0 & .1 & .2 & .3 & .4 & .5 & .6 & .7 & .8 & .9 & 1 \\ 1 & .9 & .8 & .7 & .6 & .5 & .4 & .3 & .2 & .1 & 0 \end{bmatrix} \tag{4.38}$$

As before compute $\mathbf{B} = \mathbf{SM}$ giving

$$\mathbf{B} = \begin{bmatrix} 0 & .1 & .2 & .3 & .4 & .5 & .6 & .7 & .8 & .9 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & .9 & .8 & .7 & .6 & .5 & .4 & .3 & .2 & .1 & 0 \end{bmatrix} \tag{4.39}$$

In this example, both source *1* and source *2* have frequency bands not shared by their counterpart, and both also have a point in time at which they are soloists.

4.12 Solution by Inspection

4.12.1 Method

While we have thus far proved the existence and uniqueness of a solution if the proper constraints are met, we have not discussed how to find this solution. It turns out that if the conditions which we have set forth are met, then factorization can be performed by inspection!

Let us look at a previous example.

(4.40) $$\mathbf{B} = \begin{bmatrix} 0 & .1 & .2 & .3 & .4 & .5 & .6 & .7 & .8 & .9 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & .9 & .8 & .7 & .6 & .5 & .4 & .3 & .2 & .1 & 0 \end{bmatrix}$$

The theorem states for the nonnegative factorization to be unique, each source must be a soloist for at least one time instant and have at least one frequency component which is not shared by any other source.

The effect of these conditions on the product will be that embedded somewhere in the product matrix will be the elements of a matrix with form similar to the identity matrix $\mathbf{I}$, of size corresponding to the number of columns in S and number of rows in M. *Those rows which contain elements of* $\mathbf{I}$ *are proportional to the modulation vectors. Those columns which contain elements of* $\mathbf{I}$ *are proportional to the source signature vectors.*

In the matrix above, the top left corner contains a zero of $\mathbf{I}$, therefore, the top row is proportional to one of the modulation vectors. In fact, it is the rising ramp m_1 we used in the example. The bottom right corner contains another zero of $\mathbf{I}$, therefore, the bottom row is proportional to another of the modulation vectors. It is the falling ramp m_2 we used in the example.

(4.41) $$\mathbf{m}_1 = \begin{bmatrix} 0 & .1 & .2 & .3 & .4 & .5 & .6 & .7 & .8 & .9 & 1 \end{bmatrix}$$

(4.42) $$\mathbf{m}_2 = \begin{bmatrix} 1 & .9 & .8 & .7 & .6 & .5 & .4 & .3 & .2 & .1 & 0 \end{bmatrix}$$

Since a zero of $\mathbf{I}$ appears in the left column of the matrix, therefore the left column is proportional to one of the source vectors. It is s_2, one of the source vectors we used earlier. And since a zero of $\mathbf{I}$ appears in the right column of the matrix, therefore the right column is proportional to another of the source vectors. It is s_1, the other source vector we used earlier.

(4.43) $$\mathbf{s}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

(4.44) $$\mathbf{s}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

4.12.2 Formal Proof

The proof of this is easiest to see by using block matrix multiplication. If one partitions a matrix into separate rectangular blocks, they can be multiplied in the same manner as individual elements are multiplied in regular matrix multiplication. (In general, there are some conditions on the shapes of the blocks, but for our purposes, one can easily verify that the method works the way we will use it.)

To begin we note that the soloist conditions on the modulation vectors can be stated as a requirement that there be an identity matrix appended or embedded into the rest of the modulation vector. I.e., we can write the matrix M as

$$\mathbf{M} = \begin{bmatrix} \mathbf{I} & \mathbf{M}_R \end{bmatrix} \tag{4.45}$$

where $\mathbf{I}$ is the appropriate $r \times r$ sized identity matrix and $\mathbf{M}_R$ a block containing all the remaining columns of $\mathbf{M}$. The reason for this is that having a single source on and all other sources off for one instant of time means that in the column of $\mathbf{M}$ representing that time instant there may be only one nonzero element, corresponding to the single source which is on. This is equivalent to one of the columns of an identity matrix. Since each source must have one such instant, there must be present all the columns of an identity matrix. Note that although the off-diagonal elements must be 0, the diagonal elements do not necessarily have to be 1. However, this will not affect anything, other than multiplying the entire row in question by a constant in the final result. Also note that although the columns of the identity matrix need not be next to each other in the usual order, this will also not affect anything, since transposing the columns to form a standard shaped identity matrix will only transpose those same columns in the final result. They can then be transposed back to where they were originally.

Similarly we can write the unique frequency conditions on the source signatures as

$$\mathbf{S} = \begin{bmatrix} \mathbf{I} \\ \mathbf{S}_R \end{bmatrix} \tag{4.46}$$

for exactly the same reasons. In this case since each source must have a unique frequency shared by no other, there will be a row corresponding to the unique frequency of that source which contains a nonzero element in the column containing the spectral signature of that

source, and with zeros at all other positions. That row is equivalent to a row of $\mathbf{I}$. Since all sources must have such a frequency, there must be present all the rows of $\mathbf{I}$.

We then have

$$\mathbf{B} = \mathbf{SM} = \begin{bmatrix} \mathbf{I} \\ \mathbf{S}_R \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{M}_R \end{bmatrix} \tag{4.47}$$

Multiplying these two matrices, we obtain

$$= \begin{bmatrix} \mathbf{I} & \mathbf{M}_R \\ \mathbf{S}_R & \mathbf{S}_R\mathbf{M}_R \end{bmatrix} \tag{4.48}$$

But the first r rows of this block matrix are nothing other than the set of modulation vectors. The first r columns are none other than the source signatures. So the proof is complete.

4.12.3 Implications

An intuitive understanding why this should be true can be gained from thinking about the conditions for the theorem. If we have a frequency which is unique to a source, then we can simply read off the modulation vector by looking at the row in the matrix B which corresponds to that frequency. It will not contain a contribution from any other source, so it will be proportional to the modulation vector of that source, only. Similarly, at a time in which a source is a soloist, no other source makes a contribution. We can therefore read off the relative spectral weights of that source by looking at the column of B which occurs at that point in time. It must be proportional to the source signature.

4.13 Alternating Iterative NNMF Algorithm

We introduce an additional method for performing non-negative matrix factorization on matrices which meet the constraints we have outlined. While the solution by inspection gives the most insight into the nature of the problem and the interaction of the various matrices, the following algorithm is useful for cases in which noise or other non-idealities are present. This is because it finds the best-fitting factorization given the circumstances at hand. The solution by inspection might not be as useful if the rows are not exact multiples of the same modulation vector, or the columns are not perfect multiples of the identical source signatures. In those cases, the basis vectors might not be immediately recognizable. In Chapter 3 we discussed that this

might indeed be the case for real instruments and real voices. We noted that harmonics do not always move in lock step for reasons we discussed.

We present without formal proof the following alternating iterative algorithm which decomposes a matrix into two nonnegative factors provided the uniqueness conditions of the theorem are met. If the conditions are not met and multiple solutions exist, it may or may not converge to some valid solution, but the results are unpredictable.

The algorithm requires the user to specify the rank of the matrix to be decomposed. This is used to set the sizes of the factors, since an $n \times m$ matrix can be factored into two matrices of arbitrary sizes $n \times q$ and $q \times m$ where q can be any integer.

The steps are as follows:

1) Start with matrix $\mathbf{B}$ of size $n \times m$. Specify rank of $\mathbf{B}$ as r.

2) Size $\mathbf{S}$ as $n \times r$, and $\mathbf{M}$ as $r \times m$.

3) Initialize $\mathbf{S}$ with positive uniformly distributed [0,1] random values. (Matlab®: `rand`).

4) Check rank of $\mathbf{S}$. If less than r, add random values (as before) onto $\mathbf{S}$. (This prevents errors from singular matrices.)

5) Compute least squares solution for $\mathbf{M}$ to $\mathbf{SM} = \mathbf{B}$. (Matlab®: `M=S\B` where '`\`' is the Matlab® *mldivide* operator).

6) Set any negative values in $\mathbf{M}$ to 0.

7) Check rank of $\mathbf{M}$. If less than r, add random values (as before) onto $\mathbf{M}$. (This prevents errors from singular matrices.)

8) Compute least squares solution for $\mathbf{S}$ to $\mathbf{SM} = \mathbf{B}$. (Matlab®: `S=B/M` where '`/`' is the Matlab® *mrdivide* operator).

9) Set any negative values in $\mathbf{S}$ to 0.

10) Compute ℓ^2 norm of each column of $\mathbf{S}$ and divide each column by its respective normalization factor so each column is separately normalized to 1.

11) Multiply the corresponding rows of $\mathbf{M}$ by the same set of normalization factors used for columns of $\mathbf{S}$ so that each product of column i of $\mathbf{S}$ multiplied by row i of $\mathbf{M}$ is unchanged.

12) Iterate from Step 4 until there are no more negative values in $\mathbf{S}$ or $\mathbf{M}$.

4.14 Results: Amplitude-Modulated Harmonic Sets

4.14.1 Description of Signals

Figure 22 shows results for a synthesized mixture of 2 sets of 5-harmonic sequences. The first had a fundamental of 200 Hz, and the second a fundamental of 400 Hz. Sampling rate was 10 KHz.

The frequencies of each set were as follows[4]:

	Set 1	Set 2
Fundamental	200	400
2nd Harmonic	400	800
3rd Harmonic	600	1200
4th Harmonic	800	1600
5th Harmonic	1000	2000

Table 2. Frequencies of components in each of two 5-harmonic sets.

Each harmonic in both sets had unity amplitude and zero phase.

In keeping with the conditions of the theorem that each source needs to be a soloist for at least one point in time, the first set is started earlier than the second set, and the second set ends after the first set. In addition, each set has some frequency components which are unique to that set, and not found in the other set. Both sets contain the frequencies 400 Hz and 800 Hz. Spectrograms of the two original sets are shown in the two plots in the left-hand column of the figure. Proceeding to the plot in the second column from left, we show a spectrogram of the sum of the two harmonic sets.

[4] Note: Frequencies of fundamentals and harmonics were adjusted slightly to fall in center of FFT bins (discussed in Section 6.3), which for a 10 KHz sampling rate and length 256 FFT, are at multiples of 39.0625 Hz. Hence, true frequencies of fundamentals fall at 195.3125 Hz and 390.6250 Hz, respectively. This has no impact

4.14.2 Explanation of Results

The two plots in the third column from the left of Figure 22 show the result of the algorithm's attempt to factor the spectrogram of the sum into two modulation vectors and two source signatures. Note the sharp drop in level of the first modulation vector at 0.667 seconds which corresponds to the turn-off step of the first harmonic set. Similarly, the second modulation vector exhibits a sharp rise in level at 0.333 seconds which corresponds to the turn-on step of the second harmonic set. Note that the peaks in the plots of the two recovered spectral signatures correspond with high accuracy to the frequencies of the sources.

The final column shows the reconstruction of the sources from the pair of recovered spectral signatures and modulation vectors. The first spectral signature is multiplied by the first modulation vector to obtain the first recovered source spectrogram, and the second spectral signature is multiplied by the second modulation vector to obtain the second recovered source spectrogram.

4.14.3 Evaluation of Performance and Sources of Error

For this simple case, the results are in excellent agreement with the original spectrograms. Reconstructed spectrograms are virtually indistinguishable from the original plots. The modulation vectors show slight deviations from true step shapes, but this appears to have negligible effect.

Note that we have not created an actual audio recording from the output spectral signature and modulation vectors. Since the algorithm depends on nonnegativity, we must use the absolute or squared value (magnitude only) of the FFT coefficients. Because we discard phase, we cannot obtain a true output signal with this method. It is, of course, possible to synthesize signals at the frequency of the peaks in the spectral signatures, and to use the modulation vectors to modulate those signals, thus giving an artificial reconstruction of the sources. More exact methods for reproducing a signal from a spectrogram or other auditory representations have been examined by (Slaney, Naar and Lyon, 1994), and (Yang, Wang and Shamma, 1992).

---

on discussion.

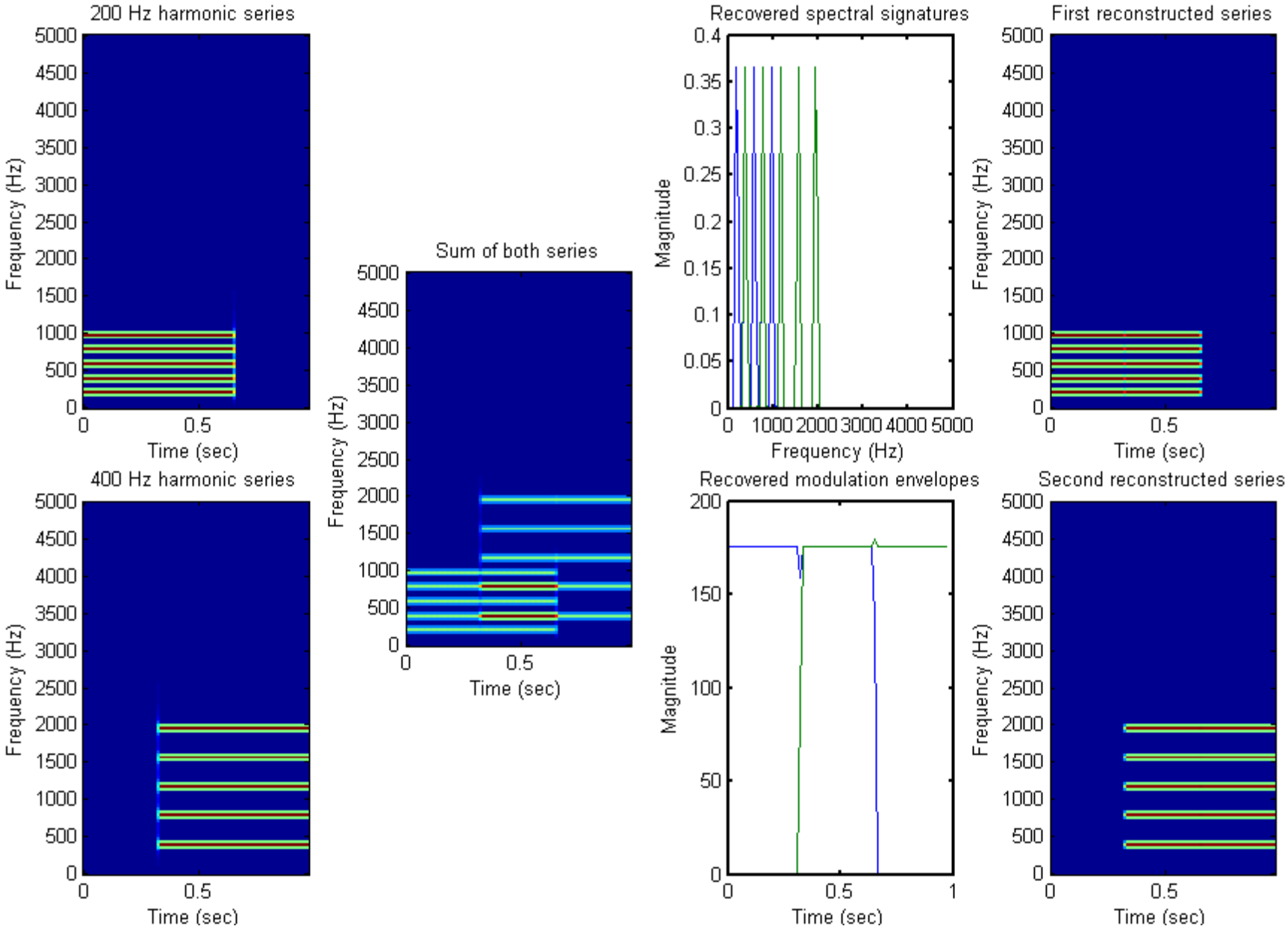
200 Hz harmonic series
Frequency (Hz)
Time (sec)
400 Hz harmonic series
Frequency (Hz)
Time (sec)
Sum of both series
Frequency (Hz)
Time (sec)
Recovered spectral signatures
Magnitude
Frequency (Hz)
Recovered modulation envelopes
Magnitude
Time (sec)
First reconstructed series
Frequency (Hz)
Time (sec)
Second reconstructed series
Frequency (Hz)
Time (sec)

Figure 22. (Previous page). First Column, Top: Original Spectrogram of 200 Hz 5-harmonic series. First Column, Bottom: Original Spectrogram of 400 Hz 5-harmonic series. Second Column: Spectrogram of mixture of both series. Third Column, Top: Recovered spectral signatures of each series. Third column, bottom: Recovered modulation vectors of each series. Fourth Column, Top: Separated and reconstructed spectrogram of first 5-harmonic series produced by taking outer product of recovered blue spectral signature in top plot of third column with recovered blue modulation vector in bottom plot of third column. Fourth Column, Bottom: Separated and reconstructed spectrogram of second 5-harmonic series produced by taking outer product of recovered green spectral signature in top plot of third column with recovered green modulation vector in bottom plot of third column.

4.15 Results: Mixture of Clarinet and Oboe

4.15.1 Description of Signals

We next looked at a more realistic auditory scene produced by recordings of actual musical instruments from the McGill University Master Samples (MUMS) collection. We selected the recording of the bass clarinet playing A#2 and the oboe playing C5, the two of which we have discussed in Chapter 3. Both of these instruments do not exhibit frequency variation within a note, and meet the criteria of our theorem.

The sampling rates were 22,050 Hz for each. The onset of the oboe was delayed slightly with respect to that of the clarinet to meet the conditions of the theorem, as before.

The results are shown in Figure 23. Note that the clarinet, being a very low-pitched instrument, has a low fundamental, and very tightly spaced harmonics. The oboe has a higher fundamental, and thus more loosely spaced harmonics.

4.15.2 Explanation of Figures and Results

As before, the first column in the figure shows the original spectrograms of the two instruments. The second column shows the spectrogram of the sum. The next column shows the recovered spectral signatures and the modulation vectors for each instrument. The last column shows the product of the modulation vectors and the spectral signatures.

The format of the plots for this and succeeding cases are presented in exactly the same format as in the previous case, and need not be re-explained.

4.15.3 Evaluation of Performance and Sources of Error

We note visually that separation is fairly good. One of the reconstructed sources has mainly closely spaced harmonics, corresponding to the clarinet, while the other has more widely

spaced harmonics, corresponding to the oboe. However, there is slightly more residual energy at the off times than in the previous trial. In addition, there is more residual energy from the competing instrument. We can attribute some of this to noise. However, in actuality, there is another reason, as well. We have assumed that instruments are in fact perfectly comodulated, i.e., that the amplitudes of all harmonics move in lock step. However, as we discussed in the Chapter 3, that is not exactly the case. Harmonics of a given instrument do not always start and stop at exactly the same times, and are not perfectly correlated.. This may also account for some of the error, as it then makes it impossible to find a single modulation vector that fits all the bands, and instead the algorithm will be forced to average or find the best fit.

4.15.4 Tremolo

We note that the modulation vectors computed by the algorithm demonstrate fairly prominent amplitude fluctuations. These correspond well to the sound actually heard in the recording. As we discussed in Chapter 3, this musical technique is known as tremolo, in which a musician varies the amplitude of the instrument rhythmically to give a pleasing effect. It is especially pronounced in the oboe. This gives further confidence in the ability of the algorithm to capture realistic acoustic parameters.

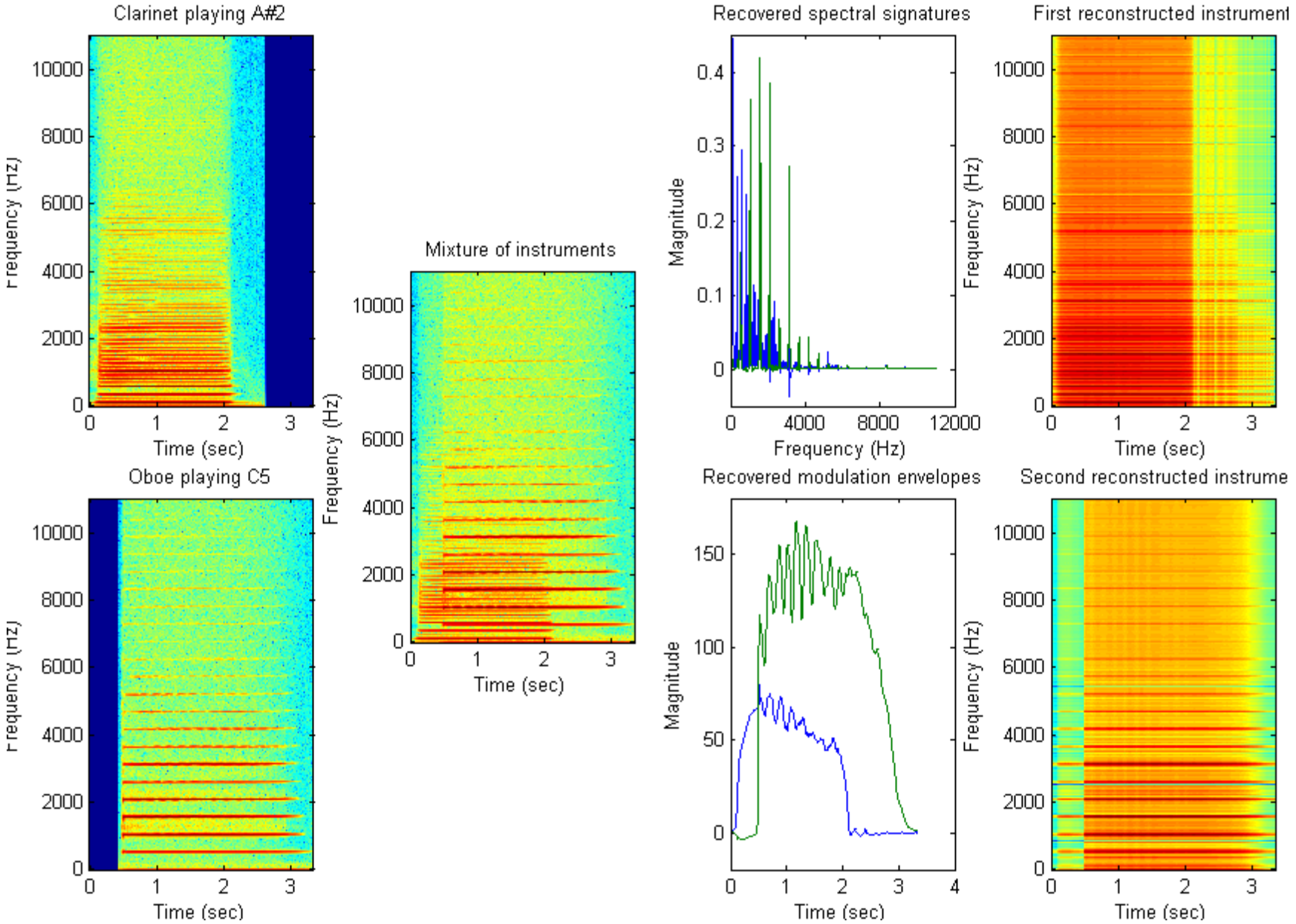
Clarinet playing A#2
Frequency (Hz)
Time (sec)
Oboe playing C5
Frequency (Hz)
Time (sec)
Mixture of instruments
Frequency (Hz)
Time (sec)
Recovered spectral signatures
Magnitude
Frequency (Hz)
Recovered modulation envelopes
Magnitude
Time (sec)
First reconstructed instrument
Frequency (Hz)
Time (sec)
Second reconstructed instrume
Frequency (Hz)
Time (sec)

Figure 23. (Previous page). First Column, Top: Original Spectrogram of clarinet playing A#2. First Column, Bottom: Original Spectrogram of oboe playing C5. Second Column: Spectrogram of mixture of both instruments. Third Column, Top: Recovered spectral signatures of each instrument. Third column, bottom: Recovered modulation vectors of each instrument. Fourth Column, Top: Separated and reconstructed spectrogram of first instrument produced by taking outer product of recovered green spectral signature in top plot of third column with recovered green modulation vector in bottom plot of third column. Result bears similarity to clarinet in top plot of first column in terms of timing and harmonic spacing. Fourth Column, Bottom: Separated and reconstructed spectrogram of second instrument produced by taking outer product of recovered blue spectral signature in top plot of third column with recovered blue modulation vector in bottom plot of third column. Result bears similarity to oboe in bottom plot of first column in terms of timing and harmonic spacing.

4.16 Results: Mixture of Violin and Oboe

4.16.1 Description of Signals

We next discuss the performance of the algorithm in the case of a frequency-modulated instrument. We use the recording of the violin playing G6 which we have analyzed in Chapter 3. As before, the recording was obtained from the McGill University Master Samples database. The note contains prominent vibrato. We mixed that with the recording of the oboe that was used in the previous trial.

The sampling rates were 22,050 Hz for each. As previously, the onset of the oboe was delayed slightly with respect to that of the violin to meet the soloist conditions of the theorem. However, the constant-frequency requirement of the theorem is deliberately not satisfied by our choice of the violin due to the vibrato, as we will discuss.

4.16.2 Evaluation of Performance and Sources of Error

As can be seen in Figure 24, the results are extremely poor, as predicted. Looking at the rightmost column, one does not see any evidence that separation has occurred. Both plots look essentially the same, as far as the amplitude and frequency of the harmonics is concerned.

The reason for this is that the frequency variation prevents the use of a constant source-signature term to describe the relative amplitude of the spectral components. In any one band, the amplitude will rise and fall with time as a given harmonic enters or leaves the passband of that filter. The previous representation of fixed ratios between the amplitudes of each band no longer applies. As a result, the modulation vectors cannot multiply any one source component for the duration of the note. Instead they see various bands getting louder and softer and turning on and off in a seemingly unrelated manner. As the harmonic leaves one band, it enters

another band, and so one band is falling in amplitude at the same time another is rising in amplitude. The amplitude variation of one band does not correlate with the amplitude variation of other bands.[5] Because of this, the modulation vectors degenerate into useless rapid variations which merely look like noise as can be seen in the lower plot in the 3rd column from the left. When multiplying by the source vectors which themselves are imprecise because of movement in frequency, the result is the unstructured and incorrect reconstruction in the right hand plots.

[5] It may actually be negatively correlated with the adjacent band, and possibly this can be exploited in some manner, although not with the current formulation.

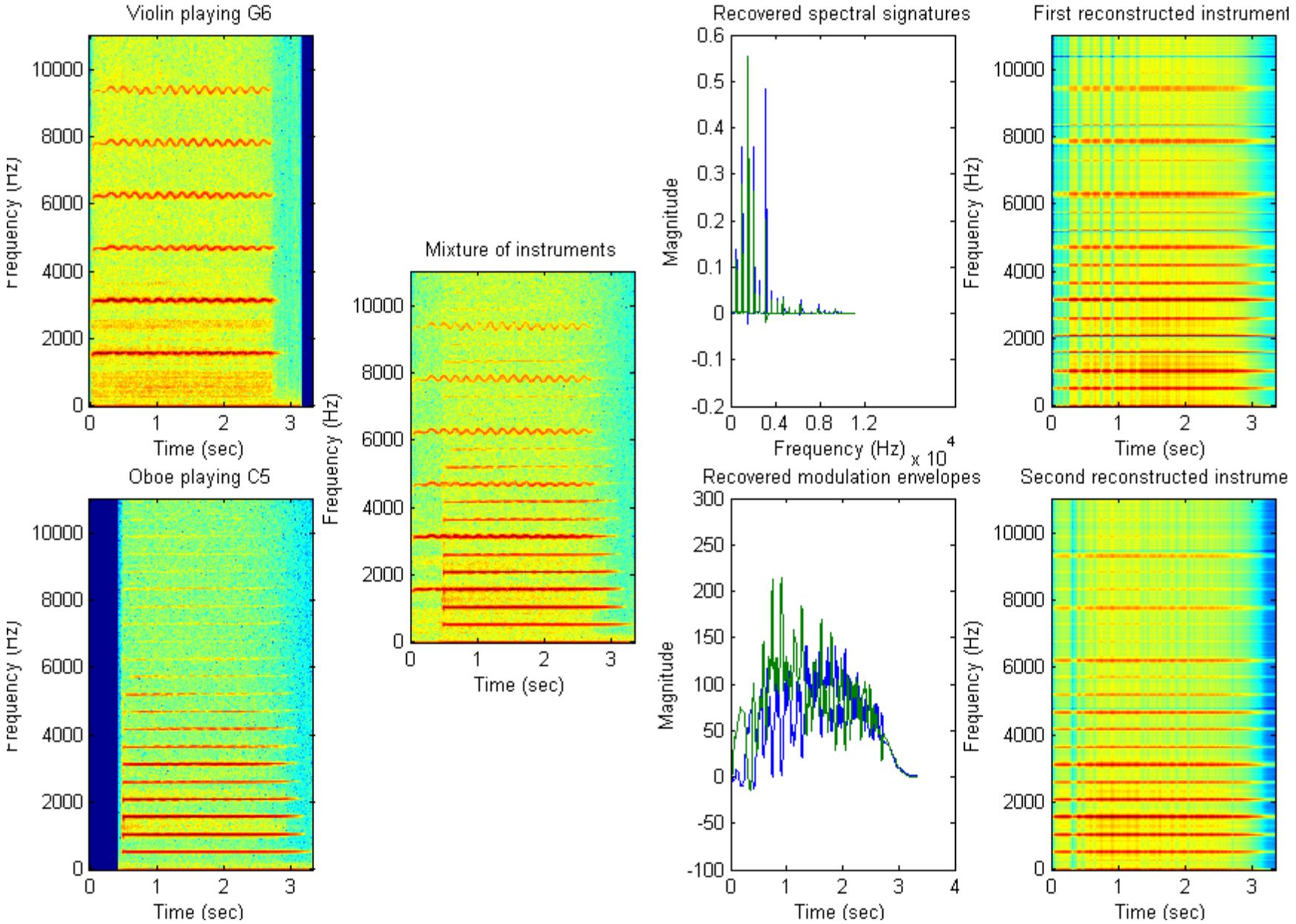
Violin playing G6
Frequency (Hz)
Time (sec)
Oboe playing C5
Frequency (Hz)
Time (sec)
Mixture of instruments
Frequency (Hz)
Time (sec)
Recovered spectral signatures
Magnitude
Frequency (Hz) x 10^4
Recovered modulation envelopes
Magnitude
Time (sec)
First reconstructed instrument
Frequency (Hz)
Time (sec)
Second reconstructed instrume
Frequency (Hz)
Time (sec)

Figure 24. (Previous page). First Column, Top: Original spectrogram of violin playing G6. First Column, Bottom: Original spectrogram of oboe playing C5. Second Column: Spectrogram of mixture of both instruments. Third Column, Top: Recovered spectral signatures of each instrument. Third column, bottom: Recovered modulation vectors of each instrument. Fourth Column, Top: Separated and reconstructed spectrogram of first instrument produced by taking outer product of recovered green spectral signature in top plot of third column with recovered green modulation vector in bottom plot of third column. Result bears some similarity to oboe in bottom plot of first column in terms of delayed onset time, but little separation is noted in spectrum, as discussed in text. Fourth Column, Bottom: Separated and reconstructed spectrogram of second instrument produced by taking outer product of recovered blue spectral signature in top plot of third column with recovered blue modulation vector in bottom plot of third column. Result bears some similarity to the violin in top plot of first column in terms of early onset time, but little separation is noted in spectrum, as discussed in text.

4.17 Results: Out of Phase Harmonic Sets

4.17.1 Description of Signals

To illustrate further the potential problems caused by an algorithm which ignores phase, we show what happens when we repeat the algorithm on the harmonic sets of Section 4.14, as before, but with one major change. We reverse the phase of all of the members of the first set of 5 harmonics.

4.17.2 Evaluation of Performance and Sources of Error

As can be seen from the plot of the summed signals in the second column from left, in those regions where there is overlap of harmonics of the two sets, there is complete cancellation of the signal. In effect, those harmonics have been turned off for that amount of time. The algorithm must then fit this unusual pattern in which some harmonics start at time $t = 0$, and end at $t = 2/3$ seconds, some start at $t = 1/3$ and continue until $t = 1$ second, and still others start at $t = 0$, continue until $t = 1/3$ seconds, and then restart at $t = 2/3$ and continue until $t = 1$ second. But we have requested of the algorithm to look for only two sets of modulation vectors and two sets of source signatures. The result will have to be some compromise. As can be seen in the plots in the 3rd column from left, the modulation vectors that were computed have nonuniform heights rather than the simple on or off step function shapes that they are supposed to have. This is because there is less energy at the points of overlap as compared to the ends. Similarly, the source signature vectors are of unequal heights. This too, is because the overlapping bands seem to have less energy overall than the other bands, since they appear to be off during the central portion of the signal duration.

The algorithm can only find the best fit, and is forced to give compromised results given the constraint of only two source signature and modulation vectors.

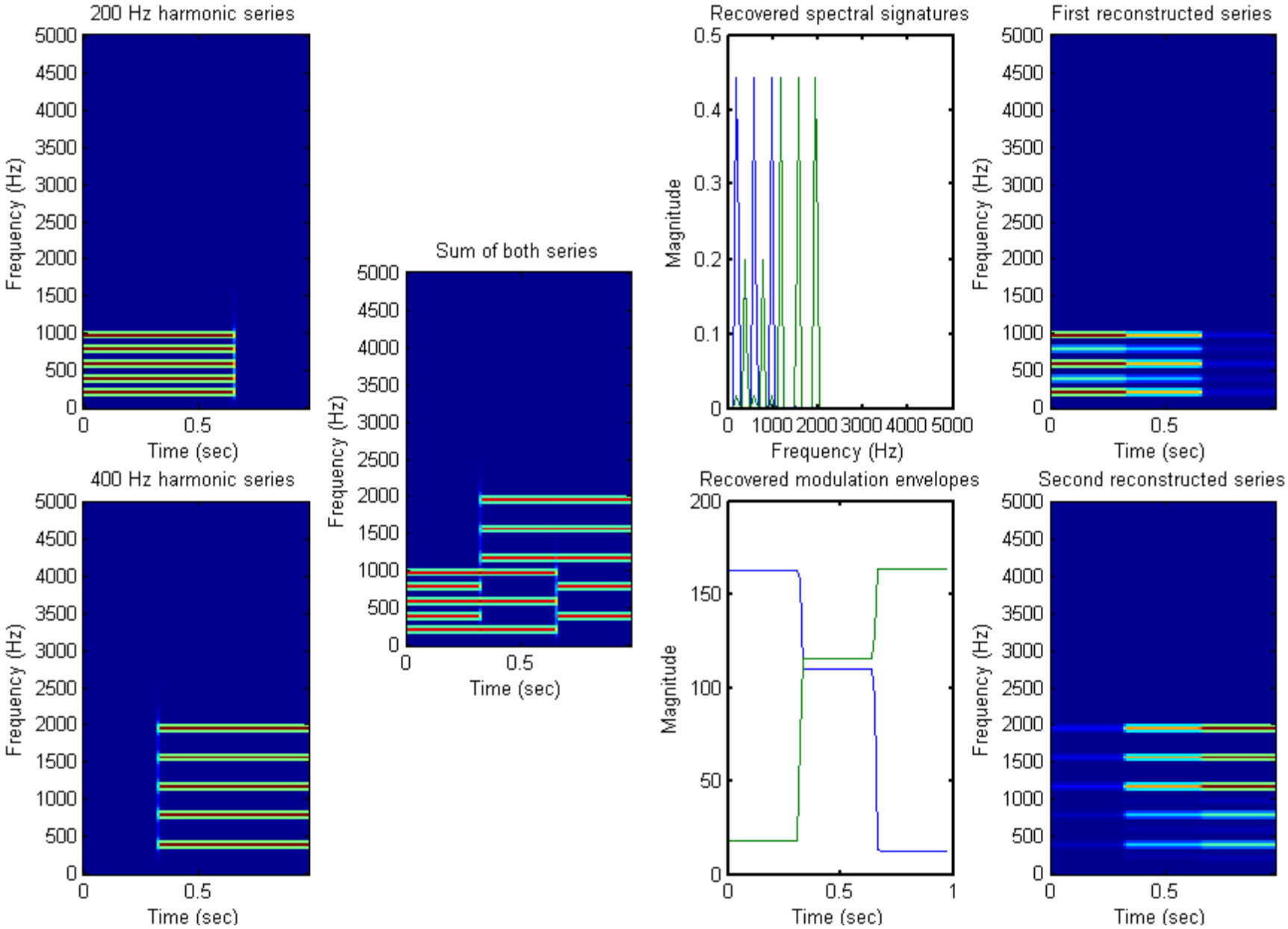
200 Hz harmonic series
Frequency (Hz)
Time (sec)
400 Hz harmonic series
Frequency (Hz)
Time (sec)
Sum of both series
Frequency (Hz)
Time (sec)
Recovered spectral signatures
Magnitude
Frequency (Hz)
Recovered modulation envelopes
Magnitude
Time (sec)
First reconstructed series
Frequency (Hz)
Time (sec)
Second reconstructed series
Frequency (Hz)
Time (sec)

Figure 25. (Previous page). First Column, Top: Original Spectrogram of 200 Hz 5-harmonic series, now with phase of pi for each harmonic. First Column, Bottom: Original Spectrogram of 400 Hz 5-harmonic series with phase of 0 for each harmonic . Second Column: Spectrogram of mixture of both series. Note cancellation in regions of overlap. Third Column, Top: Recovered spectral signatures of each series are not of 5 equal amplitudes due to cancellation in regions of overlapping harmonics. Third column, bottom: Recovered modulation vectors of each series do not show simple on/off step behavior due to regions of cancellation in time. Fourth Column, Top: Separated and reconstructed spectrogram of first 5-harmonic series produced by taking outer product of recovered blue spectral signature in top plot of third column with recovered blue modulation vector in bottom plot of third column. Fourth Column, Bottom: Separated and reconstructed spectrogram of second 5-harmonic series produced by taking outer product of recovered green spectral signature in top plot of third column with recovered green modulation vector in bottom plot of third column. Results show unequal amplitudes among harmonics of both series, and show tri-level behavior, rather than correct bi-level on/off behavior.

We further discuss the issue of phase and present an additional example in Section 4.19.

4.18 Relation to Auditory Scene Analysis

4.18.1 Onsets

A point underscored by the theorem is the importance of onsets in the separation process. As we have had occasion to mention previously, many current auditory scene analysis algorithms employ onset detectors in one way or another based on psychophysical evidence seeming to indicate that they are important factors. In Chapter 1, we mentioned studies of mistimed harmonics on number of sources perceived, as such an example. Perhaps we are now in a position to understand mathematically why this might be so. An onset in a particular band represents a point where there is a change from zero to nonzero in the modulation of that band. This serves as the most unambiguous marker for a separable source.

4.18.2 Comodulation Masking Release

From the requirements of the theorem, we also see the importance of a well demarcated spectral signature in the separation process. This may explain results found by (Grose and Hall III, 1996) who studied comodulation release from masking (CMR). The concept of comodulation masking release is that if a tone is buried under a region of noise whose frequency range overlaps the frequency of the tone, subjects will have trouble hearing the tone. However, if similarly modulated noise is added whose frequency range is outside the region of the tone (flanking bands of noise), the subject is better able to detect the tone. The explanation is most likely that the subject groups the similarly modulated noise regions together, and is therefore able to differentiate the unmodulated tone from the blanketing comodulated noise. In this particular

study, involving detection of a multicomponent signal presented against a background of multiple modulation patterns, the authors found that subjects did better when the independent modulation patterns were restricted to relatively discrete frequency regions. They concluded that the release from masking is more effective when the noise regions are narrow in bandwidth, as compared to when the noise regions are very wide in bandwidth. According to our theorem, having a frequency component in which a source is a soloist is a requirement for separation based on comodulation. If the noise regions are too wide, then there may not be such a frequency or range of frequencies.

4.18.3 Future Refinements

The application of these theoretical results to actual auditory scene analysis will require refinement before they can be implemented in a practical system. In the real world, sources change in frequency as well as amplitude, as for example in the case of speech. In addition, the issue of how to partition the frequency axis arises, since according to the theorem, the finer we partition the axis, the greater the chance of achieving uniqueness, as we require each source to have at least one frequency which is not overlapped by other sources. On the other hand, if the partition is made too fine, than we sacrifice time resolution necessary for tracking the modulation envelopes accurately. The filters then become so narrow that the rise time is too slow to respond quickly enough to amplitude changes. This is an example of the time/frequency tradeoff.

4.19 Phase

As we have seen in the last example presented (Section 4.17), our work tacitly assumed that the sources were coherent and in phase. In the event that the sources are out of phase, the situation becomes more complicated. As a further example, consider 2 sources of amplitude 1 volt which are 90° out of phase. If one source is modulated by an upward going ramp, and the other by a downward going ramp, the resultant envelope will begin at 1 volt, dip to 0.7071 volts, and then rise back to 1 volt, as shown in Figure 26. This should be compared with the similar Figure 21 where the sources were in phase. Unless one uses a complex-number representation, the addition appears nonlinear, and the algorithm cannot be applied to achieve separation since linearity is assumed throughout. On the other hand, if complex numbers are to be allowed, then

one is faced with the question of how to formulate a sufficient constraint to insure uniqueness, as nonnegativity is not satisfied when the domain is the entire set of complex numbers, since that includes negative numbers, as well.

While for real signals true coherence would be unlikely over a sustained period of time, however, one can alternately view the modulation vectors as representing squared amplitude quantities, corresponding to energy rather than amplitude of the signals in the mixture. For addition to work properly in this interpretation, we must assume that signals are incoherent with respect to each other. Then the energies of each component can be considered to add in a linear fashion. This can be understood by considering the definition of energy of a signal as

$$\begin{aligned} E &= \frac{1}{T}\int_0^T x^2(t)dt \\ &= \frac{1}{T}\int_0^T \left[s_1(t)+s_2(t)\right]^2 dt \\ &= \frac{1}{T}\int_0^T {s_1}^2(t)+{s_2}^2(t)+\cancel{2s_1(t)s_2(t)dt} \end{aligned} \tag{4.49}$$

where the cross-term integrates to 0 over a period provided that the signals are incoherent with respect to each other.

In summary, if one decides to treat all sources as incoherent, and to consider energy alone as the parameter of interest in order to remain in the domain of positive numbers only, then additions appear to be nonlinear whenever there is a coherent phase or frequency difference between overlapping spectral components, and in addition one can no longer faithfully reconstruct the waveform, as phase has been irreversibly destroyed. If, instead, one decides to use complex amplitude as the parameter of interest, then one needs to find some other way to insure uniqueness in the matrix factorization.

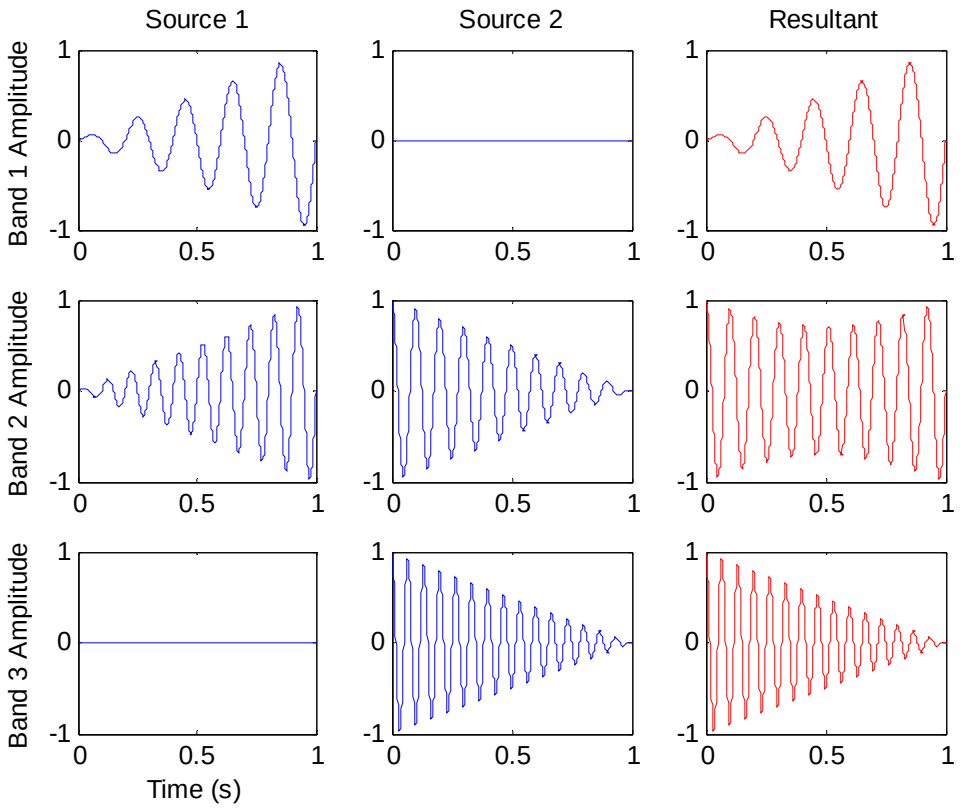

Figure 26. Example of 2 ramped sources which overlap in middle band of a 3 band system, as in Figure 24, but are now 90^0 out of phase. In this case, the resultant envelope of the middle band is no longer flat due to partial cancellation in the phasor sum.

Before leaving the topic, we point out that the seemingly simple phasor interaction in Figure 26 actually brings out some paradoxical questions about the concept of purely amplitude-modulated sources and their instantaneous frequency. We have explicitly assumed all sources to be constant in frequency, varying only in amplitude. However, Figure 26 seems to indicate that this is not the case for the sum. The fact that the resultant signal in the overlapping band (middle) begins in sine phase and gradually transitions to cosine phase, necessarily implies that its frequency, being the time derivative of its phase angle, must also be changing. The signal

appears to be a single source which is both amplitude and frequency-modulated, rather than a sum of two oppositely ramped quadrature sources. This problem relates to the broader conceptual difficulty in defining instantaneous frequency for multi-component signals. An analytic example is given in (Mandel, 1974), and further discussion is found in (Jones and Boashash, 1990).

4.20 Other Issues

4.20.1 Requirements for Real Time Operation

The algorithm we presented suffers from a deficiency in that it requires the entire data set to be present before it begins, which makes it inherently off-line. The reason for this is not simply computational. One of the conditions for uniqueness is that every source must be a soloist for at least one moment in time, even if that time is the last second of a one hour concert. But clearly the human ear is able to separate sounds with high probability even under circumstances in which the rigorous conditions of the theorem have not yet been or will not ever be completely fulfilled.

In addition, continuity in the time structure is not exploited in arriving at a solution. To see this, note that if we were to scramble the columns of $\mathbf{B}$ in some random order, the source signatures found would be exactly as before, but the columns of the modulation matrix $\mathbf{M}$ would be scrambled in the same manner as the columns of $\mathbf{B}$ were scrambled. This means that no advantage is taken of the time progression of real world sources, whose modulation is often continuous or piecewise continuous, in finding the correct solution.

4.20.2 Effect of Noise

Real world signals will always be corrupted to some extent by noise which will affect the rank of the matrix $\mathbf{B}$. However, methods for computing rank involving the singular value decomposition allow one to specify a threshold below which small deviations from ideality are ignored.

In addition, by its very nature of trying to find the best fitting set of source signatures and modulation vectors, the algorithm effectively averages out minor fluctuations due to noise, as we have seen. To some extent, this provides a certain immunity. As we have discussed in

Section 3.5, the redundancy of the multiple harmonics of many common auditory signal sources probably provides an additional advantage in characterizing and separating signals. This would not be possible if pure tones were the favored method of communicating.

4.20.3 Other Work on Non-Negative Matrix Factorization

At around the same time this work was presented by (Jacobson, Cauwenberghs and Litvak, 2001), independently (D. D. Lee and Seung, 1999) published a paper in Nature on the use of NNMF for learning the shapes of faces. They used their own algorithm for performing the factorization, and described a way to estimate the rank of the system. They compared results found using a number of other approaches (ICA, PCA) for obtaining a basis. They concluded that while a number of valid bases can be used to reconstruct faces from linear combinations of different images, the basis found by NNMF seemed to contain elements that most resembled discreet, identifiable parts of a face, like a nose, etc. In personal communication it appeared that Seung was unaware of the fact that the decomposition by NNMF is in general, nonunique. However, it could be that in a 2-D structure like a face, which has a large number of pixels, it is likely that there are regions that have nonzero values which do not appear in other faces. Hence it is possible that the comparable soloist conditions may have been fulfilled at any rate, or else that their algorithm found one solution, but may have also been able to find additional ones, as well. (Tropp, 2003) provides a nice overview of current work in the field of NNMF. It seems that most authors he cites do not have a clear understanding of the uniqueness issues, and how to reliably find the solution when it exists.

4.21 Summary

We have looked at the issues involved in separating constant-frequency amplitude-comodulated signals using a matrix factorization approach under the assumption that each source can be described by a single modulation vector which captures time variation, and a single source signature vector which captures spectral weighting. We further assumed that the envelopes of the sources add linearly. If these hold true, we proved that a mixture of such sources can be separated under two constraints: each source is a soloist for at least one moment in time, and each source has at least one frequency band which is not shared by any other source. This result makes use of the technique of Non-Negative Matrix Factorization, which has

recently been applied to various problems in biology and learning. Our work appears to provide a correct understanding of the uniqueness issues associated with this approach, which have previously eluded the research community, and we applied this knowledge to develop a number of reliable algorithms to obtain the solution when it exists.

We demonstrated with examples that the method works in constant-frequency coherent cases, since they meet the above assumptions, but will fail in frequency-varying or out-of-phase cases, as the assumptions are not met.

We note that extension of this method by allowing the modulation vector to become complex, thereby incorporating phase information is inviting, but unfortunately would require a completely new framework of analysis. Restricting the modulation vector to positive, real values along with the row and column constraints we have applied leads to a unique solution, as we have shown. However, relaxing the non-negativity requirement readmits an infinity of solutions as in Equation 4.3, and it is not at all clear if it is possible to limit the choices to a single, physically meaningful solution. We therefore broaden our search for more generally applicable methods in the coming chapters.

Chapter 5

Estimating Instantaneous Frequency in Mixtures of Sinusoids Using Multiple Filters

5.1 Introduction

In the previous chapter, for the constant-frequency, coherent case we developed a matrix formulation to describe a mixture of amplitude-comodulated sources. We explored uniqueness issues and found that if certain conditions are satisfied, then it is possible to separate the spectrogram of a mixed set of sources into separate spectrograms. A major deficiency with this approach is that phase information is not represented. This can lead to errors in separation by incorrectly accounting for interference when sources are out of phase or differ in frequency. Although waveforms add linearly, the envelopes do not. Furthermore, lack of phase information can cause difficulty in reconstructing the source waveforms. In addition to the phase issue, we also desire to handle the wider class of signals which are frequency-varying. We saw that the approach of Chapter 4 was unable to properly represent a violin note due to the applied vibrato.

In this chapter we develop techniques that can analyze waveforms with time-varying parameters, and extract instantaneous amplitude, frequency and phase of mixtures of multiple sinusoids of this type. These methods make use of slight differences in weighting that are applied to the spectral components of a mixture as it is passed through an overlapping filter bank with properties we will discuss. This differential weighting produces small differences in the waveforms of adjacent channels and corresponding small shifts in the positions of the local maxima. We demonstrate that given only the coordinates of the local maxima in the various

channels, it is possible to obtain the instantaneous frequency, amplitude and phase of all spectral components. We suggest that this might possibly explain the importance of phase-locking in the auditory system. We furthermore show that, in general, a mixture of as few as two sinusoids may have multiple images across channels depending on the characteristics of the filters, thereby confusing efforts at discovering the true number of sources based on comparison of the channel waveforms. To overcome this, we demonstrate methods of dimensionality reduction that yield the correct number of source components by efficiently consolidating this disparate data from large numbers of channels. These methods further simplify the problem by making it necessary to examine positions of the local maxima in only a small subset of the original channels, thereby greatly reducing the computational burden, and demonstrating that there is much redundant information in the various channels. We suggest that this redundancy can possibly be exploited to provide improved noise immunity.

5.2 Resolution Issues in Source Separation

Our previous work was restricted to coherent, constant-frequency AM sources. A more realistic class of signals is frequency-modulated as well as amplitude-modulated. In speech, the fundamental frequency (pitch of the voice), the frequencies of the formants (vocal-tract resonances), and the type of excitation (whether voiced, or produced via various kinds of turbulent airflows and bursts) are continually changing to convey phonetic information In music, instruments play many different notes over time, and even during the course of a single note, there is often deliberately introduced frequency variation in the form of vibrato to give the note a richer sound.

Whereas in Chapter 4 our approach was based on *a priori* applications of comodulation constraints, where the requirement of comodulation assisted in determining the parameters of the spectral components of the sources in the mixture; in the following discussion, we apply comodulation in an *a posteriori* manner. We first determine with as much accuracy as possible the parameters of all spectral components in the mixture, using methods we will describe. Then we may group those with common frequency and/or amplitude trajectories together. The final regrouping and resynthesis task is not attempted in this thesis.

The problem of determining the exact frequencies of closely spaced, time-varying signals presents a general problem due to the uncertainty principle, which mandates a tradeoff between time and frequency resolution. If one wants to compare amplitudes of spectral components to determine whether they are commonly modulated for purposes of source separation, one clearly needs accurate time resolution. In addition, one needs to know whether any instantaneous amplitude fluctuation within a particular channel is truly indicative of the amplitude variation of a single source, or whether it is an artifact caused by beating of multiple sources, and carries no information about any one source alone. We illustrated an occurrence of this type of ambiguity in Section 3.5.

Similarly, if one wants to compare frequency variations of particular spectral components for grouping purposes, one needs precise, high-resolution, instantaneous-frequency information on all spectral components. In general, however, conventional spectral estimation methods are limited by the general rule of thumb (Kay and Marple Jr, 1981) that the length of the recording needs to be equal to the reciprocal of the desired frequency resolution. (We provide a simple derivation in Section 6.3.) In other words, to separate signals whose frequencies differ by 0.1 Hz, one needs 10 seconds of recording time. However, that is clearly way too long to capture the fast amplitude and frequency variations that occur in spoken speech. Typical time scales for the movements of the articulators are on the order of 20 Hz, i.e., amplitudes can change in 50 milliseconds. Stop consonants can be shorter, still. Hence, much less time is available to capture a particular phoneme and separate it from competing speech. We also note that the usage of the term resolution often implies merely the ability to discern separate peaks in a spectrum. However, being able to determine the actual parameters of the signals, such as the correct amplitude of each, can often require even longer recording segments.

In analogy with the problem of the corruption of amplitude envelope information due to interference from multiple sources (beating), there is also corruption of instantaneous frequency information in the presence of multiple sources, as well. As we will discuss more fully in Chapter 6, and will graphically illustrate in Chapter 7, the normal definition of instantaneous frequency often gives nonsensical results in the case of mixtures.

For the foregoing reasons we seek to develop systems that improve on current techniques in their ability to provide both short-time and fine-grained resolution in the process of parameter

estimation. In the next section we look at actual mixtures of speech waveforms to graphically illustrate the necessity of these requirements.

5.3 Considerations in Filtering: The Speech Case

Whereas much of our previous discussion in Chapter 3, and the tests of algorithms in Chapter 4 were devoted to musical instruments, we now turn to speech waveforms. Musical instruments have more regularity than speech, since they play a single well-defined note (ignoring polyphony) at any given time. Even when there is applied vibrato, the deviation from the center frequency of the note is generally less than the interval between it and any other note. Because we wish to develop methods which are generally applicable to all types of audio signals, and because speech has less constraints on its structure, we will now examine some of the complex issues in the field of speech separation. A further motivation, as has been noted by (Halpin, 1997) is that when patients come to a clinic for assistance with hearing disorders, they are primarily concerned with being able to understand speech. Loss of ability to hear other sounds is never as bothersome or as detrimental to their quality of life as is difficulty in communicating with other human beings.

5.3.1 Spectrograms

We now examine samples of actual speech signals to illustrate the complexities of difficulties we are facing. The following are spectrograms of a male speaker, a female speaker and a mixture (summed waveform) of the two speakers. Both sentences are voiced. The male speaker is saying the sentence, *"Nanny may know my meaning."* The female speaker is saying the sentence, *"Why were you away a year, Roy?"* The sampling rate for both recordings was 10 KHz, and the spectrograms were computed using an FFT length of 512 with a Hamming window of 30 ms, and an overlap of 75%.

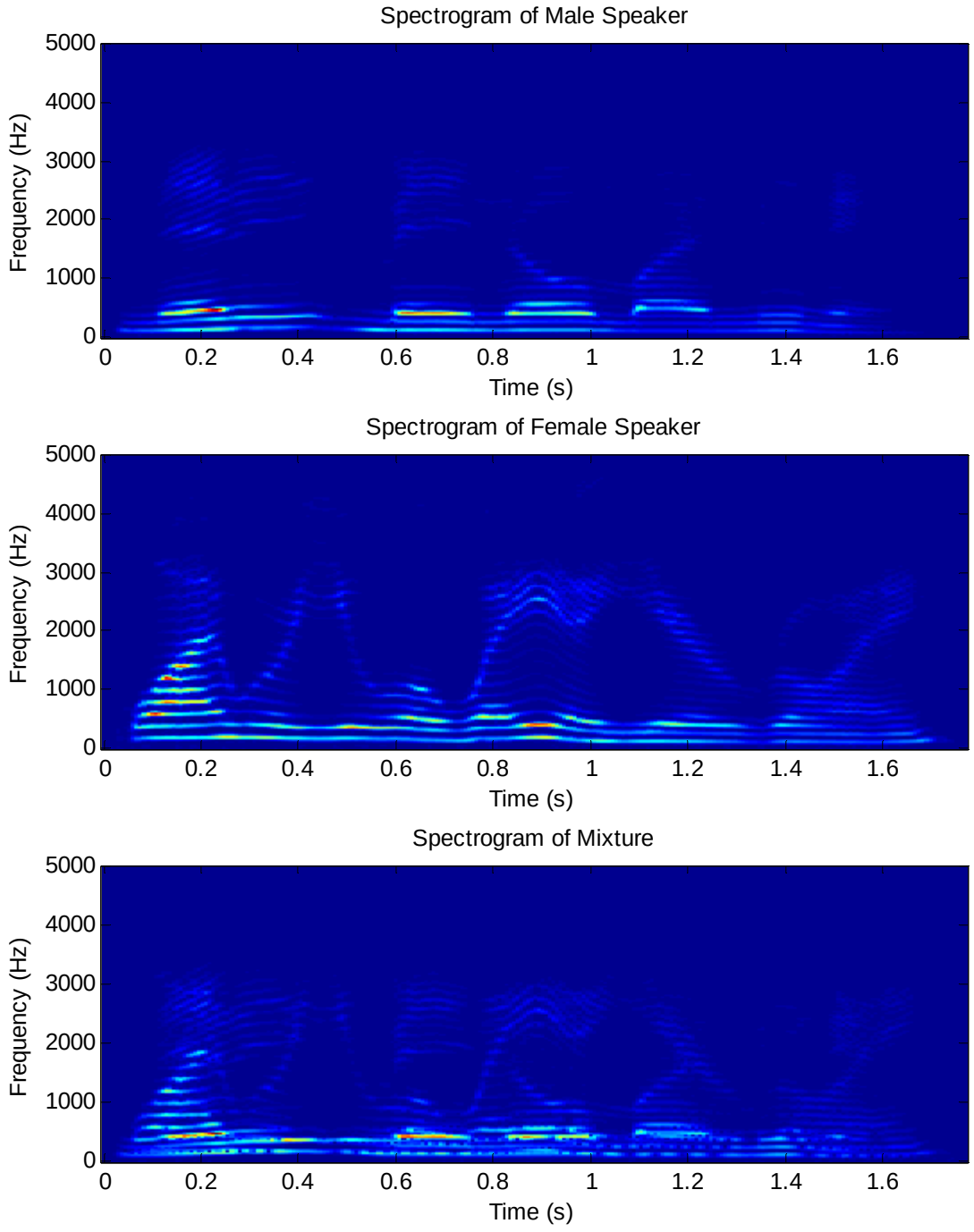

Figure 27. **Top: Spectrogram of male speaker saying, *"Nanny may know my meaning."***
Middle: Spectrogram of female speaker saying, *"Why were you away a year, Roy?"*
Bottom: Spectrogram of mixture.

We note the following features of these spectrograms:

1) The general structure of harmonic traces throughout. Without the effect of the vocal-tract filter, one would expect to see equally strong traces for all harmonics looking in the vertical direction at any given time (discounting the gradual roll off of the glottal source at high frequencies.) The effect of the vocal tract is to make some regions of harmonics stronger than others, thus, the appearance of regions at particular times in which certain harmonics seem to be missing or barely visible. The regions in which a group of harmonics are especially strong are those which are near the formant frequencies.

2) The phenomenon of frequency comodulation is quite noticeable. An especially prominent manifestation of this is at time 0.9 seconds in the female recording; the entire series of harmonics rises and falls in a coordinated manner.

3) The spacing between harmonic traces in the male recording is less than that in the female recording due to the lower fundamental frequency of the male's voice. Consequently, all multiples are at lower frequencies than in the female case. The frequency separation in either case, is of course, equal to the fundamental frequency.

4) In the spectrogram of the mixture, there is noticeable beating in the lower harmonics. This is due to interference between harmonics of the two speakers at regions in which they are close together in frequency. We will look more carefully at this shortly. The reason why this is not as prominently visible in the higher harmonics is probably due to an effective averaging of the shorter duration beating envelopes which are less than the window length used by the FFT calculation. They are shorter, because higher harmonics beat at multiples of the beat frequency of the fundamental, as we discussed in Chapter 3.

5.3.2 Waveforms

We next examine the actual channel waveforms by passing the recordings through bandpass filters located in certain regions of the spectrum so that only a single harmonic of each speaker is expected to be present to any significant extent. As we will discuss, the effect of beating will become more clearly noticeable. For comparison, we first show the complete, unfiltered

waveforms of the recordings in Figure 28, and then a shorter, unfiltered region of interest of each in Figure 29. We follow this with filtered versions of the signals.

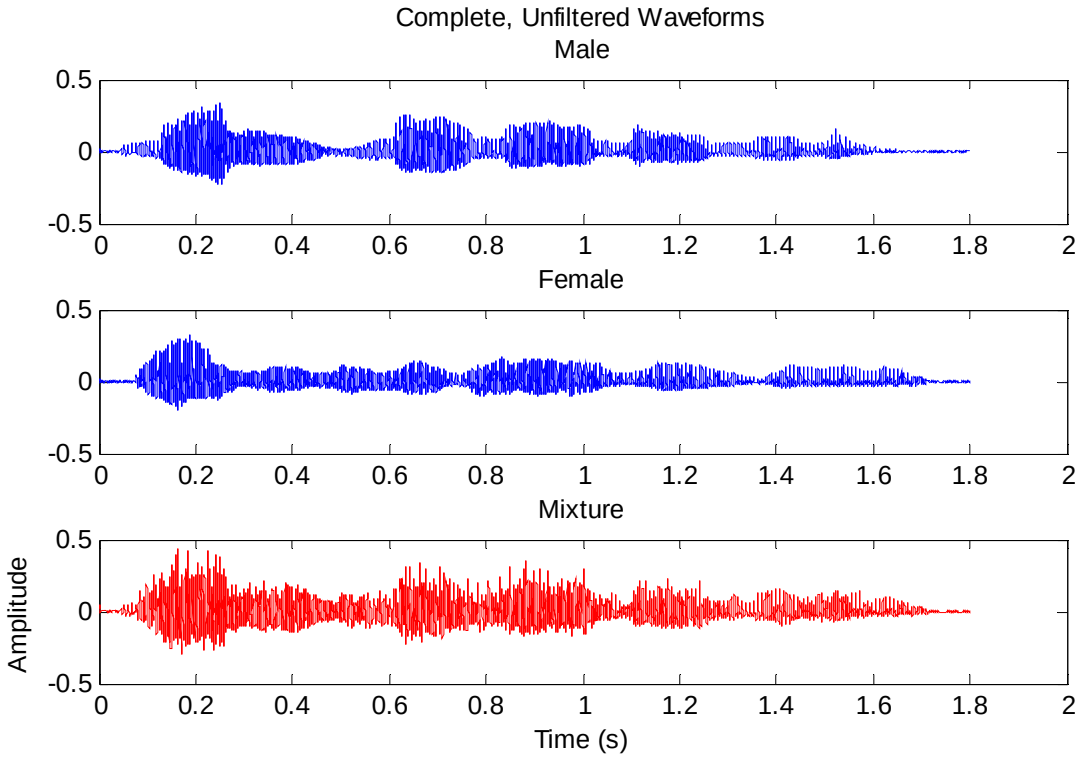

Figure 28. Plots of complete, unfiltered waveforms of the three recordings.

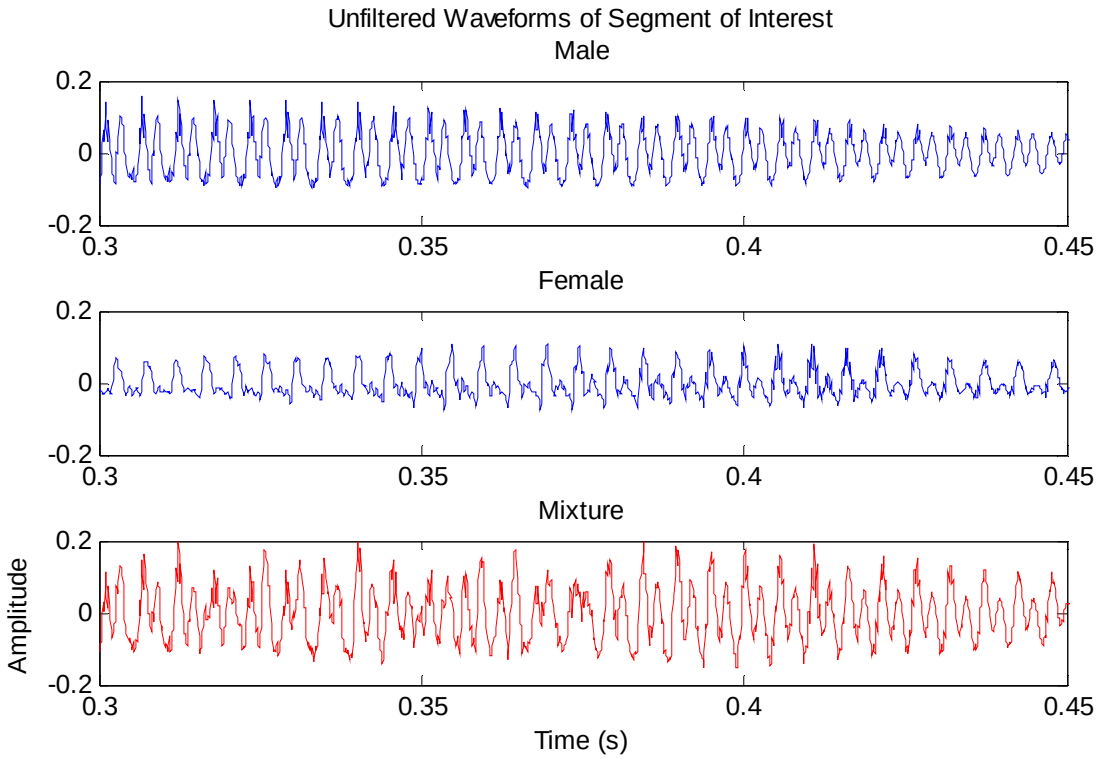

Figure 29. Typical unfiltered segment from the three recordings.

We note that in both the male and female waveforms, there is a fair amount of similarity between one cycle and the next, provided the two are not separated too far in time. We would expect this within a given vowel from our discussion of scaling in Chapter 3, although we concluded that perfect scaling requires the rather stringent and uncommon condition of phase comodulation. It is clear, however, that the male and female motifs are noticeably different, and distinguishable from each other anywhere within the duration shown. As the vowel changes, however, we expect the relative amplitudes of the harmonics to change, thus altering the composite waveshape. We also note that in the mixture waveform there is much less regularity, and that there is an appearance of roughness due to the inharmonicity of the combined signal. This is manifest in the bottom plots of Figure 28 and Figure 29.

5.3.3 AM vs. FM Characteristics Under Filtering

We next look at each of the above three waveforms after filtering by 100-Hz-wide 5th order Butterworth filters so as to pass only a single harmonic of each speaker. We show results using successively higher center frequencies (CF's) beginning with 150 Hz, at increments of 25 Hz in Figure 30, Figure 31 and Figure 32.

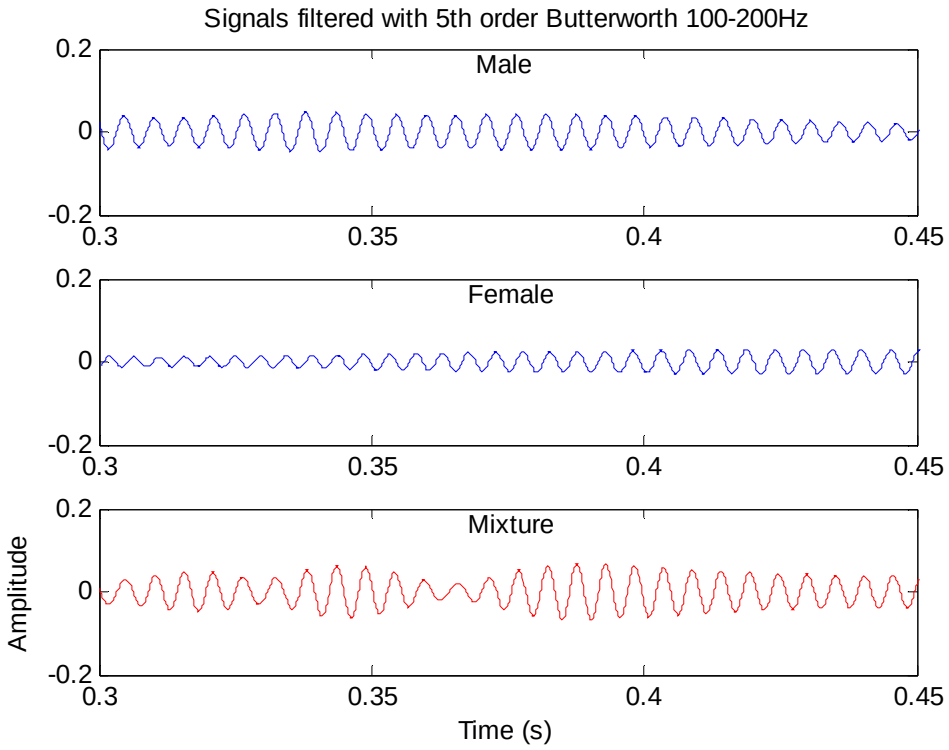

Figure 30. Waveforms of male, female and mixture at output of filter with passband 100-200 Hz.

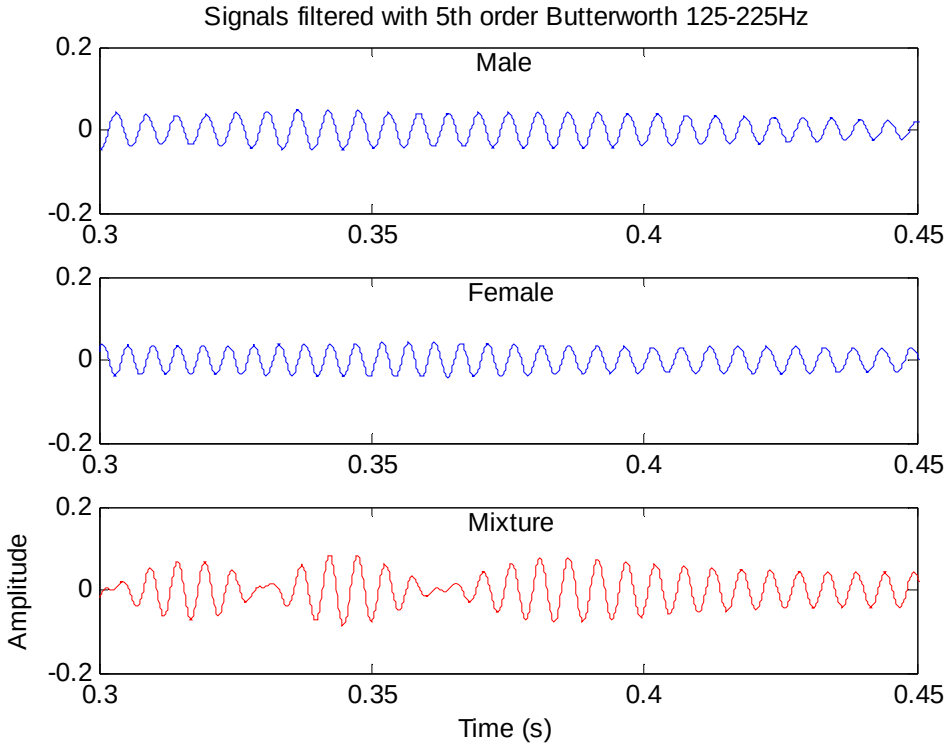

Figure 31. Waveforms of male, female and mixture at output of filter with passband 125-225 Hz.

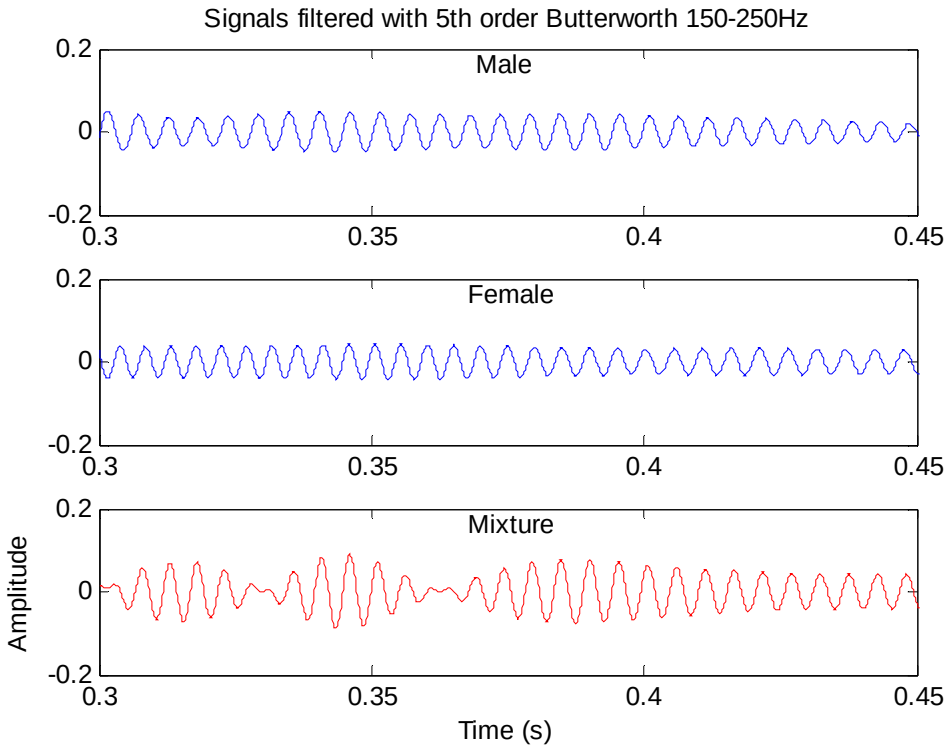

Figure 32. Waveforms of male, female and mixture at output of filter with passband 150-250 Hz.

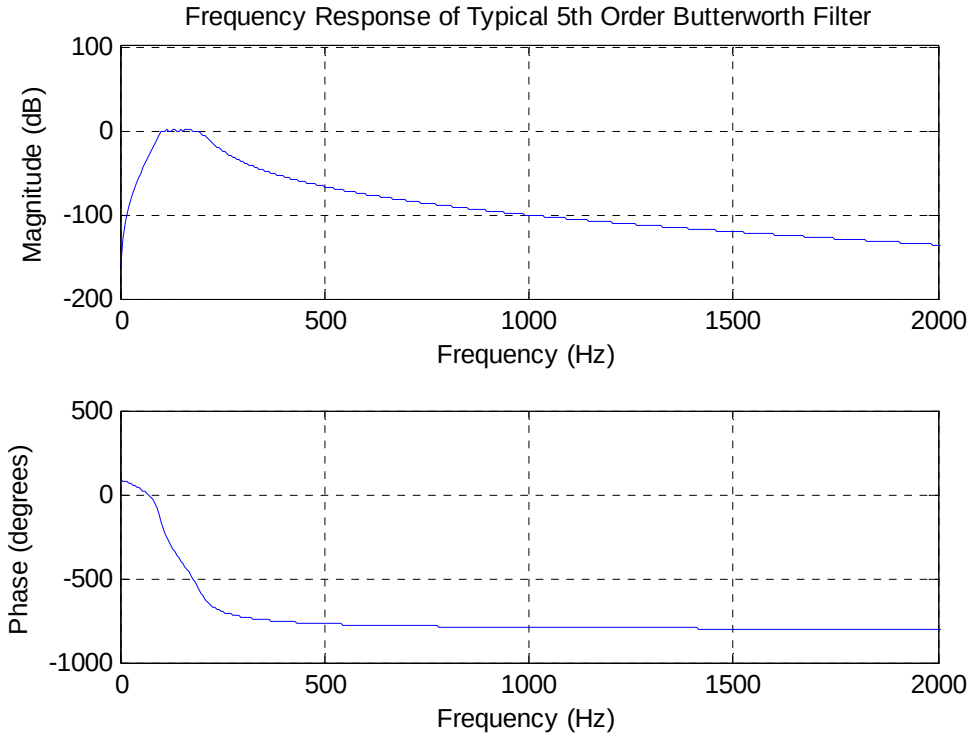

Figure 33. The characteristics of a typical 5th order Butterworth filter with passband from 100-200 Hz.

The characteristics of the first of these filters is shown in Figure 33. It has a nonlinear phase response, and displays a rather flat top with some ripple within the passband region, and has gradually sloping sides.

The frequency ranges in Figure 30, Figure 31, and Figure 32 are expected to be close to the fundamental frequencies of both speakers. More accurately, an FFT determines that the fundamental frequency of the female speaker is approximately 210 Hz in this segment, whereas that of the male is about 180 Hz. Therefore, in Figure 30, the female voice is slightly outside the usable bandwidth of the filter. For this reason, the amplitude in the middle plot of that figure shows a weak response, initially, which increases over time. This would appear to be indicative of an amplitude-modulated signal with increasing envelope. However, as an illustration of the complexity of the problems attendant in source separation, let us compare with the middle plot of the third filter in the series (Figure 32) in which the bandwidth is now 150-250 Hz. Here the female voice is actually seen to be decreasing in amplitude at the same instant. The explanation of this enigma is that, as can be seen by looking at the period of the waveform, the frequency of the female voice is decreasing. This brings it down to within range of the first filter (Figure 30). As the frequency moves away from the sidewalls towards the center frequency of the filter, the filter attenuates less and less. Therefore, in the first filter (Figure 30), the amplitude seems to be increasing. However, in the third filter (Figure 32), the frequency is fairly well centered within the passband so that changes in frequency do not shift the signal close enough to the sidewalls to be attenuated significantly. The same is true of the second filter (Figure 31), as well.[6]

The point of this example is that by looking at a single filter and a single harmonic of each speaker there are various ambiguities which are difficult to resolve. However, by using information from multiple filter channels we can get a clearer assessment of the situation. This will be our approach to this problem, and the theme of much of the remaining work in this thesis. We note that the problem of separating AM and FM contributions to a waveform has

[6] It is interesting that early investigators cited in (Hartmann, 1998) p. 454 including (Zwicker, 1956) and (Maiwald, 1967b; a)proposed that exactly such a method of converting FM to AM on the upper sidewall of an auditory filter and tracking the resulting amplitude fluctuations may be the mechanism actually used by the auditory system to detect FM. However, three pairs of later investigators (Hartmann and Hnath, 1982), (Demany and Semal, 1986) and (Moore and Sek, 1992; 1995) found that such a unified AM-FM detection model apparently is insufficient to account for all the data in experiments using mixed modulation and other stimuli. One pair of later investigators (Edwards and Viemeister, 1994) dissented on the basis that the detection statistic d' in mixed modulation tests equals the sum of the d' values obtained in AM and FM tests alone.

generated a large amount of literature in other contexts, as well (Quatieri, Hanna and O'Leary, 1997), (Torres and Quatieri, 1999), and we noted an additional example of such ambiguity in Section 4.18.

5.3.4 Constructive and Destructive Interference (Beating)

Another important feature which is evident in the plots we have shown is the beating that occurs in the summed waveform in the lower plot of each figure. We observed this in the spectrogram previously, but it is now more clearly visible. If we look at Figure 32, in both of the individual plots of the male and female waveforms we do not see any rapid amplitude fluctuations, the waveforms are smooth and well-behaved. However, in the summed waveform, there is prominent variation in amplitude which is due to constructive and destructive interference between the male and female signals, and seemingly not due to any behavior in either one alone that would account for it.

Note that the beating is most pronounced in the second and third filters of the series. The reason being that in these filters, the amplitude of the male and female voices are approximately equal to each other. In the first filter, the female signal is much lower than the male due to the fact that the frequency of the female signal lies very close to one of the sidewalls of the filter, as we discussed earlier. Beating is most prominent when the signals are equal in amplitude. When one signal is much greater than the other, the smaller signal has little effect on the larger by simple arithmetic. The most the smaller signal can affect the larger is to add an amount equal to its positive excursion and to subtract an amount equal to its negative excursion.

It is important to note that the relative amplitudes of the two components will only affect the amplitude or depth of the beating envelope, but not the beat frequency or phase. Any shift in the beating patterns between two different filter outputs is likely due to a phase shift that has been applied by one or both filters.

5.3.5 Comparison of Higher Harmonics

The following figures are plots from higher frequency filters whose bandwidth is expected to pass the second harmonic of the previous speech waveforms. Let us examine the behavior of the second harmonic of these signals.

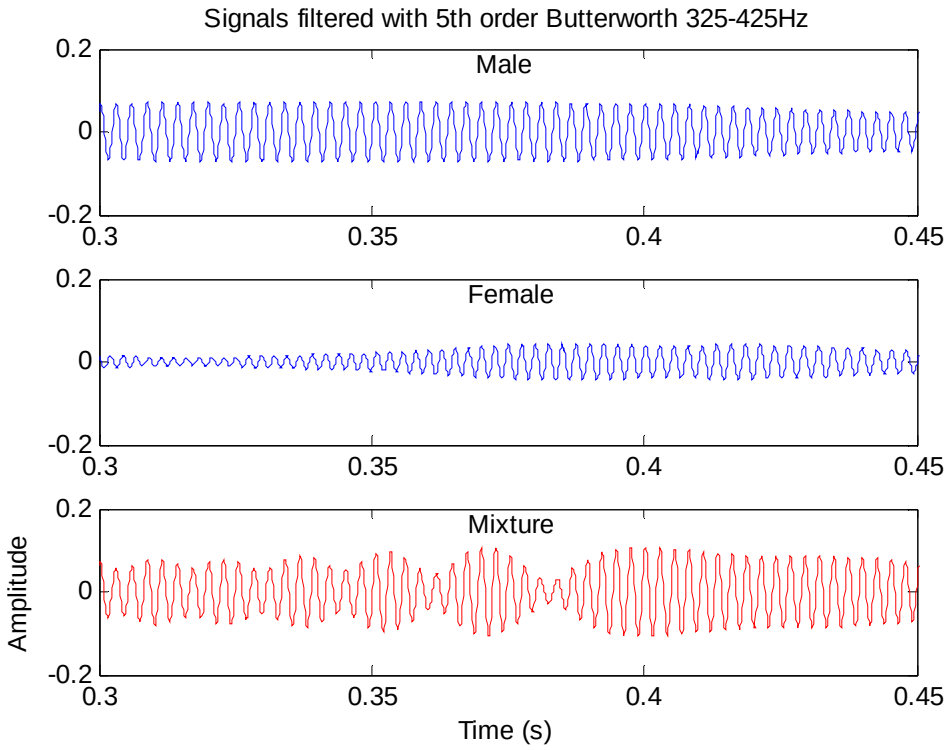

Figure 34. The speech waveforms passed through a 5th order Butterworth filter with bandwidth from 325-425 Hz. Filter is expected to pick up second harmonic of speech signals.

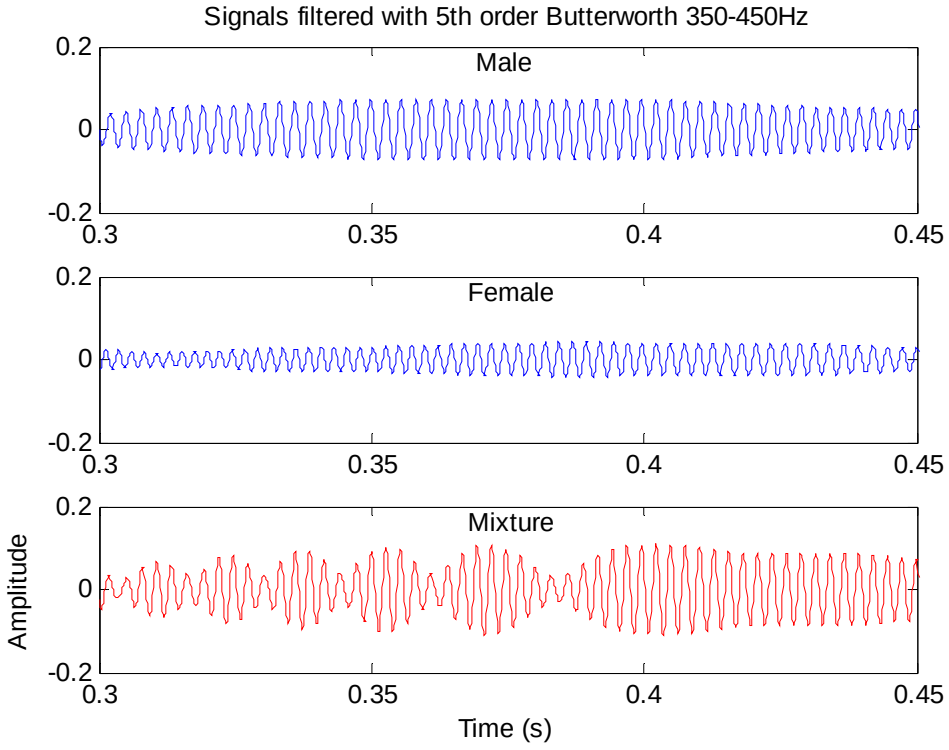

Figure 35. The speech waveforms passed through a 5th order Butterworth filter with bandwidth from 350-450 Hz. Filter is expected to pick up second harmonic of speech signals.

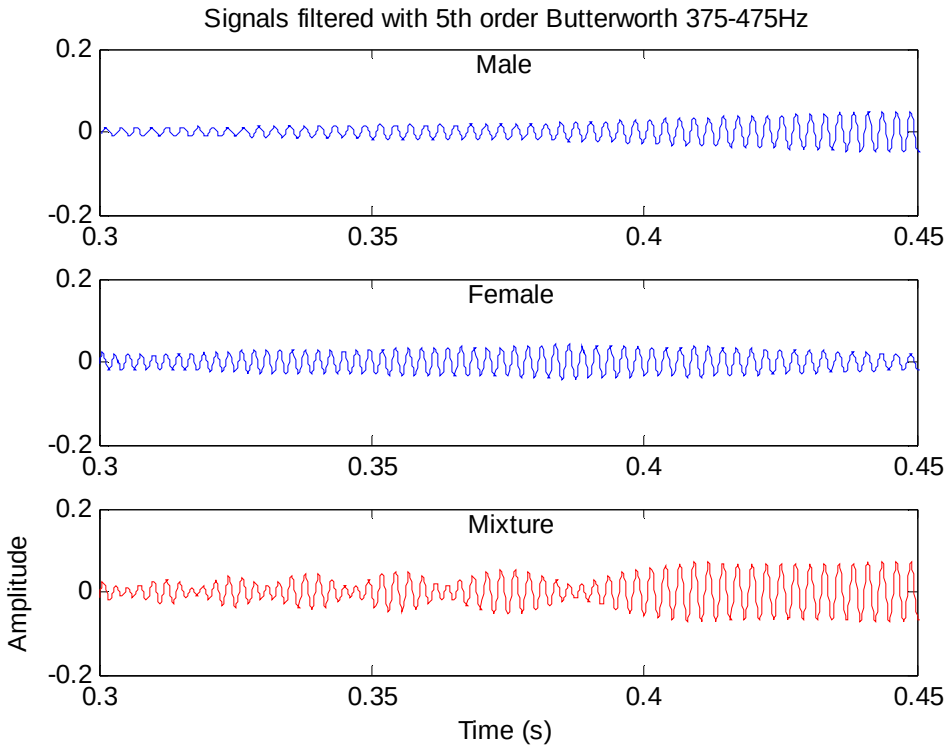

Figure 36. The speech waveforms passed through a 5th order Butterworth filter with bandwidth from 375-475 Hz. Filter is expected to pick up second harmonic of speech signals.

Our analysis in Chapter 3 concluded that if the envelope amplitude variation of the second harmonic of a comodulated signal is twice that of the fundamental frequency, it indicates a mixture, whereas if the envelope amplitude variation is the same as that of the first harmonic, a single signal is indicated. It is apparent by comparing Figure 35 with Figure 32 that we do indeed have a mixture. In Figure 35, the beat frequency of the second harmonics in the region from 0.3 seconds to 0.38 seconds is about 5 cycles/0.08 seconds or approximately 60 Hz, whereas in Figure 32 the beat frequency for the first harmonics in the same region is about 2.5 cycles/0.08 seconds or approximately 30 Hz. This of course corresponds with the difference between the FFT estimates of 180 Hz and 210 Hz for the fundamentals of the male and female speakers, respectively, as it must.

Although in Chapter 3 we established that speech cannot be considered to be truly amplitude-comodulated, nevertheless, it would be very difficult to generate some rhythmic modulation pattern using the muscles of speech which would vary the first harmonic at 30 Hz (quite rapid, considering the articulators move at about 20 Hz) and would simultaneously vary the second harmonic at 60 Hz. Aside from the impracticality, it would be nearly impossible from an

acoustic perspective, since the formants are generally wide enough so that they enhance a number of harmonics together. It would be quite improbable to manipulate the vocal tract into some configuration in which the first and second harmonics could be modulated independently of each other. Furthermore, we concluded that at the glottal-source level, in all likelihood the harmonics are fairly well amplitude-comodulated, leaving mainly the vocal tract to differentiate the amplitudes. We therefore conclude that this unusual behavior could only arise from a mixture of sources, and not from a single source. This is a useful observation, as even the task of determining the number of speakers alone, without separating, is known to be difficult. (Quatieri, 2004), (D. L. Wang, 2005).

5.3.6 Summary

Speech waveforms are continually changing in amplitude and frequency. The use of multiple harmonics may give greater insight into the underlying composition of a signal and in determining its origin, whether from one source or from a mixture. Tracking the trajectories of individual harmonics is difficult because of the coloring of the filter. The behavior of mixtures is more complicated, still, due to interference effects. While we have found that interference patterns can shed some light on the nature of a signal, we seek more precision, since any method based on the general envelope shape will take quite a few cycles until the shape of the envelope becomes sufficiently clear.

5.4 Spectral Estimation Based on Local Maxima

In the next few sections, we discuss three algorithms, all of which are based on the idea of comparing the positions of local maxima in differently weighted mixtures of sinusoids to improve resolution. After motivating the progressive stages of development, we will distinguish between three increasingly more realistic models for each. The first is where the weighting of the respective versions of the signals are known in advance. The second is where the weighting is unknown, but is based on mathematically-ideal multiplicative factors that simulate the steady-state response of a filter bank. The third is where the weighting is produced by convolution with actual filters.

5.4.1 Temporal Processing for Frequency Resolution

To overcome some of the limitations of conventional Fourier methods, we have developed parametric methods for frequency estimation utilizing the local maxima in the signals of each channel. In contrast to conventional methods which treat each channel as independent, these methods seek to pool information among channels in order to get more precise and better localized results.

It has been pointed out by (Cariani, 2005) and others that there is a paradox in understanding the frequency resolution of the auditory system. On one hand, neural data seems to show that the higher the amplitude of a tone, the wider the response area of a nerve fiber to the tone. In other words, the frequency response broadens and the fiber responds to more frequencies than it does at lower amplitudes. The implication is that an individual fiber becomes less selective, and that more fibers of other frequencies will respond, as well. This should seem to lead to poorer resolution or discrimination at higher amplitudes. But psychophysical data show just the opposite; people do better on frequency selective tasks when the stimulus is louder. Simple spectral-filtering models of the function of the cochlea do not seem to be able to explain this. Cariani therefore believes that temporal analysis plays an important role in understanding the function of the cochlea. He proposes using autocorrelation calculations in multiple channels to explain the frequency-resolving abilities of the auditory system.

While we agree that temporal processing can provide advantages in parameter estimation, our methods try to avoid the multiplicative operations inherent in computing autocorrelations. We believe that the use of long sequences of past data points negatively impacts temporal resolution, and note that this is a drawback of Fourier methods, as well, which rely on the multiplication of a finite-length sequence of data points by a set of complex exponentials. This inclusion of past history limits the ability of any method to respond to instantaneous changes. We therefore attempt to avoid this, and suggest an alternate approach.

5.4.2 Estimation of Parameters

Our approach is based on the elementary fact that the parameters of a pure sinusoid can be estimated from two successive local maxima in its waveform, as follows:

1) Amplitude is the height of the peaks.

2) Frequency is the reciprocal of the period, which is the elapsed time between two successive peaks.

3) Phase is the offset of the first peak from the time origin as a fraction of the period.

The above is a true characterization of instantaneous amplitude, frequency and phase for monocomponent signals only. For such signals, the height of all peaks will be the same. Our approach will be to assume that the signals in a given channel are initially mixtures, or multicomponent, but in the course of multiple iterations of the algorithms, will become more monocomponent in character. Since we will initially be dealing with complex waveforms, we will average the heights of two successive peaks to determine amplitude. Similarly, the presence of other components distorts frequency estimation as well, and can cause estimates to come out lower or higher than the true frequency. We elaborate on this in Chapter 7.

5.4.3 Behavior of Local Maxima in Mixtures

The intuition on which the algorithm rests is that if one has a mixture of more than one sine wave, in general, the resultant peaks will fall somewhere in between the locations in which they would otherwise fall for each sinusoid alone (a compromise). If one sinusoid is weighted more heavily than the other, then the peak locations will move closer to those of the dominant component. When considering multicomponent signals passing through a filter bank, the peak locations in each channel will reflect the weighting of that particular filter to the frequencies of the sinusoids present. For example, in the case of a mixture of a lower-frequency sinusoid and a higher-frequency sinusoid, in lower-frequency channels, the peaks will lie closer to the locations in which they would lie for the lower-frequency signal alone, while in higher-frequency channels, they will lie closer to the locations in which they would lie for the higher-frequency

signal alone.[7] However, as we will now show, the peaks are actually biased toward the higher-frequency component, all other things being equal.

If we have

$$x = a_1 \sin(\omega_1 t + \phi_1) + a_2 \sin(\omega_2 t + \phi_2) \tag{5.1}$$

then

$$\frac{dx}{dt} = \omega_1 a_1 \cos(\omega_1 t + \phi_1) + \omega_2 a_2 \cos(\omega_2 t + \phi_2) \tag{5.2}$$

At local maxima, the derivative $\frac{dx}{dt}$ will equal zero. This will occur when

$$\omega_1 a_1 \cos(\omega_1 t + \phi_1) = -\omega_2 a_2 \cos(\omega_2 t + \phi_2) \tag{5.3}$$

Let us therefore plot both sides of the equation and examine points of intersection which are the solutions of the system.

[7] It is actually not true that the presence of a lower frequency signal in the mixture can only pull the estimate lower, and the presence of a high frequency can only pull the estimate higher. Rather, the reverse can happen, as well, for reasons we describe in Chapter 7 on analytic solutions. For now, we ignore this, as it doesn't affect operation of algorithms.

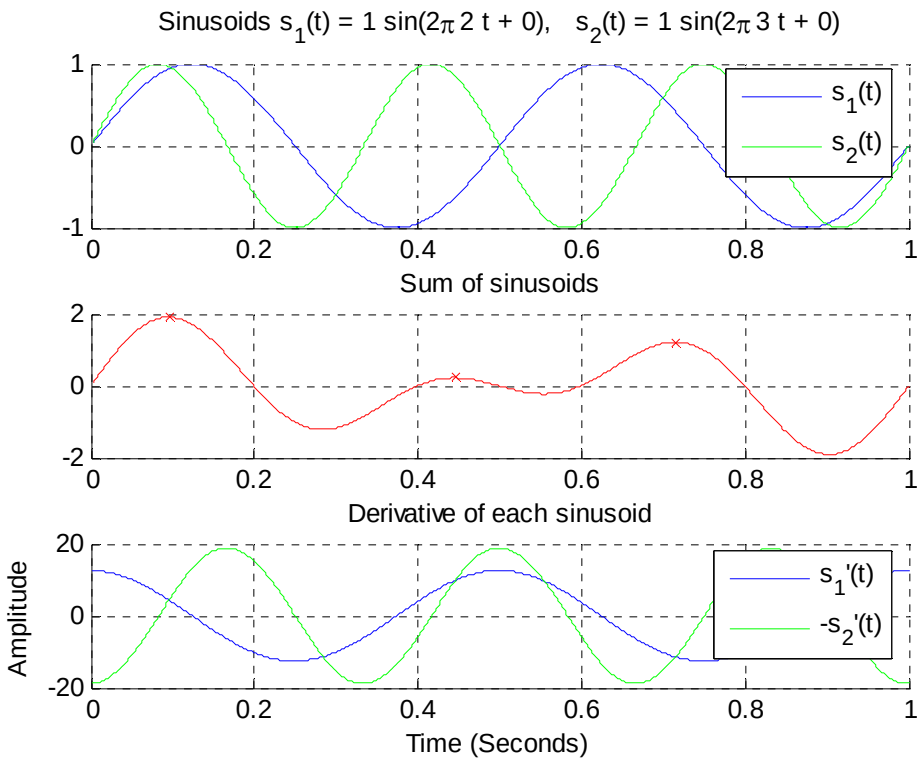

Figure 37. Behavior of local maxima of a pair of sinusoids with a1=1, f1=2, p1=0 and a2=1, f2=3, p2=0. Upper plot shows the two sinusoids. Middle plot shows the sum, with an 'x' marking the local maxima. Lower plot shows the derivative of each sinusoid, with the higher frequency derivative flipped 180 degrees. The local maxima of the sum correspond to points in time at which the two derivative curves intersect each other. Number of peaks in sum corresponds to number of peaks in higher-frequency signal.

In Figure 37, the first plot in this series, we have used two sinusoids of frequencies 2 and 3 Hz, respectively. Because the derivative terms in the equations are weighted by a frequency factor, the derivative of the higher-frequency sine will be of greater amplitude than that of the lower-frequency sine. Therefore, the derivative of the higher-frequency sine will cross the derivative of the lower-frequency sine twice per every higher-frequency cycle, as can be seen in the lower plot. One of these crossings will produce a local maximum, and the other a local minimum, as can be seen by comparing the middle and lower plots. The end result is that the peaks will, in a sense, be dominated by the higher frequency waveform. Note also that the location of the peaks in the resultant falls closer to the location of the peaks in the higher-frequency waveform than to the location of the peaks in the lower-frequency waveform, as can be seen by comparing the upper and middle plots.

If we increase the amplitude of the lower frequency plot to 1.5 to compensate for the frequency ratio of 3:2, the situation then looks like that of Figure 38.

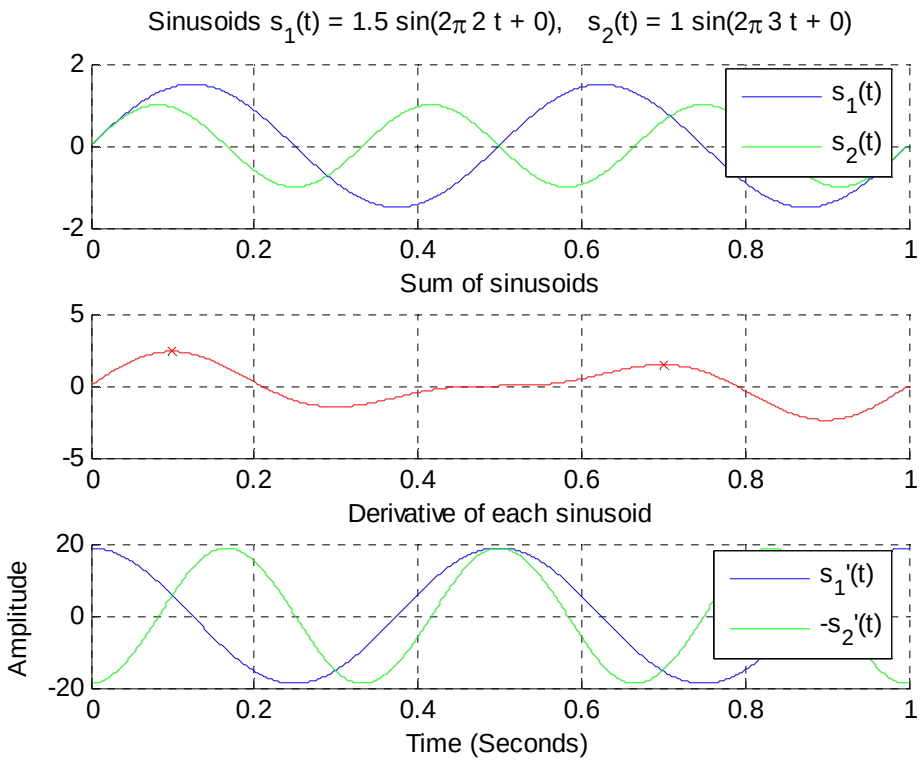

Figure 38. Sinusoids of Figure 37 with a1 increased to 1.5. Derivatives are now of equal amplitude. Number of peaks in sum now corresponds to number of peaks in lower frequency signal.

The derivatives are now of equal amplitude. Note that whereas in the previous plot, the derivative of the higher-frequency signal intersected the derivative of the lower twice per higher-frequency cycle, now, at time 0.5 seconds, the derivatives intersect only once in that cycle, thus producing a point of inflection, rather than a local maximum or a local minimum, as can be seen by comparing the middle and lower plots. The end result is that there are now only two local maxima, corresponding to the number of maxima in the lower frequency signal, rather than the number in the higher frequency signal, as before.

If we increase the amplitude of the lower frequency signal still further, to 2 volts, overcompensating for the 3:2 frequency ratio, then the situation is as in Figure 39.

In this case, as can be seen in the lower plot, the amplitude of the derivative of the lower-frequency signal exceeds that of the higher frequency signal. At time 0.5 seconds, the lower-frequency derivative is "out of reach" of the higher-frequency derivative, and is not intersected by it. Therefore, here again, there is no peak at that location, and the number of peaks in the resultant corresponds to the number in the lower-frequency signal, rather than to the number in the higher-frequency signal, as in the first case.

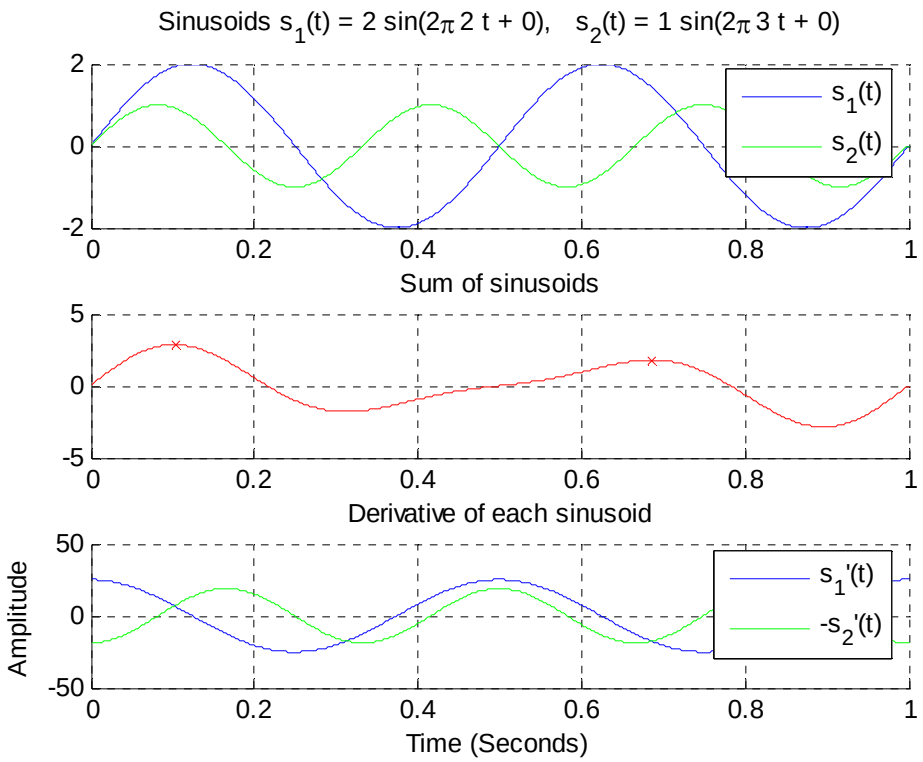

Figure 39. Sinusoids of Figure 37 with a1 increased to 2. Amplitude of derivative of lower frequency signal now exceeds that of higher frequency signal. Number of peaks in sum corresponds to number of peaks in lower frequency signal.

Finally, in Figure 40 we show an example where the frequency ratio is now 5:1, rather than 3:2, and where we keep the amplitude of the lower-frequency signal the same as in the last example, at 2 volts, while continuing to hold the amplitude of the higher-frequency signal at 1 volt. The peaks of the higher-frequency signal dominate, since the amplitude ratio is less than the reciprocal of the frequency ratio. The effect is similar to a ripple in which a high-frequency signal rides on a low-frequency signal, with the peaks corresponding almost entirely to the peaks of the higher-frequency signal. As in the first case, the derivative of the higher-frequency signal intersects that of the lower-frequency signal twice per higher-frequency cycle.

To sum up, the locations of peaks within a mixture of two signals are a compromise between the locations they appear in either signal alone. Both the relative amplitudes and frequencies of the signals influence this positioning. Our goal is to unravel this information to recover the individual parameters.

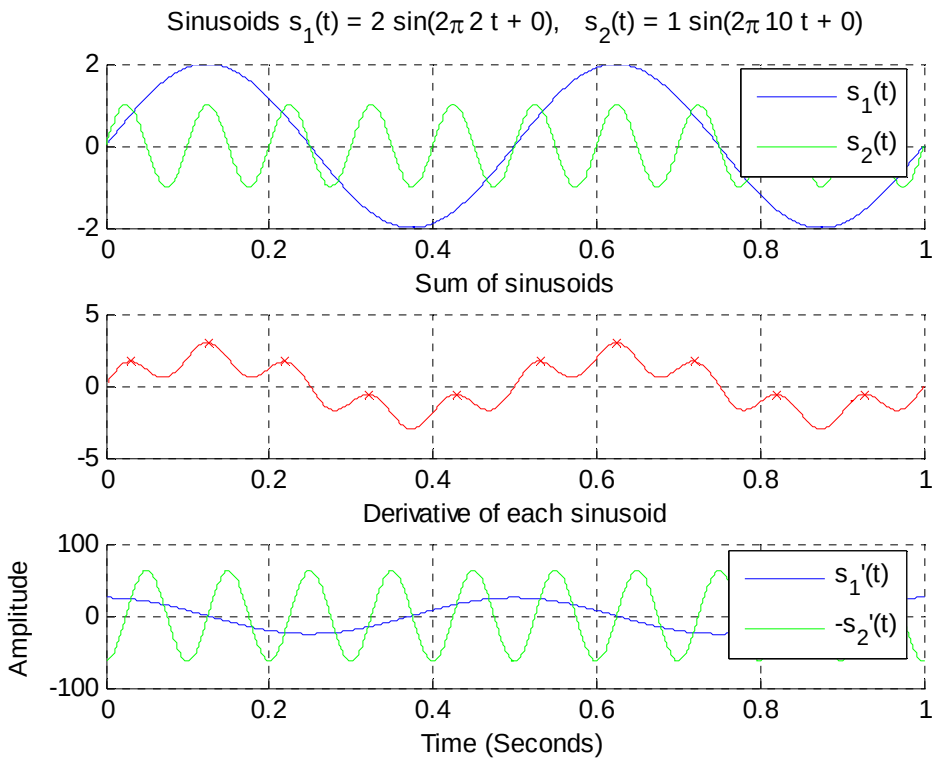

Figure 40. The situation of Figure 37 when f2>>f1 (f2 is 5x greater than f1) while the amplitude of a1 is undercompensated (a1 is only 2x greater than a2). Number of peaks in sum now corresponds to number of peaks in higher frequency signal.

We will now present a number of algorithms all of which are based on using local maxima from multiple filters in an attempt to improve time and frequency resolution over what can be obtained with conventional methods.

5.5 Iterative-Subtraction Algorithm

We first look at a single channel attempt at sinusoidal separation using local maxima in Section 5.5.1. We obtain an interesting result that although the parameters recovered can faithfully reproduce the input signal for the duration of interest, they are not accurate if the frequencies of the mixture are too close, for reasons we explain. Modification of the algorithm to use multiple channels produces more accurate results in 5.5.2.

5.5.1 Single Channel—Introduction and Motivation

Our first attempt to estimate parameters of a mixture via coordinates of local maxima is based on the discussion in 5.4.3 that if one has two sines which are additively combined, the resultant will have peaks that are expected to be asymmetrically weighted toward the higher frequency

component, assuming the amplitudes are comparable. Because of this asymmetry, if one computes a frequency estimate based on peaks of the resultant, one might expect this estimate to be closer to the frequency of the higher-frequency component. One could then construct a signal having the amplitude, frequency and phase of this first estimate, and subtract it from the original sum. One would then expect to have left over a residual which is closer to the lower-frequency component, since an initial estimate of the higher-frequency component has been removed. One would then form an estimate of this lower-frequency residual, again using the peaks of the waveform, and subtract from the original mixture, expecting the residual of this operation to be a better estimate of the higher-frequency component than the first estimate. Iterating, one might expect to get more and more accurate estimates of both components in an alternating manner. Note that no filters have been used in this algorithm. We operate directly on the signal in question, which in this early stage of our work, consists of only two sines.

5.5.2 Algorithm

The following steps were implemented:

1) Given $x = a_1 \sin(2\pi f_1 t + p_1) + a_2 \sin(2\pi f_2 t + p_2)$.

2) Estimate $\hat{a}_1, \hat{f}_1, \hat{p}_1$ using peaks of x.

3) Construct $\hat{s}_1 = \hat{a}_1 \sin(2\pi \hat{f}_1 t + \hat{p}_1)$.

4) Compute residual $R_2 = x - \hat{s}_1$.

5) Estimate $\hat{a}_2, \hat{f}_2, \hat{p}_2$ of R_2 using peaks.

6) Construct $\hat{s}_2 = \hat{a}_2 \sin(2\pi \hat{f}_2 t + \hat{p}_2)$.

7) Compute residual $R_1 = x - \hat{s}_2$

8) Estimate $\hat{a}_1, \hat{f}_1, \hat{p}_1$ of R_1 using peaks.

9) Repeat from step 3 until convergence.

Figure 41 shows this schematically. Note that in the diagram Estimate *1* refers to $\hat{s}_1$ and Estimate *2* refers to $\hat{s}_2$. Initially, both estimates are zero. The input is applied directly to the Amplitude, Frequency and Phase (AFP) Estimator block on the top which uses a desired pair of successive peaks to form estimates of these parameters, as described above in Section 5.4.2. Using these estimates, the Oscillator block generates a sine wave with these parameters. This yields Estimate *1*. This estimate is then subtracted from the original mixture (Input) and the residual is then fed into the AFP Estimator and Oscillator blocks on the bottom. This yields a new sine wave which is Estimate *2*. This estimate is then subtracted from the original mixture (Input) and the residual is then fed into the AFP Estimator and Oscillator blocks on the top, and the whole process is repeated until convergence.

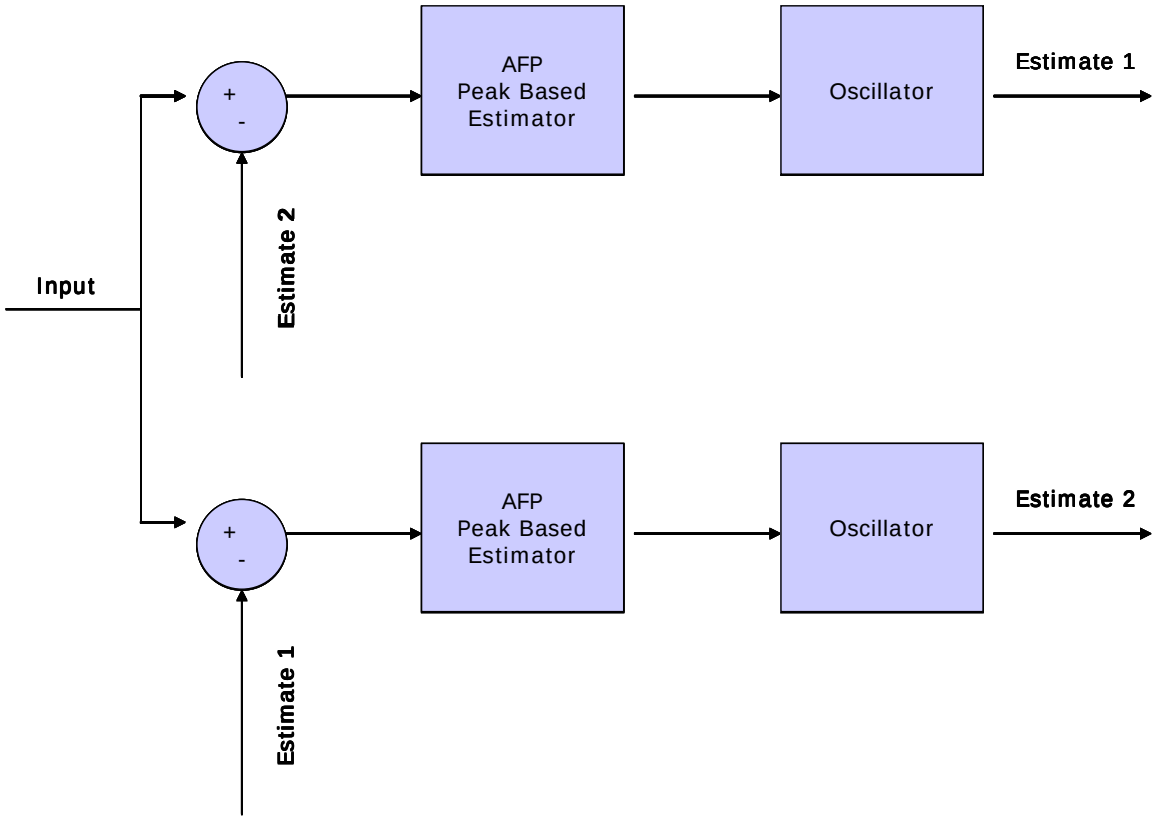

Figure 41. An attempt at sine-wave separation using iterations on a single input.

5.5.3 Results

In practice, the farther apart the frequencies, the better the estimates. If the frequencies are too close, the method fails. We illustrate an example of each case. The reason for the difference in performance is probably due to a manifestation of the uncertainty principle. It is possible for

two different sets of sines to combine so that the resultants are indistinguishable from each other over the shorter term, but diverge in the longer term. This implies that to get accurate frequency estimates, one must observe for a longer time. In other words, the uncertainty principle can be thought of as a statement of nonuniqueness. More than one combination of sines can fit the waveform under study over the short term. How long is considered to be short term depends on the frequency difference between the sines. The greater the frequency difference, the shorter the duration over which the waveforms can match.

Table 3 illustrates the original and recovered parameters for a pair of sines separated by 0.5 Hz. Phase is given as a percentage of π, the maximum possible deviation.

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
6.00	5.76	4.00	1.00	0.61	39.0	0.0	+0.12	4.0
6.50	6.45	0.80	1.00	1.39	39.0	0.0	-0.05	1.5

Table 3. Comparison of original and recovered parameters for a mixture of sines using a single input.

The algorithm correctly finds parameters that fit the waveform, but these differ from those used to generate the waveform. In a sense, the algorithm performs properly, but is limited by the uncertainty principle discussed earlier.

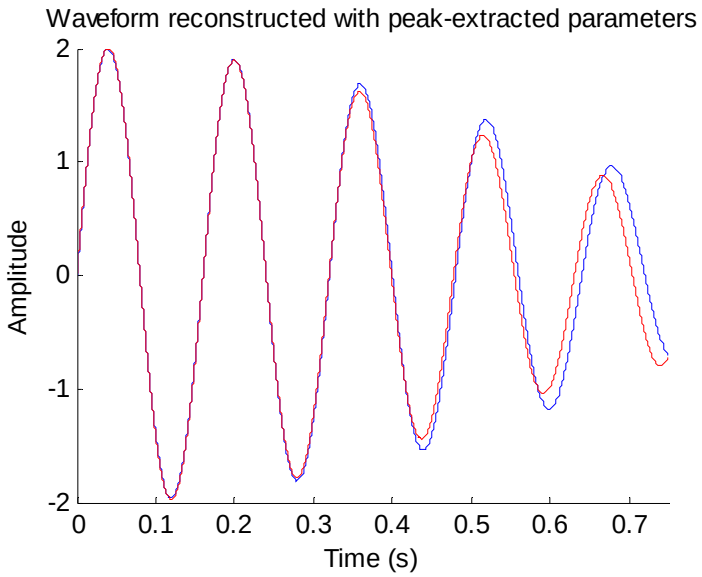

Figure 42. Results of separation on mixture of sines using single input. The frequencies were 6 and 6.5 Hz, respectively, with unity amplitude and zero phase.The recovered signal (red) tracks the original (blue) over the first two peaks. Divergence begins to occur later in time.

Figure 42 illustrates that during the early portion of the waveform in between the first and second peaks which were used for the estimation process, the original (blue) and recovered (red) waveforms are virtually indistinguishable. However, as we move further in time, they begin to deviate. Note that the waveforms track very closely even at areas in between the peaks, thus supporting the notion that, in analogy with the Nyquist sampling theorem, under the correct circumstances local maxima may serve to fully characterize the signal, and carry all information necessary to reproduce the waveform. Work towards a proof of this is discussed in Chapter 7.

Next, we examine a situation in which the original signals were closer in frequency, with separation 0.1 Hz.

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
6.00	4.83	20.0	1.0	.007	99	0.0	+0.32	10.0
6.10	6.05	0.80	1.0	1.99	99	0.0	-0.00	0.0

Table 4. Comparison of original and recovered parameters for a more closely spaced pair of sines than in Table 3. Results were much poorer, although waveforms appear to match better as in next figure. (See text for explanation of paradox.)

The original and recovered waveforms for this case are shown in Figure 43.

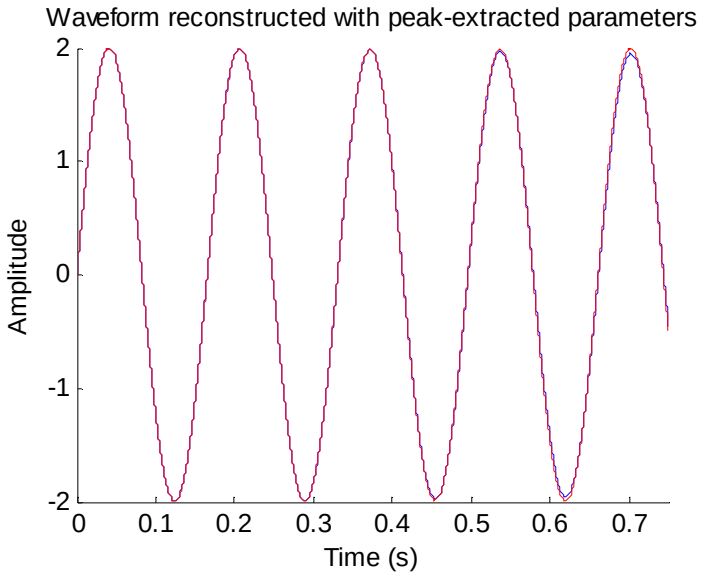

Figure 43. Original (blue) and recovered (red) waveforms for mixture of more closely spaced sines. The frequencies were 6 and 6.1 Hz, respectively, with unity amplitude and zero phase. Visual match is better, although recovered parameters are worse, as in Table 4. (See text for explanation of paradox.)

The estimates found by the algorithm were very poor as compared to the true parameters. However, the reconstructed waveform is very similar to the original, and divergence becomes apparent only after a longer time than in the first case. The reason again appears to be the uncertainty principle. In order to separate closer frequencies, we need a longer observation time.

5.5.4 Multiple Channels

To improve results, it was decided to study the effect of differentially weighting the higher and lower frequency components of the mixture to simulate the effect of passing the mixture through a filter bank, and using the output of two different channels for analysis, rather than a single version, as before.

If we know the weighting of the filter bank on the respective signal components we can proceed as follows:

We are given two versions of the sine mixture, which should be considered as outputs of two channels.

$$x_1 = g_1 s_1 + s_2$$
$$x_2 = s_1 + g_2 s_2 \tag{5.4}$$

The coefficients g_1 and g_2 represent the gain factors, the amount by which a signal is boosted as compared to the other in a particular filter. These factors are assumed to be greater than 1.

To solve, we form estimates of amplitude, frequency and phase at the output of the first filter, i.e., using the peaks of x_1. However, the estimate that we have now obtained has in effect been boosted by the gain factor g_1 for that band. To compensate we must normalize by this same factor. This removes the bias or contribution of that filter. We then subtract this normalized estimate from the output of the second filter, x_2. The rationale is that since the lower-frequency filter has presumably emphasized the lower-frequency sinusoid, we would like to remove that sinusoid in accordance with our best knowledge of its parameters, thus leaving a residual which contains a purer version of the higher-frequency sinusoid. We then estimate parameters of this residual and normalize, as before, by the gain factor of the second filter. We expect that

this will yield a more accurate estimate of the lower-frequency sinusoid. We continue to iterate back and forth in this manner until convergence.

More formally, the steps are:

1) Given $x_1 = g_1 a_1 \sin(2\pi f_1 t + p_1) + a_2 \sin(2\pi f_2 t + p_2)$.

2) Given $x_2 = a_1 \sin(2\pi f_1 t + p_1) + g_2 a_2 \sin(2\pi f_2 t + p_2)$.

3) Estimate $\hat{a}_1, \hat{f}_1, \hat{p}_1$ using peaks of x_1.

4) Normalize $\hat{a}_{1norm} = \hat{a}_1 / g_1$, since we know *a priori* that it has been emphasized in x_1 due to gain factor g_1.

5) Construct $\hat{s}_1 = \hat{a}_{1norm} \sin(2\pi \hat{f}_1 t + \hat{p}_1)$.

6) Compute residual $R_2 = x_2 - \hat{s}_1$.

7) Estimate $\hat{a}_2, \hat{f}_2, \hat{p}_2$ of R_2 using peaks.

8) Normalize $\hat{a}_{2norm} = \hat{a}_2 / g_2$, since we know *a priori* that it has been emphasized in x_2 due to gain factor g_2.

9) Construct $\hat{s}_2 = \hat{a}_{2norm} \sin(2\pi \hat{f}_2 t + \hat{p}_2)$.

10) Compute residual $R_1 = x_1 - \hat{s}_2$.

11) Estimate $\hat{a}_1, \hat{f}_1, \hat{p}_1$ of R_1 using peaks.

12) Repeat from step 4 until convergence.

Figure 44 shows a schematic diagram of the modified algorithm. Note that here, as well, Estimate 1 refers to $\hat{s}_1$ and Estimate 2 refers to $\hat{s}_2$.

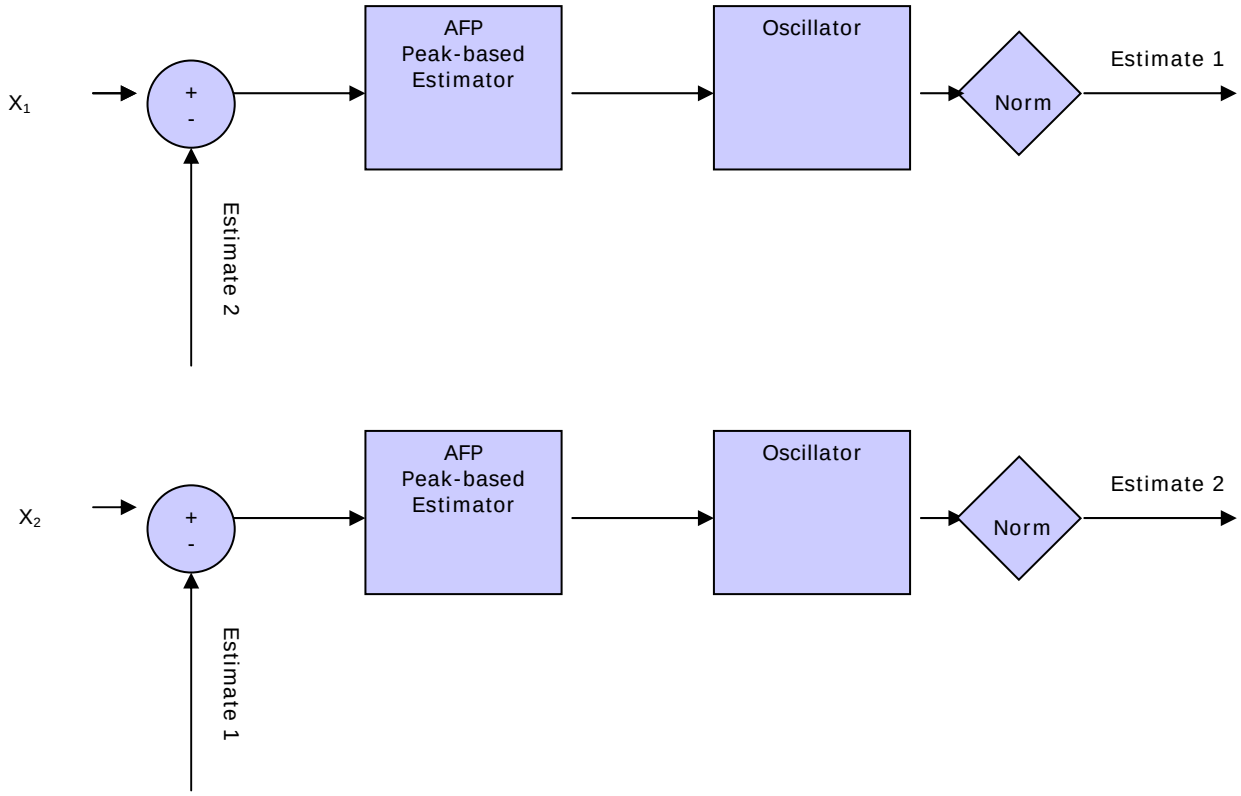

Figure 44. Modification of original algorithm to use multiply weighted versions of input mixture.

The first weighted mixture x_1 is applied to the upper AFP block and the resulting parameter estimates are used by the Oscillator block to generate a sine wave with these parameters. The amplitude of this sine is then normalized by the known gain factor g_1 and this yields Estimate *1*. This estimate is subtracted from the second weighted mixture x_2 and the residual is then fed into the lower AFP, Oscillator and Norm blocks, as before, yielding Estimate *2*, after normalization by g_2. This is then subtracted from the original first weighted mixture x_1 and the residual is again fed into the upper AFP, Oscillator and Norm blocks, and the process is repeated until convergence.

Results were much improved, even with extremely closely spaced frequencies and with significant amplitude differences, provided gain factors were known *a priori*. (It could even separate two signals of equal frequency but different phase, assuming phase-sensitive filters are available such that they would provide a net gain to one phase over another phase. Such a realization would involve nonlinear filters or a PLL type of arrangement.)

Following are representative results using two closely spaced sines. The equations are as before:

(5.5)
$$s_1(t) = a_1 \sin(2\pi f_1 t + \phi_1)$$
$$s_2(t) = a_2 \sin(2\pi f_2 t + \phi_2)$$

(5.6)
$$x_1(t) = g_1 s_1(t) + s_2(t)$$
$$x_2(t) = s_1(t) + g_2 s_2(t)$$

The parameters are:

$a_1 = 3.5,\ f_1 = 100,\ \phi_1 = +.5$

$a_2 = 2.5,\ f_2 = 100.1,\ \phi_2 = -.5$

Gain factors $g_1 = 1.1,\ g_2 = 1.1$

Sampling rate = 500 KHz.

We deliberately chose very closely spaced frequencies and very low gain factors to illustrate the performance of the algorithm in difficult cases. As we will demonstrate, we can obtain perfect results in a similar example to that used in Section 5.5.3 by using multiple channels.

The following 3 figures show the convergence of the algorithm

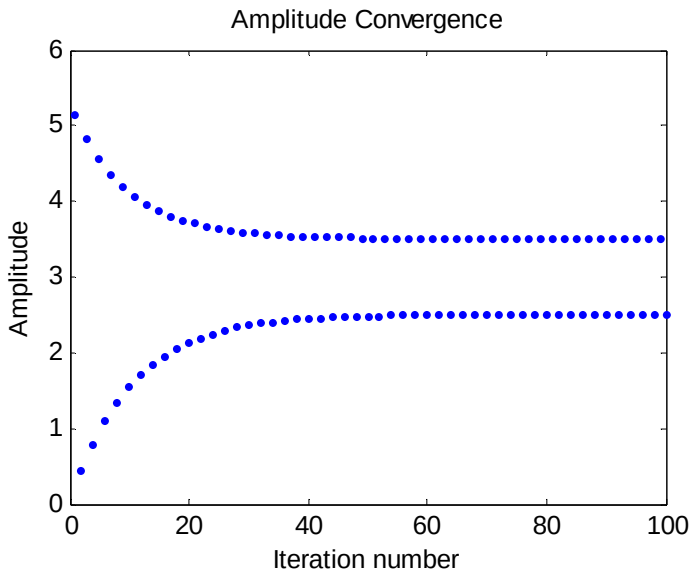

Figure 45. Convergence of amplitude estimates with increasing iterations. Estimates alternate between a1 and a2. Original values were a1=3.5, a2=2.5.

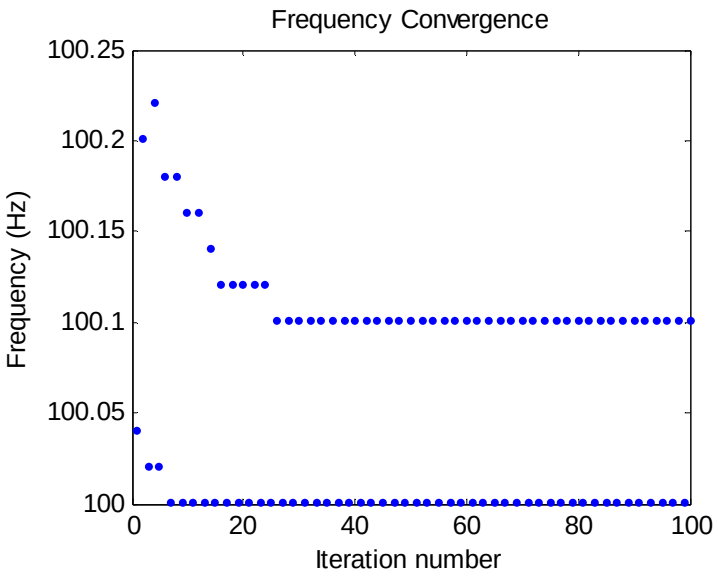

Figure 46. Convergence of frequency estimates with increasing iterations. Estimates alternate between f1 and f2. Original values were f1=100 Hz, f2=100.1 Hz.

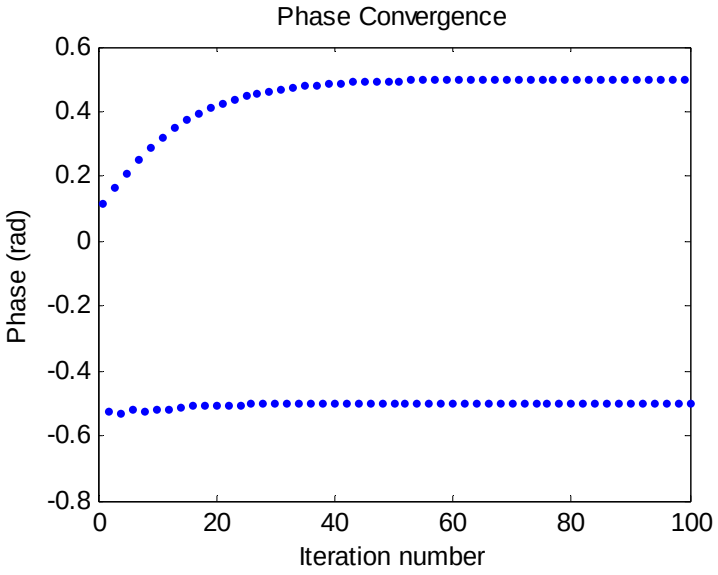

Figure 47. Convergence of phase estimates with increasing iteration. Estimates alternate between p1 and p2. Original values were p1=+0.5, p2=-0.5.

Results are shown in Table 5.

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
100.0	100.0	0.00	3.50	3.50	0.09	+0.50	+0.50	0.08
100.1	100.1	0.00	2.50	2.49	0.31	-0.50	-0.50	0.01

Table 5. Results for multiple channel Iterative-Subtraction algorithm.

Although we do not have a formal proof, we note that convergence was virtually perfect in all cases tried (many of which are not listed here), including mixtures of multiple sinusoids, provided the gain factor was known and was sufficiently high. [We note similar statements in the literature in other contexts in which convergence is difficult to prove (Tropp, 2003) citing (Cadzow, 1988) based on work of (Zangwill, 1969)]. Numerical results were identical to original within roundoff error. Although these results could be found algebraically by solving Equations 5.4 which are two functional equations in two unknowns, what is noteworthy with this method is that it is not necessary to have the entire waveforms, only the locations of the local maxima of each.

As demonstrated here, in many cases, even a gain factor as low as 1.1 between the filters is enough to separate adequately.

5.5.5 FM Signals

Since with real-world signals, frequency may change continually, we examine the performance of the algorithm in frequency-varying situations. The following series of plots show how the successive peaks of a mixture can be used to track the instantaneous frequency of an FM signal. The mixture consists of a pair of sines, one with constant frequency, and one frequency-varying according to the equations below.

(5.7)
$$s_1(t) = a_1 \sin\left[2\pi\phi_{fm}(t)\right]$$
$$s_2(t) = a_2 \sin(2\pi f_2 t + \phi_2)$$

(5.8)
$$x_1(t) = g_1 s_1(t) + s_2(t)$$
$$x_2(t) = s_1(t) + g_2 s_2(t)$$

where

$a_1 = 3.5$, $\phi_{fm} = \int (8 + 10t)dt$

$a_2 = 2.5$, $f_2 = 9$, $\phi_2 = -.5$

Gain factors $g_1 = 2$, $g_2 = 2$

Sampling rate = 100.001 KHz.

We note that sampling rate has been lowered here to reflect lower frequencies of signals. There will still be an adequate number of samples per cycle to locate the local maxima accurately

enough. We occasionally used odd numbers such as 100.001 KHz, to see if certain roundoff errors could be reduced by decreasing the symmetry of the sampling points.

The algorithm uses two peaks at a time. The advantage of using such a short time-analysis interval is that each pair can be analyzed independently, without the effect of past history. One obtains a set of parameters estimates for each inter-peak interval. Although this leads to slight discontinuities, these can later be smoothed. In addition, if harmonicity holds, one may track behavior of higher-frequency harmonics which have more peaks per unit time and divide by the harmonic number to get an additional, and possibly more accurate estimate of the time course of the fundamental.

Each of the following diagrams in Figure 48 shows the original (blue) and reconstructed (red) waveforms for the mixture. The reconstructed waveform in each figure treats the calculated frequency for that interval as a constant frequency throughout. Therefore, correct agreement is only obtained within that local interval. As can be noted, the interval in which the red and blue lines match shifts from one figure to the next, occurring at the time corresponding to the peak pair upon which the analysis was performed.

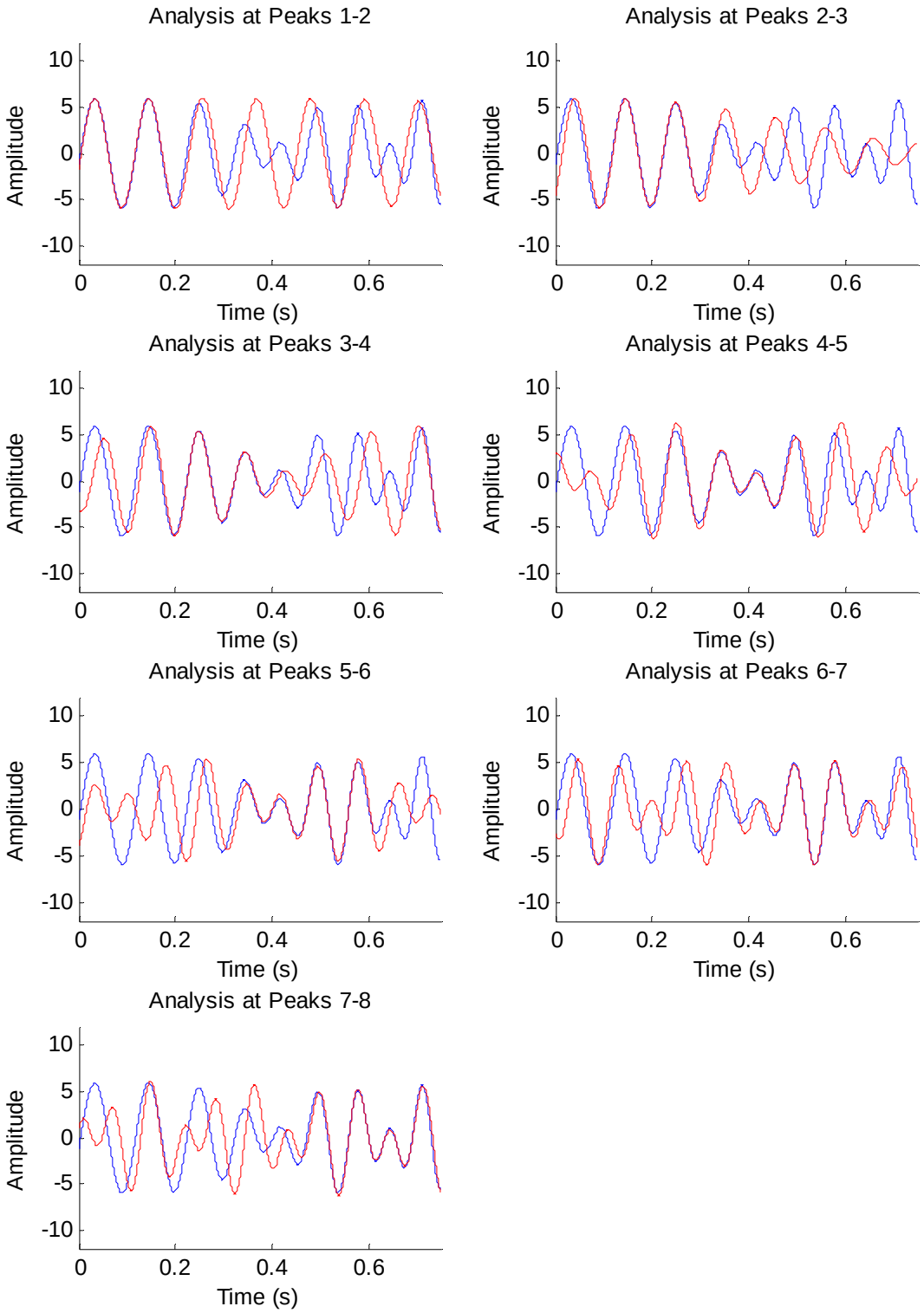
Analysis at Peaks 1-2
Analysis at Peaks 2-3
Analysis at Peaks 3-4
Analysis at Peaks 4-5
Analysis at Peaks 5-6
Analysis at Peaks 6-7
Analysis at Peaks 7-8
Amplitude
Time (s)
10
5
0
-5
-10
0
0.2
0.4
0.6

Figure 48 (Previous page). A sequential series of plots showing performance of Iterative-Subtraction algorithm on a mixture of a constant-frequency 9 Hz signal with a signal whose frequency varies as 8+10t. Analysis is performed using coordinates of two successive peaks at a time beginning with 1st and 2nd and continuing until 7th and 8th as shown in corresponding plots. For each plot, the original (blue) and reconstructed (red) signals are shown. Accurate matches are obtained within the region between the peak pairs used for each analysis, supporting the idea that the method may be useful for capturing instantaneous parameters of time-varying signals.

Table 6 and Table 7 show the numerical results obtained for each pair.

Signal 1			
Peaks	Amplitude	Frequency	Phase
1-2	3.50	8.84	-0.11
2-3	3.49	9.90	-1.06
3-4	3.57	10.87	-2.53
4-5	3.56	11.69	2.09
5-6	3.43	12.73	-0.72
6-7	3.50	13.38	-2.76
7-8	3.55	14.04	1.15

Table 6. Extracted parameters for frequency-varying component of mixture. Note increasing frequency with time (peak number).

Signal 2			
Peaks	Amplitude	Frequency	Phase
1-2	2.50	9.09	-0.56
2-3	2.46	9.15	-0.66
3-4	2.44	8.94	-0.45
4-5	2.84	8.83	-0.06
5-6	2.19	9.57	-2.24
6-7	2.51	9.02	-0.51
7-8	2.74	9.12	-0.96

Table 7. Extracted parameters for fixed frequency component of mixture. Estimated frequency remains in the neighborhood of 9 Hz, but fluctuates somewhat with time, as described in text.

Note that the calculated frequency of the first signal increases from one peak pair to the next, as it should. The calculated frequency of the second signal remains close to the correct value of 9, but deviates somewhat from pair to pair. The amplitude and frequency estimates at pair 5-6 are somewhat poorer than expected, as is the visual match between the waveforms in the corresponding figure. The reason for this may be due to the fact that the peaks in the weighted signals x_1 and x_2 may have become "out of sync" with each other such that the closest peak pairs at the specified time were farther apart from each other than is usual. In other words, since low-frequency signals have less peaks than high-frequency signals, one may need to look over different time regions of the respective signal when searching for the closest peak pair to a

given time point. In one channel, the closest pair could conceivably be located mostly to the left of the desired point in time, while the in the other it might be mostly to the right.

In general, it is apparent that the effect of the FM signal was to cause fluctuations in the estimate of the constant-frequency signal, as well. This is undesirable, but possibly unavoidable due to our use of a staircase constant-frequency approximation, rather than an actual smooth ramp. We will have more to say about this in Chapter 6.

5.5.6 Multiple Signals

For 3 or more sinusoids, the algorithm is similar. We require at least one channel per sinusoid or we will merge parameter estimates in error. (Using more channels will just produce extra estimates of value zero, which will not affect accuracy.) For simplicity, we assume that channel i contains a weighted mixture x_i consisting of the sum of a boosted version of sine s_i by an appropriate gain factor g_i plus the remaining sines with unity weights each.

1) Given $x_1 = g_1 a_1 \sin(2\pi f_1 t + p_1) + a_2 \sin(2\pi f_2 t + p_2) + a_3 \sin(2\pi f_3 t + p_3)$.

2) Given $x_2 = a_1 \sin(2\pi f_1 t + p_1) + g_2 a_2 \sin(2\pi f_2 t + p_2) + a_3 \sin(2\pi f_3 t + p_3)$.

3) Given $x_3 = a_1 \sin(2\pi f_1 t + p_1) + a_2 \sin(2\pi f_2 t + p_2) + g_3 a_3 \sin(2\pi f_3 t + p_3)$.

4) Start with $\hat{s}_1 = \hat{s}_2 = \hat{s}_3 = 0$.

5) Estimate $\hat{a}_1, \hat{f}_1, \hat{p}_1$ using peaks of x_1.

6) Normalize $\hat{a}_{1norm} = \hat{a}_1 / g_1$, since know *a priori* that it has been preemphasized by gain factor g_1 in x_1.

7) Construct $\hat{s}_1 = \hat{a}_{1norm} \sin(2\pi \hat{f}_1 t + \hat{p}_1)$.

8) Compute residual $R_2 = x_2 - \hat{s}_1 - \hat{s}_3$. Note: last term is initially 0.

9) Estimate $\hat{a}_2, \hat{f}_2, \hat{p}_2$ using peaks of R_2.

10) Normalize $\hat{a}_{2norm} = \hat{a}_2 / g_2$, since know *a priori* that it has been preemphasized by gain factor g_2 in x_2.

11) Construct $\hat{s}_2 = \hat{a}_{2norm} \sin(2\pi \hat{f}_2 t + \hat{p}_2)$.

12) Compute residual $R_3 = x_3 - \hat{s}_1 - \hat{s}_2$.

13) Estimate $\hat{a}_3, \hat{f}_3, \hat{p}_3$ of using peaks of R_3.

14) Normalize $\hat{a}_{3norm} = \hat{a}_3 / 2$, since know *a priori* that it has been preemphasized by gain factor g_3 in x_3.

15) Construct $\hat{s}_3 = \hat{a}_{3norm} \sin(2\pi \hat{f}_3 t + \hat{p}_3)$.

16) Compute residual $R_1 = x_1 - \hat{s}_2 - \hat{s}_3$.

17) Estimate $\hat{a}_1, \hat{f}_1, \hat{p}_1$ using peaks of R_1.

18) Repeat from step 6 until convergence.

We now show an example of separation on a mixture of 3 signals. Parameters of original and recovered sinusoids are as in Table 8. The 3 gain factors are each 1.1, which is again a deliberately chosen low value to demonstrate the strength of the algorithm. Sampling rate is 100.001 KHz. Figure 49 shows the 3 differently-weighted mixtures, and the original (unweighted) signal. Note that the differences in the peak locations are so small as to be almost unnoticeable in this figure. In Figure 50 we have zoomed in on the 6th and 7th peaks, the pair arbitrarily chosen for analysis, to better illustrate the small differences in peak coordinates between the differently weighted versions.

Orig Amp	Recov Amp	Pct Error	Orig Freq	Recov Freq	Pct Error	Orig Phase	Recov Phase	Pct Error
3.50	3.50	0.10	13.00	13.00	0.01	0.50	0.50	0.05
2.50	2.50	0.05	14.10	14.11	0.04	-0.50	-0.51	0.43
5.00	5.00	0.04	14.00	14.00	0.01	0.00	0.01	0.22

Table 8. Original and recovered parameters for a mixture of 3 sinusoids using the Iterative-Subtraction algorithm.

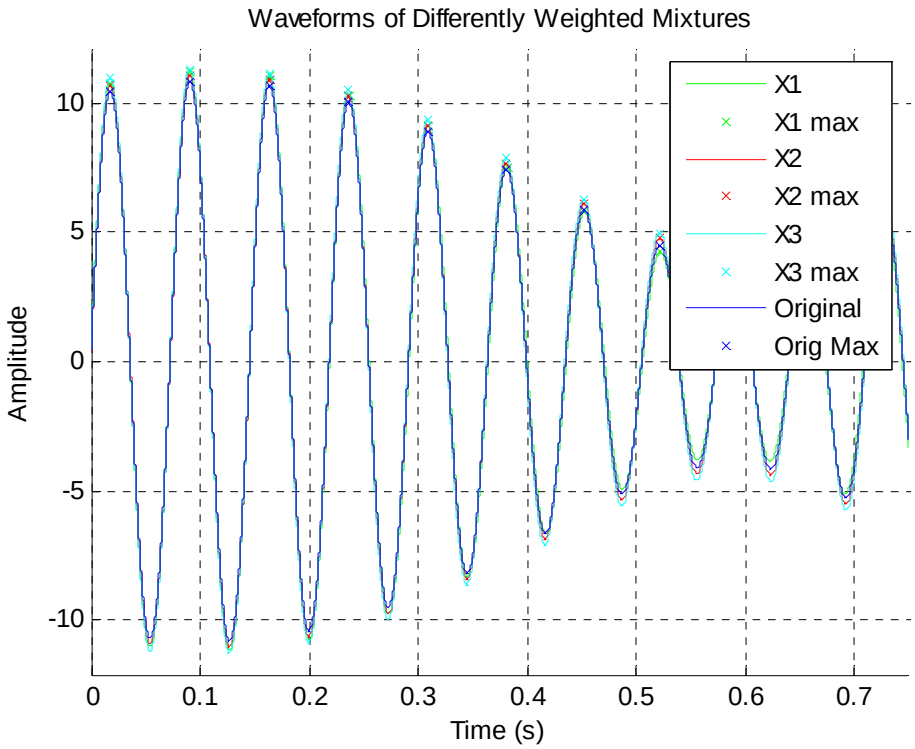

Figure 49. Comparison of three differently weighted versions and original unweighted version of mixture of three sinusoids. Note that differences in the shapes and peak locations are so minor as to be nearly unnoticeable, yet algorithm is able to correctly determine parameters of all three sinusoids based solely on positions of 6th and 7th local maxima in each of the three weighted versions of the mixture.

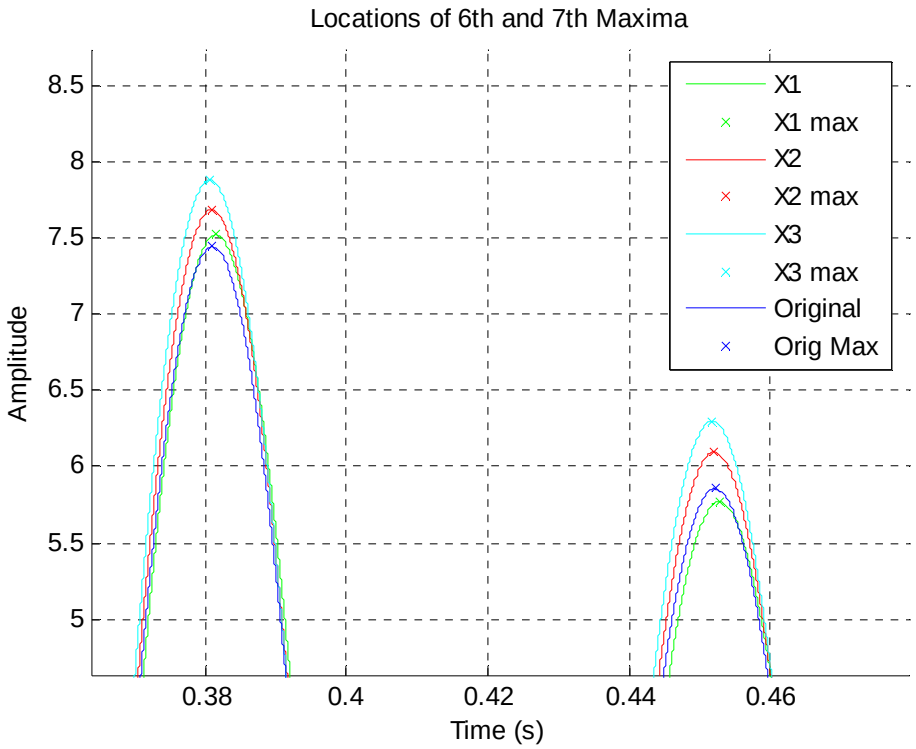

Figure 50. The locations of the 6th and 7th local maxima in the three differently weighted versions and original unweighted version of the mixture of three sinusoids. Slight shifts in vertical and horizontal positions can be seen in this zoomed view of Figure 49. These differences are used by the algorithm to compute the parameters of each sinusoid.

Figure 51, Figure 52 and Figure 53 illustrate the convergence trajectories of amplitude, frequency and phase, respectively. We have added the use of color to better distinguish the parameters of the three sinusoids from each other. Indexing of iterations now refers to each color individually, and hence total number of iterations is three times the value indicated.

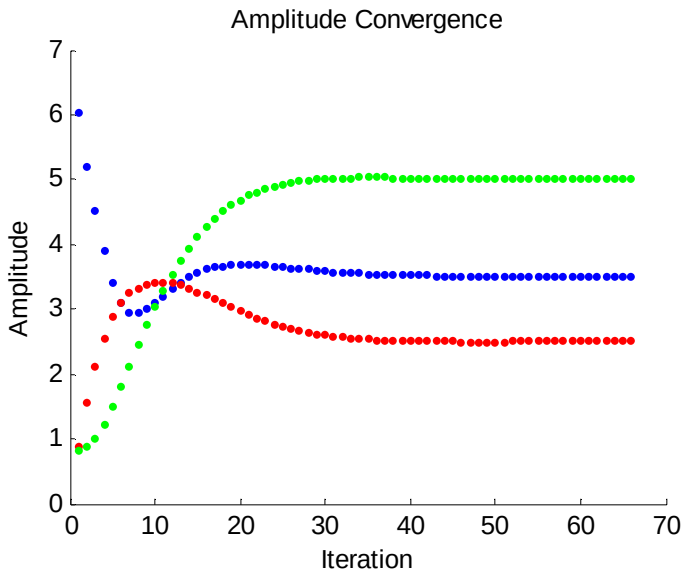

Figure 51. Convergence of amplitude estimates with increasing iterations. (Iterations are now indexed per color, so total number is three times value indicated when comparing with Figure 45.) Original values were a1=3.5, a2=2.5, and a3 =5.0.

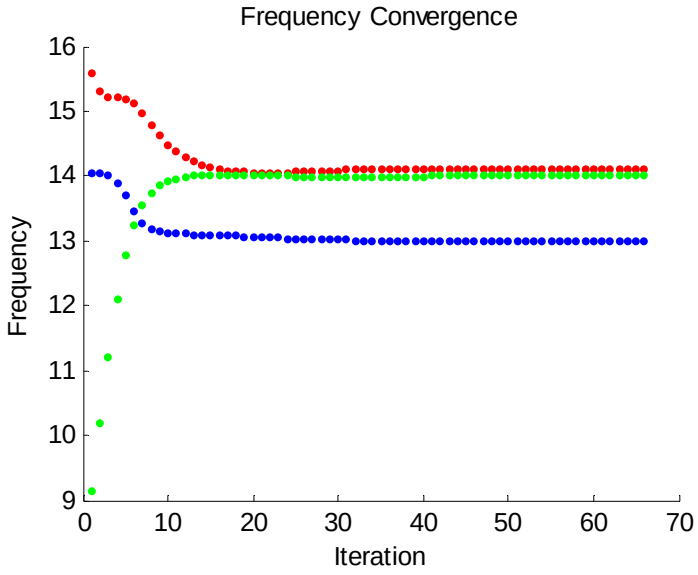

Figure 52. Convergence of frequency estimates with increasing iterations. (Iterations are now indexed per color, so total number is three times value indicated when comparing with Figure 46.) Original values were f1=13.0, f2=14.1, and f3 =14.0 Hz.

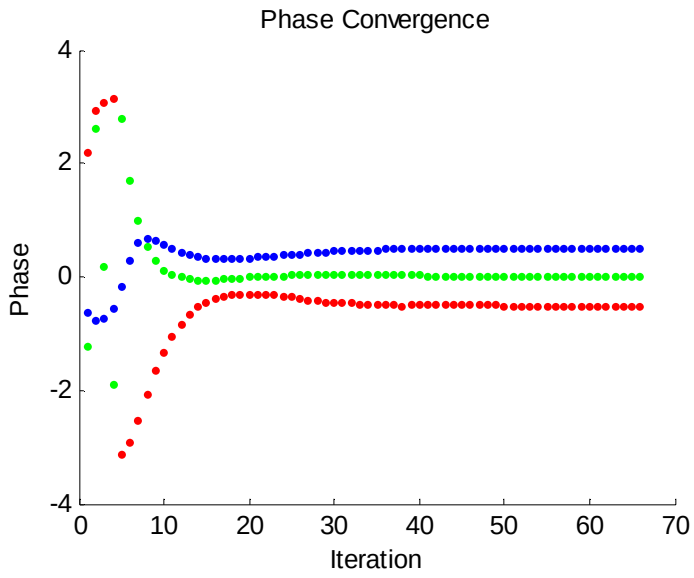

Figure 53. Convergence of phase estimates with increasing iterations. (Iterations are now indexed per color, so total number is three times value indicated when comparing with Figure 47.) Original values were phi1=0.5, phi2=-0.5, and phi3=0.0.

5.5.7 Mathematical Simulations of Actual Filters

Although the results of the above algorithms are highly accurate, they are not realistic simulations of actual filters. The reason is that in a real filter the gain factor is frequency-dependent, not a constant as we have used up till now. While we may indeed have a perfectly characterized description of the gain of the filter $H(f)$ for each frequency, however, frequency is initially an unknown parameter, and hence the appropriate gain factor is unknown, as well. To take that into account, one needs to modify the algorithm so that the gain factor is represented as $H_i(f_j)$ which is the gain of the i^{th} filter to the j^{th} frequency component.

The steps are as follows for an example using 3 bands and 3 sinusoids:

1) Given filter shapes $H_1(f), H_2(f), H_3(f)$ for all values of f in frequency domain.

2) Given

$$x_1 = H_1(f_1)a_1\sin(2\pi f_1 t + p_1) + H_1(f_2)a_2\sin(2\pi f_2 t + p_2) + H_1(f_3)a_3\sin(2\pi f_3 t + p_3)$$

3) Given

$$x_2 = H_2(f_1)a_1\sin(2\pi f_1 t + p_1) + H_2(f_2)a_2\sin(2\pi f_2 t + p_2) + H_2(f_3)a_3\sin(2\pi f_3 t + p_3)$$

4) Given

$$x_3 = H_3(f_1)a_1 \sin(2\pi f_1 t + p_1) + H_3(f_2)a_2 \sin(2\pi f_2 t + p_2) + H_3(f_3)a_3 \sin(2\pi f_3 t + p_3)$$

5) Start with $\hat{s}_1 = \hat{s}_2 = \hat{s}_3 = 0$.

6) Estimate $\hat{a}_1, \hat{f}_1, \hat{p}_1$ using peaks of x_1.

7) Normalize $\hat{a}_{1norm} = \hat{a}_1 / H_1(\hat{f}_1)$ to remove filter bias.

8) Construct $\hat{s}_1 = \hat{a}_{1norm} \sin(2\pi \hat{f}_1 t + \hat{p}_1)$.

9) Compute residual $R_2 = x_2 - H_2(\hat{f}_1)\hat{s}_1 - H_2(\hat{f}_3)\hat{s}_3$. Note: last term is initially 0.

10) Estimate $\hat{a}_2, \hat{f}_2, \hat{p}_2$ using peaks of R_2.

11) Normalize $\hat{a}_{2norm} = \hat{a}_2 / H_2(\hat{f}_2)$ to remove filter bias.

12) Construct $\hat{s}_2 = \hat{a}_{2norm} \sin(2\pi \hat{f}_2 t + \hat{p}_2)$.

13) Compute residual $R_3 = x_3 - H_3(\hat{f}_1)\hat{s}_1 - H_3(\hat{f}_2)\hat{s}_2$.

14) Estimate $\hat{a}_3, \hat{f}_3, \hat{p}_3$ using peaks of R_3.

15) Normalize $\hat{a}_{3norm} = \hat{a}_3 / H_3(\hat{f}_3)$ to remove filter bias.

16) Construct $\hat{s}_3 = \hat{a}_{3norm} \sin(2\pi \hat{f}_3 t + \hat{p}_3)$.

17) Compute residual $R_1 = x_1 - H_1(\hat{f}_2)\hat{s}_2 - H_1(\hat{f}_3)\hat{s}_3$.

18) Estimate $\hat{a}_1, \hat{f}_1, \hat{p}_1$ using peaks of R_1.

19) Repeat from step 7 until convergence.

5.5.8 Discussion

While the underlying foundation upon which this method of parameter estimation is based seems to be firm, as we saw in a number of examples, this is true in the case of known gain factors. However, reliably translating this method to the case where gain factors are unknown and frequency-dependent, as occurs when dealing with simulations of linear filters, is more difficult. At the end of this chapter, we show comparative test results of all the algorithms. While we had occasional success with the Iterative-Subtraction method in the unknown, weighted case described in Section 5.5.7, for the most part computation was extremely slow and prone to error buildup during the tedious process of subtracting band after band. A possible reason may be that in the unknown weighted case, the gain factors can attain values less than 1.0, depending on the frequency difference between a given component and filter, since $H_i(f_j)$, a frequency-dependent variable, now assumes the role of the heretofore fixed gain factor of the known, weighted case which was always set greater than 1.0.

Nevertheless, the main accomplishment is a proof-of-concept that the use of temporal information from multiple versions of a signal, specifically, the coordinates of local maxima, could potentially be used to gain more precise knowledge of the underlying parameters. We continue with this theme, but search for better implementations.

5.6 Peak-Locus Algorithm

With the aim of improving performance, we propose a method of dimensionality reduction based on the behavior of local maxima in multiple bands. The importance of this method is that it separates the contribution of those bands with novel information on the presence of hitherto unknown frequency components from those bands which merely contain redundant repetition of information already known.

Consider a mixture of sines which is passed through a filter bank with overlapping filters. Because the characteristic frequency (CF) of each is slightly different, each successive filter is expected to weight the mixture slightly differently. As a result, there is no simple way to compare the waveforms of the separate channels, as they will all differ slightly as well, similar to the situation in Figure 49 and Figure 50, earlier. Because of this, it is very difficult to identify

the correct number of actual source components in the input space, due to the large number of dissimilar images in the output space. The maxima will also differ. However, there is a class of filter bank which will produce identical but scaled images of the mixture waveform, even with filters of different CF, under conditions we will describe. This is the exponential filter bank. We note that the following discussion assumes ideal filters that completely conform to the theoretical mathematical shapes, and further assumes that the signals are in steady state, after settling of all transients. We also assume a uniform linear phase response so that all frequencies have equal delays in all filters. The theory we develop here will be used in the Simultaneous-Equation formulation, as well, which we will introduce in Section 5.7. The proof is as follows:

Consider a filter with the following response:

$$H_1(f) = e^{-\frac{|f - cf_1|}{bw}} \tag{5.9}$$

where

cf_1 is the center frequency of the filter.

bw is the bandwidth of the filter.

This can be factored into:

$$H_1(f) = e^{-\frac{|f|}{bw}} e^{\frac{|cf_1|}{bw}} \tag{5.10}$$

provided that the frequencies are restricted to be on the same side of the filter so that the sign doesn't change within the argument of the absolute value function over the range of frequencies present.

The next filter can be written as

$$H_2(f) = e^{-\frac{|f|}{bw}} e^{\frac{|cf_2|}{bw}} \tag{5.11}$$

etc.

Therefore, if we have a pair of sines at f_1 and f_2 passing through an exponential filter bank, the complete response of the filter at cf_1 is

$$(5.12)\quad a_1 \sin(2\pi f_1 t) e^{-\frac{|f_1|}{bw}} e^{\frac{|cf_1|}{bw}} + a_2 \sin(2\pi f_2 t) e^{-\frac{|f_2|}{bw}} e^{\frac{|cf_1|}{bw}} = \left[a_1 \sin(2\pi f_1 t) e^{-\frac{|f_1|}{bw}} + a_2 \sin(2\pi f_2 t) e^{-\frac{|f_2|}{bw}} \right] e^{\frac{|cf_1|}{bw}}$$

Similarly, the complete response of the filter at cf_2 is

$$(5.13)\quad a_1 \sin(2\pi f_1 t) e^{-\frac{|f_1|}{bw}} e^{\frac{|cf_2|}{bw}} + a_2 \sin(2\pi f_2 t) e^{-\frac{|f_2|}{bw}} e^{\frac{|cf_2|}{bw}} = \left[a_1 \sin(2\pi f_1 t) e^{-\frac{|f_1|}{bw}} + a_2 \sin(2\pi f_2 t) e^{-\frac{|f_2|}{bw}} \right] e^{\frac{|cf_2|}{bw}}$$

These expressions, (5.12) and (5.13) are just scaled versions of each other, as can be seen by examining the right hand side of each, since the part within the brackets is the same in both, and the factors outside the brackets are constant. Since the signals in the two filters are identical except for an amplitude scaling, the peaks must occur at identical times.

The importance of this is that channels which have coincident peaks indicate that all frequency components are on the same side of all of the corresponding filters. If we see a change in the peak pattern in a particular filter as compared to an adjacent filter, it indicates that there is a frequency component which lies on the right side of CF of one filter, but on the left side of CF of the adjacent filter. This gives an indication of the frequency of the signal component, which must be located in between the CF of the two filters.

To see the behavior more clearly, we can plot the peak locations of each filter in the time-frequency plane. We call this a Peak-Locus plot. In regions of the plane which lie below the lowest-frequency spectral component, the peaks will be coincident in time, and will form straight vertical lines. In transition regions where there are spectral components that lie on different sides of a given filter's CF, the lines will curve or break. In regions which are above the frequency of the highest-frequency component, the peaks will again be coincident and form vertical lines.

There is one additional consideration we must mention. Recalling our discussion in Section 5.4.3, that the highest-frequency component dominates the peaks, all else being equal, it is necessary that the higher-frequency side of each filter be attenuated more steeply than the lower-side, i.e., requires use of asymmetric filters. Otherwise, components which lie below a

higher-frequency component are not likely to be noticed, as they will not contribute any significant change to the peak pattern. In a sense, they will be masked by the higher-frequency component. This will all become clearer in the following example:

Figure 54 shows a mixture of two inverted cosine signals chosen so that the resultant will have two symmetrical center peaks.

$$s_1 = -\cos(2\pi 5t) \qquad (5.14)$$
$$s_2 = -\cos(2\pi 7t)$$

Due to the deliberately designed symmetrical peaks, we can compute an equivalent monocomponent signal that passes through these two peaks. This is illustrated by the black dotted trace, which passes through the two center peaks at +/- 0.08 seconds, and is coincident at those points with the red trace of the multicomponent signal. The equation of this equivalent curve is:

$$s_{equiv} = -1.74\cos(2\pi 6.18t) \qquad (5.14a)$$

The existence of s_{equiv} clearly demonstrates that a single pair of peaks from a single channel is insufficient to uniquely determine the component parameters, and furthermore cannot even uniquely resolve the number of sources, something we might have suspected from the failure of the single channel separation algorithm in Section 5.5. (We note that the case of Section 5.5 is not exactly analogous to this illustration, as the results there were produced after an Iterative-Subtraction procedure which successively fine-tunes results, and also used the sum of two sinusoids to match the peaks, as opposed to the case here where only a single iteration and a single sinusoid were used. As a case-in-point, the entire near-term waveform closely matched the original in Figure 42 and Figure 43, while in Figure 54, the trough of the dotted-black monocomponent equivalent waveform does not match the trough of the solid-red resultant as can be seen at time 0.0 seconds.)

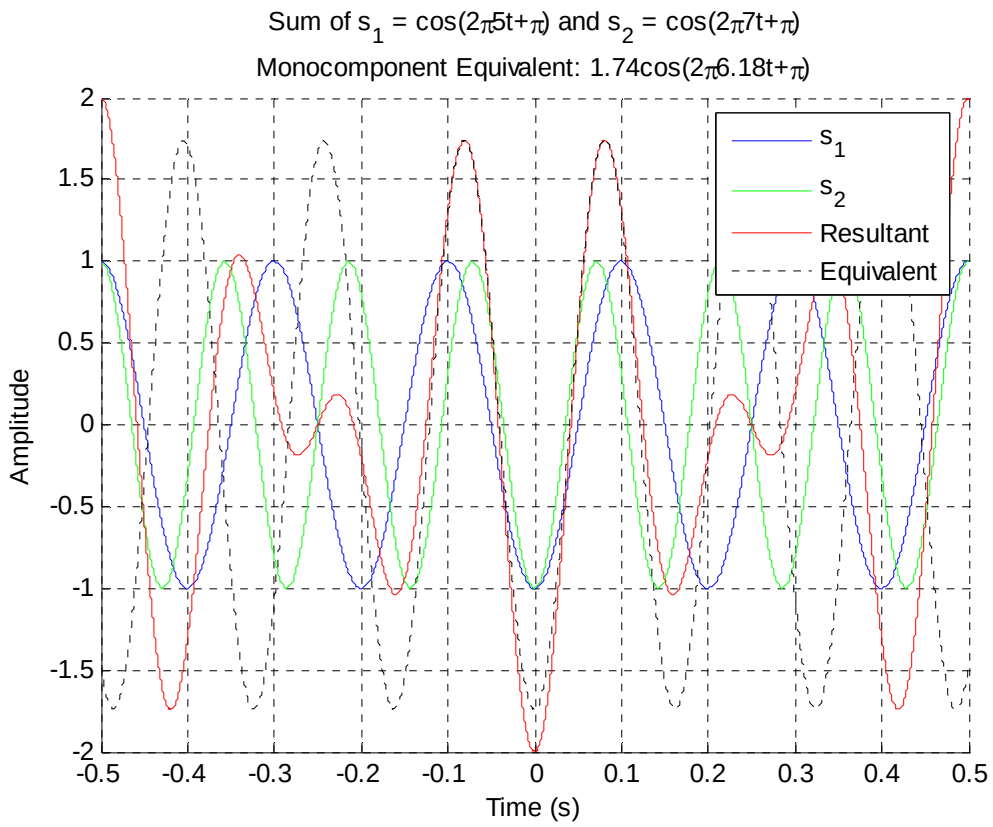

Figure 54. A plot of 5 Hz and 7 Hz inverted cosines and their resultant. Symmetry of two center peaks in resultant at t=+/- 0.08 seconds demonstrates that single-band observation over time region between these peaks is unable to distinguish resultant from the equivalent monocomponent signal 1.74cos(2π 6.18t+π).

In order to distinguish the mixture from the monocomponent case and correctly determine the component parameters, we pass the signal through an asymmetric, overlapping exponential filter bank.

Figure 55 shows such a filter bank consisting of 20 filters ranging from 0 to 10 Hz in increments of 0.5 Hz. The lower bandwidth is 10 Hz, with the higher bandwidth a sharper 1 Hz. We use bandwidth in the sense of writing the filter response $H(f)$ as:

$$\begin{matrix} e^{-\frac{|f-cf|}{bw_{low}}} : f < cf \\ e^{-\frac{|f-cf|}{bw_{high}}} : f > cf \end{matrix} \tag{5.15}$$

For this filter, $bw_{low} = 10$, $bw_{high} = 1$

When we plot the output of each successive filter to the input mixture of our example as in the style of a spectrogram, we obtain the surface of Figure 57. (We use a finer filter spacing of 0.01

Hz in these 3-D illustrations for a smoother, more polished surface, rather than the 0.5 Hz used in the 2-D plot of Figure 55 which was chosen to better bring out the shape of each individual filter.) We note that since the original signal was an even signal with the negative time-axis mirroring the positive axis, we show only the positive time axis in the 3-D plots, and extend the right-hand limit from 0.5 seconds to 1 second to show more of the waveforms. The negative side, of course, exhibits identical behavior.

Each horizontal slice of the surface (parallel to the time axis) represents a single channel output. In the regions below 5 Hz, and above 7 Hz, they are simply scaled versions of each other. In the transition region from 5 Hz to 7 Hz, there is a gradual change in shape, as the weighting of the respective components changes.

Figure 58 shows the Peak-Locus for the same mixture in 3-D. Note that similarly below 5 Hz and above 7 Hz, the lines are straight, since in those regions in which the waveforms scale with respect to each other the peaks must be coincident, and hence parallel vertical structures are formed. In the transition regions, where the waveforms do not scale, the peaks are not coincident, and the structures exhibit curvature.

In order to minimize the transition regions, and thus to bring out the exact location of a component, one might choose to further sharpen the upper sidewall. This will cause the gentle curvature in the transition region to become a sharp break, as we will see shortly. Figure 56 shows the same type of exponential filter bank as in Figure 55, except bw_{high} has now been tightened to 0.01 Hz.

Figure 59 and Figure 60 show the effect of sharpening the upper sidewall so that $bw_{high} = 0.01$. The transition region now becomes much more abrupt, so that instead of a gradual curvature, there is now an effective break in the Peak-Locus plots. It would appear that this would make the measurement of frequency much more precise. However, we note that when real filters are used, as opposed to the mathematically simulated weights used in this example, the response time becomes slower, in accordance with the uncertainty principle that compactness in time yields expansion in frequency, and vice versa, thereby lengthening the transient settling time. This itself may cause curvature of the Peak-Locus plots for a different reason, as will be explained in Section 6.8 on transient response. So the best choice of filter parameters is a

tradeoff, as it must always be. Nevertheless, the advantage of multiple filters is that information from wider-bandwidth and better time-localized filters can be pooled to improve frequency resolution over the single channel case.

Figure 61 and Figure 62 illustrate that the response to s_{equiv} is completely parallel, and can now readily be distinguished from the mixture, which was not possible using a peak pair from a single channel.

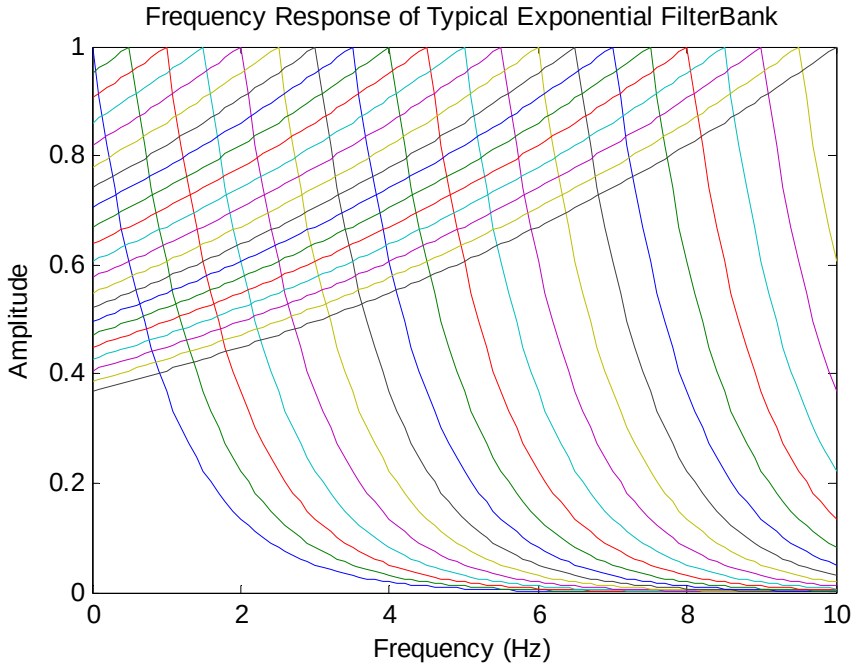

Figure 55. An asymmetrical, exponential filter bank. Lower bandwidth is 10 Hz, upper bandwidth is 1 Hz. Spacing is 0.5 Hz to avoid crowding the image. In practice, one could use as many filters as is computationally practical.

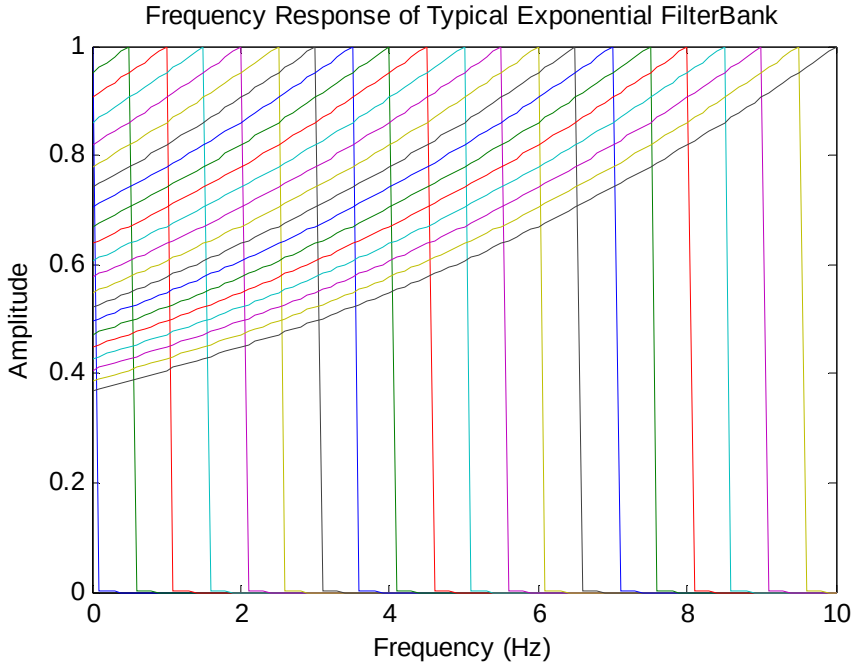

Figure 56. An asymmetrical, exponential filter bank with a tighter upper sidewall than that in Figure 55. Lower bandwidth is 10 Hz, upper bandwidth is 0.01 Hz. Spacing is 0.5 Hz to avoid crowding the image. In practice, one could use as many filters as is computationally practical.

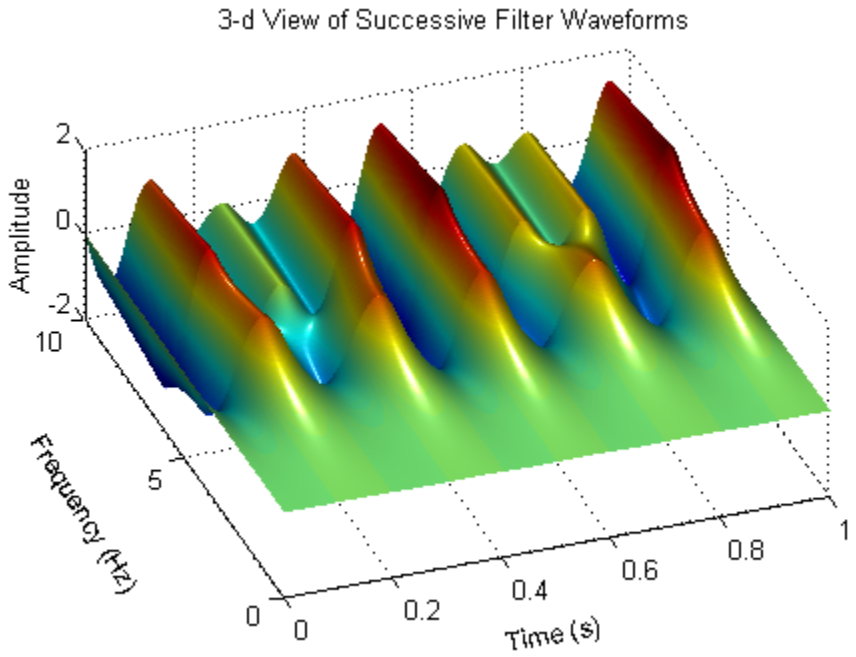

Figure 57. 3-D plot of the signal at the output of each filter in the filter bank. Input was mixture of 5-Hz and 7-Hz cosines each with unity amplitude and 180 degrees phase. Exponential filters had lower bandwidth of 10 Hz and upper bandwidth of 1 Hz. Note that in regions below CF of 5 and above CF of 7, the waveforms are scaled versions of each other. In transition region between 5 Hz and 7 Hz, there is horizontal curvature, due to the disproportionate weighting of the two cosine components.

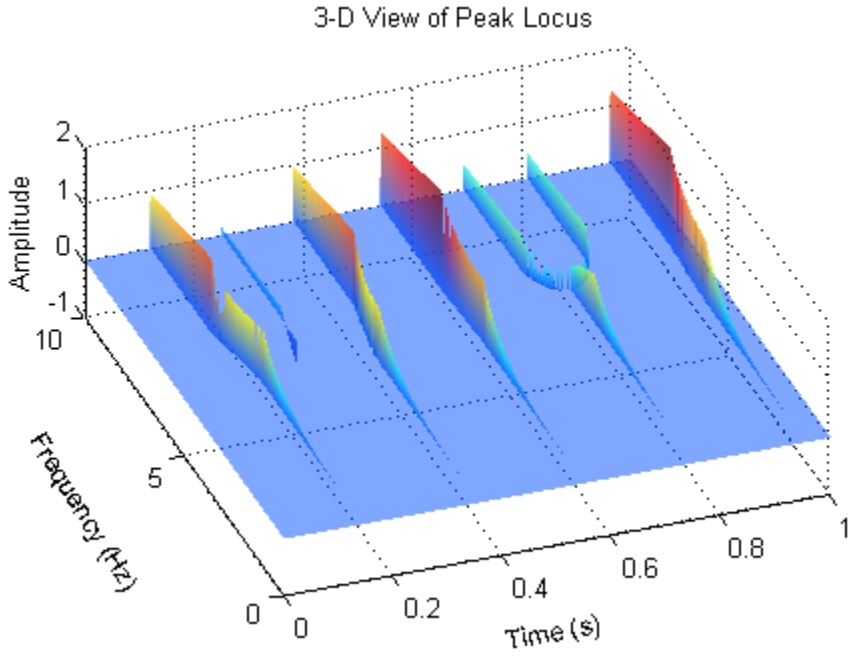

Figure 58. The Peak-Locus (locations of local maxima) of each band output in Figure 57. Note parallel structure throughout except for the curvature in the transition region (filters whose CF lies in between 5 and 7 Hz). Lowest frequency filters show 5 peaks while highest frequency filters show 7 peaks.

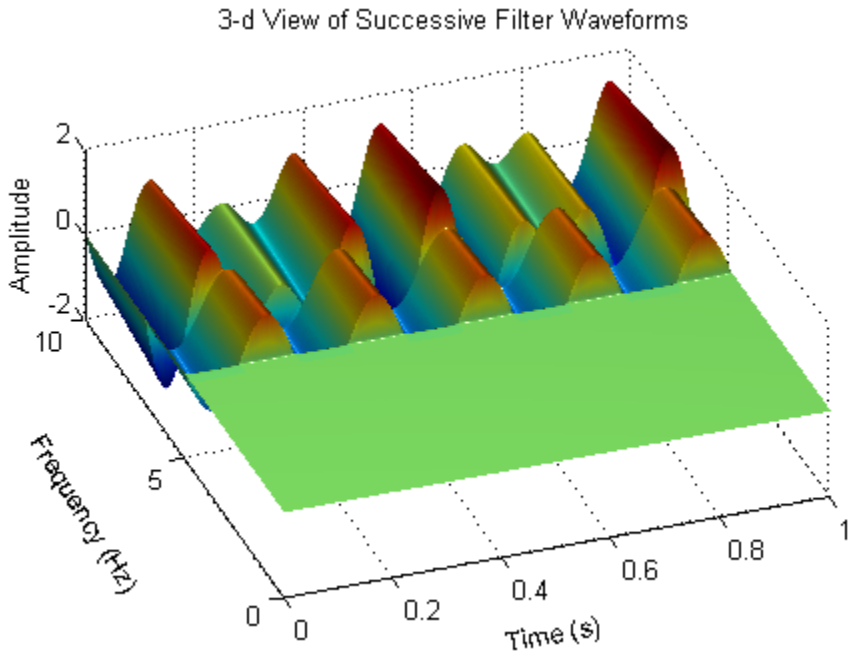

Figure 59. 3-D view of same signals as in Figure 57, but now with upper bandwidth of all filters tightened to 0.01 Hz. Note the sharp demarcations between the regimes 0-5 Hz, 5-7 Hz, and 7-10 Hz.

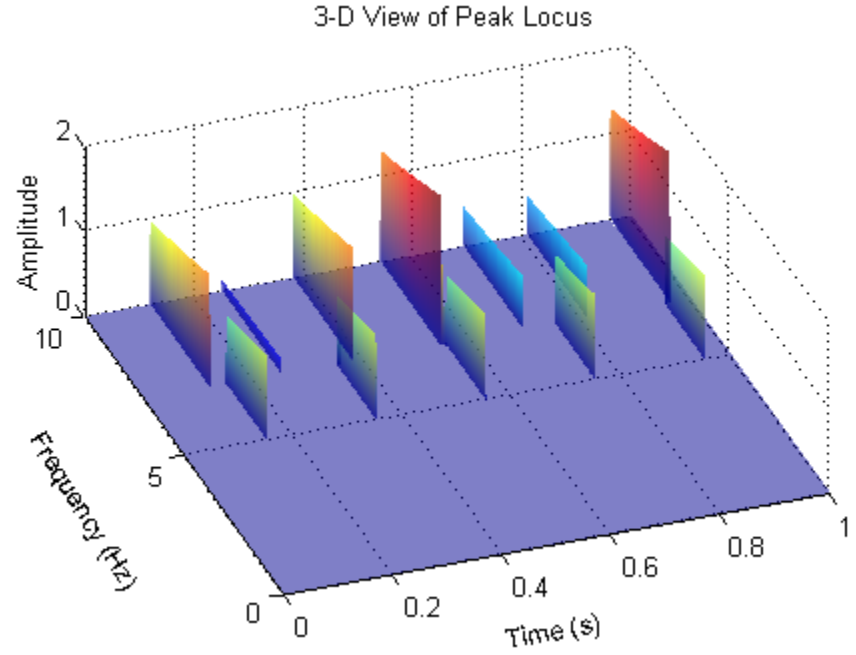

Figure 60. The Peak-Locus of the waveform plot from Figure 56. Note straightening of the 5-7 Hz transition region as compared to Figure 55 due to use of sharper upper cutoff. The 5-7 Hz region now has negligible contribution from higher frequencies, so regularly spaced and equal height peaks are seen, reflecting the monochromatic 5 Hz energy, exclusively. From 7-10 Hz there is presence of both 5 Hz and 7 Hz energy, so unevenly spaced and unequal height peaks are seen.

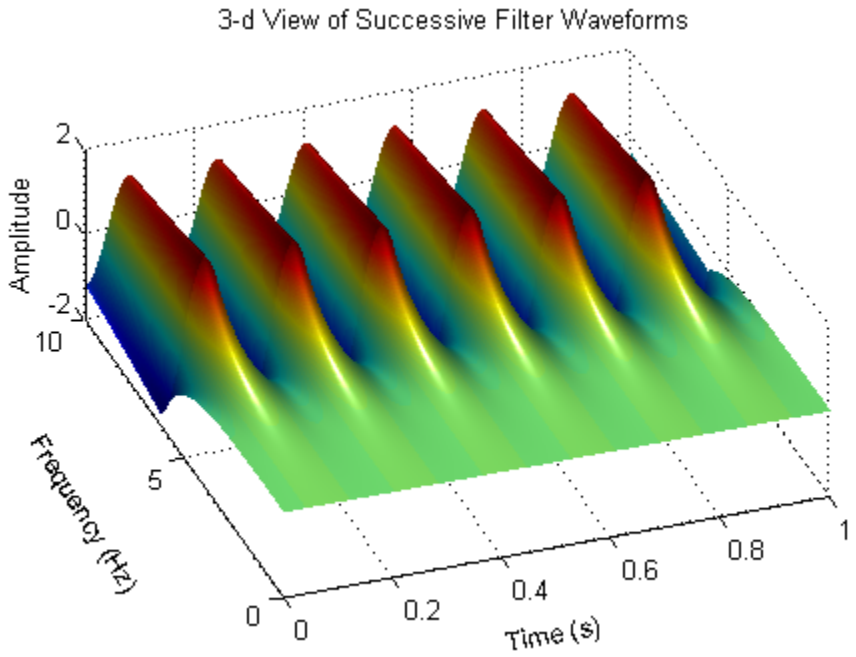

Figure 61. 3-D plot of the signal at the output of each filter in the filter bank. Input was monocomponent equivalent 1.74cos(2π 6.18t+π). Exponential filters had lower bandwidth of 10 Hz and upper bandwidth of 1 Hz. Note parallel structure throughout, thus readily distinguishing from multicomponent signal of Figure 57.

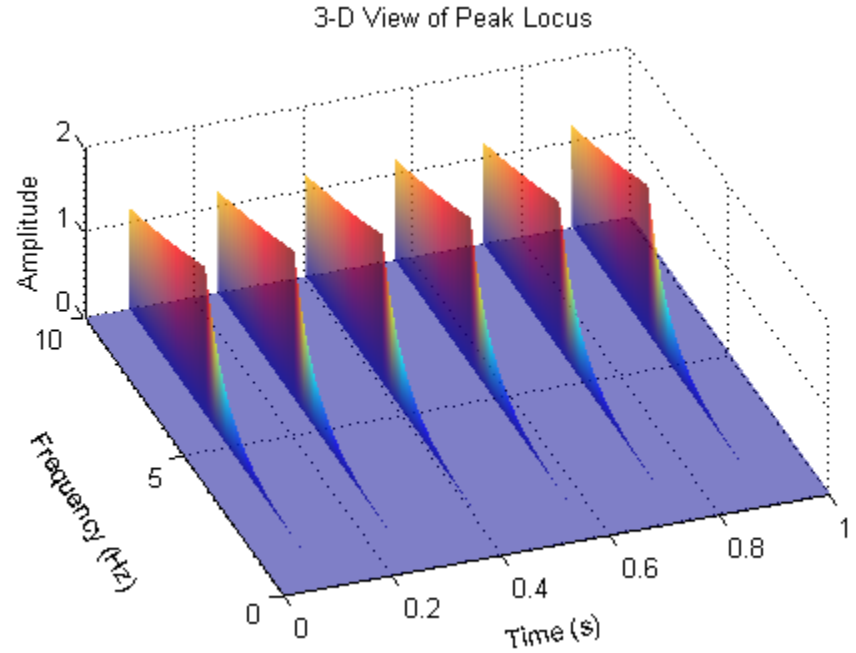

Figure 62. The Peak-Locus of the monocomponent equivalent from Figure 61. Parallel structure readily distinguishes from the multicomponent signal of Figure 58, even with use of filters with looser upper sidewalls of 1 Hz bandwidth, as in Figure 57 and Figure 58, and does not require the use of the tighter sidewall of 0.01 Hz as in Figure 59 and Figure 60 to make difference apparent. In other situations one may prefer a tighter sidewall, although the longer duration of such a filter may have other undesirable properties, such as slower response, and longer settling time for transients. These effects are not noticeable in these figures, since mathematically weighted simulations of exponential responses were used here for clarity, and not actual filters.

5.6.2 Algorithm

The previous discussion can be translated into the following algorithm for separating components:

1) Pass component mixture x_{orig} through an asymmetric exponential filter bank such as we described. Use a sharp upper cutoff and a much wider lower cutoff.

2) Estimate amplitudes, frequencies and phases of each band using peaks, as before.

3) Compute a normalization factor for each band, i.e., evaluate $H(\hat{f})$ of each band at the estimated frequency. Normalize each band.

4) Starting at filter with lowest CF and moving upwards, one should see a string of filters with the same estimated frequencies. Look for the point at which estimated frequency changes. Use the estimate obtained from the last filter in the string (properly normalized) as the estimate of the first spectral component in the mixture $\hat{s}_1 = \hat{a}_{1norm} \sin(2\pi \hat{f}_1 t + \hat{\phi}_1)$.

5) Subtract estimate from the original signal to get residual $R_1 = x_{orig} - \hat{s}_1$.

6) Pass residual R_1 through filter bank.

7) Repeat process to get estimate of second component $\hat{s}_2$. A new string of filters will be expected to be seen with a new common estimated frequency which is now higher than the one in the first iteration.

8) Subtract second estimate to obtain new residual $R_2 = x_{orig} - \hat{s}_1 - \hat{s}_2$.

9) Repeat until all components have been found and subtracted such that the residual is now negligible.

5.6.3 Discussion

1) The search for strings of filters with common frequency estimates can be performed by coincidence detection. This has an extremely plausible biological mechanism.

Since nerve fibers fire preferentially near signal maxima because of phase-locking property as in Figure 1 and Figure 2, some portion of neural processing for source separation could involve detection of coincident spikes, a computational task believed suitable for neural architecture. In addition, asymmetric exponential-shaped filters look not unlike plots of actual frequency response curves from auditory neurons, as in Figure 63. The gradually sloping left edge looks like the tail of neural response curves. These resemblances potentially make it a viable model for understanding actual auditory neural processing.

2) One can't simply use the estimated frequencies of those filters which follow each break in continuity to obtain estimates of all frequencies in one iteration, because of the reason we described earlier—estimates based on peaks are corrupted by the presence of other components. Therefore one starts with lowest CF's, which do not contain the presence of any higher-frequency components due to the cutoff from the sharp upper sidewall. These can be estimated very accurately, and then subtracted out. Each iteration will then be estimating peaks of purified components only.

3) One gets the best frequency estimate at the filter whose CF is just above the frequency of the lowest spectral component found in each pass. This has the best signal-to-noise ratio. We do not want to use an estimate from a filter whose CF is below the frequency of a spectral component, since any slight error in frequency estimate will then cause a huge error in the corresponding amplitude estimate due to the sharpness of the upper filter wall.

4) There are many possibilities for computing amplitude estimates. One can use only the last filter in a string. One can use the average of all the filters in a string suitably normalized. One can use the filter whose CF lies just above the average estimated frequency. Each has advantages and disadvantages in terms of stability.

5) There is much redundancy in the fact that many bands will give identical estimates. This could possibly be exploited to improve accuracy in the presence of noise. In real-world situations we would expect slight shifts in peak locations instead of exact

coincidence. But this variation might be averaged out to form a very accurate estimate of each spectral component.

6) As compared with the previous algorithm (Iterative-Subtraction), instead of subtracting estimates of each band, we now only subtract one representative estimate of a given spectral component; i.e., only one estimate is gleaned from each string of common frequency bands. This greatly speeds up the process, and avoids error buildup that could plague the Iterative-Subtraction algorithm.

7) On the negative side, our frequency resolution becomes strongly dependent on the filter spacing which must be close enough so that we will not admit more than one signal into any filter. This is certainly less elegant than the previous approach which was able to "purify" successive estimates even if originally there were more than one component present within a single filter. In addition, because processing is done in a single pass (one iteration per component), the performance of the algorithm depends critically on perfect exponential behavior of the filters. In practice, it is difficult to design such filters, and this causes some curvature of Peak-Locus lines. Because of this, setting a threshold becomes difficult, as making it too tight will cause minor deviations to be interpreted as an additional frequency component, while making it too loose may miss a genuine component. Therefore, the algorithm works well in the case of mathematically simulated exponential filters (unknown weighted case), but is difficult to implement with actual exponential filters designed with IFFT procedures. We will have more to say about filter design in Section 5.8.

8) The strength of the algorithm lies in its pedagogical ability to illustrate how one can collapse multiple filters into those few that actually carry information. This is an example of redundancy reduction which was enumerated as a key goal of ASA systems by a number of researchers cited in Chapter 2. The exponential filter bank acts somewhat like a lens in focusing the peaks of bands containing the same mixture into the same time point.

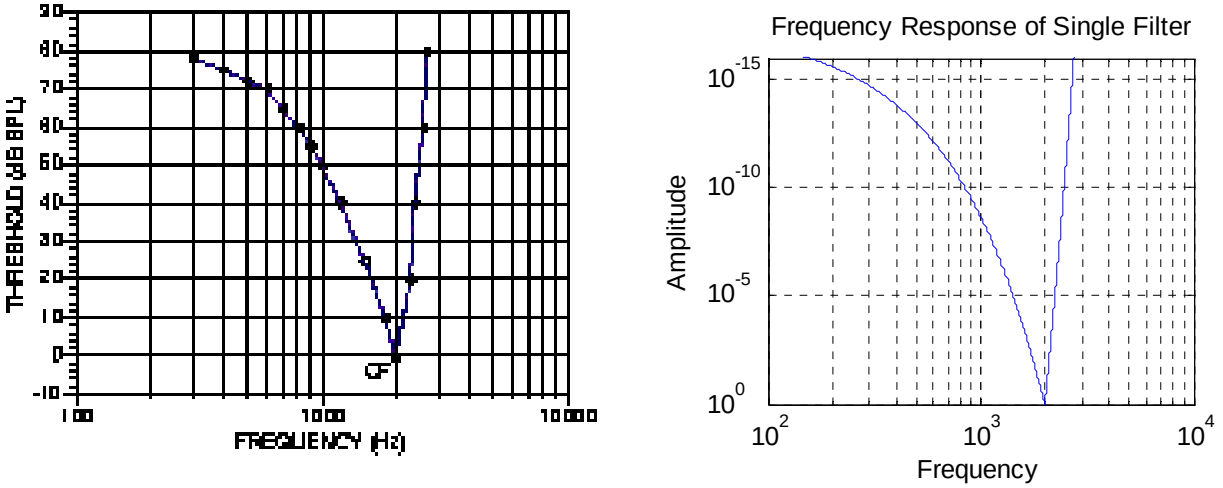

Figure 63. Left: A plot of actual frequency response data from an auditory nerve fiber. Plot is inverted, since it measures threshold of hearing, hence lowest points represent frequencies of maximum sensitivity. From (Brugge, 1996). Right: A log-log plot of frequency response of typical exponential filter illustrating the similar shape of the exponential model to the actual neural response. The lower bandwidth is 50, upper bandwidth is 20, and the CF is 2000 Hz. The filter on the right is narrower, so more orders of magnitude are shown to illustrate general resemblance of response shape.

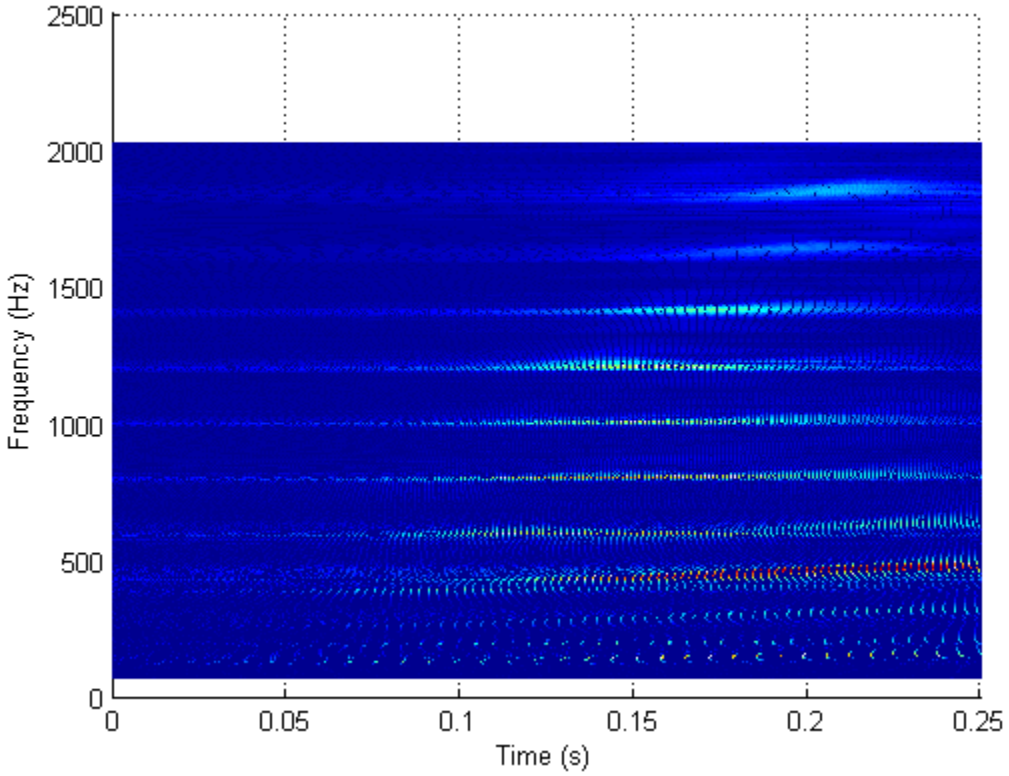

Figure 64. The unprocessed Peak-Locus plot of mixture of male and female waveforms from figure 37 for the first 2000 Hz.

Figure 64 shows raw Peak-Locus data from the speech mixture whose spectrogram was shown at the bottom of Figure 27. Filter bank consisted of overlapping exponential filters of lower

bandwidth 10 and upper bandwidth 1, with spacing of 2 Hz. Sampling rate was 10 KHz. Band data shown is without processing bands through algorithm. Nevertheless, greater detail seems to enable visual resolution of harmonic tracks that appear to overlap in the conventional spectrogram of Figure 27. This figure will be discussed further in Section 6.8.

5.7 Simultaneous-Equation Approach

We describe a third algorithm for separation of sinusoids which is more robust, and has some of the combined advantages of the previous two. This method uses a matrix formulation of simultaneous equations. We again use exponential filter banks for their desirable scaling properties.

The method is as follows for the case of 2 sinusoids. Assume the frequencies of the 2 sinusoidal components are known. If one knows the filter gain and phase characteristics $H_1(f)$, $H_2(f)$ as a function of frequency, one can then set up a set of simultaneous equations as in (5.16) relating the amplitudes and frequencies.

There are two caveats to note. First, this formulation will strictly hold true only in the steady state after transients have settled. However, in Section 6.8 we will offer some additional insights on the issue of transients that became apparent in the course of our work, and find that they are possibly less bothersome than one would expect. Second, there is an additional caveat on the use of these equations in the presence of amplitude or frequency modulation. (Bovik, Havlicek and Desai, 1993) have studied the validity of the approximation that the Fourier transform of the output can be taken to be the product of the transforms of the input and the filter when the input is not a pure sinusoid, but rather consists of amplitude and frequency terms that are functions of time *t*. They provide a bound on the error in such cases. Here, too, as we will explain in Chapter 6, this problem is possibly less of a concern than it would appear. We therefore put aside these issues for now, and continue with the derivation.

Let x_1 and x_2 be the filter outputs in two bands, and let s_1 and s_2 be the original sinusoids.

We then have, by appropriate modification of Equations 5.4 for the filtered case:

(5.16)
$$\begin{aligned} x_1 &= H_1(f_1)\, s_1(t) + H_1(f_2)\, s_2(t) \\ x_2 &= H_2(f_1)\, s_1(t) + H_2(f_2)\, s_2(t) \end{aligned}$$

We can solve this set for s_1 and s_2 since it is a linear system and the functions $H_i(f)$ are known. It can easily be extended to more than 2 sinusoids.

However, clearly, the method depends on knowing f_1 and f_2 *a priori*. If they are unknown, then the weighting of the respective filters $H_1(f)$, $H_2(f)$ which are the coefficients of s_1 and s_2 are in turn unknown, and we can't solve. This is a serious drawback in the real world, as one does not in fact have *a priori* information on the frequencies of the individual signals at every time.

We therefore propose to once again use temporal information from local maxima in an attempt to fill in the missing information. We proceed as follows:

The input sines are:

(5.17)
$$\begin{aligned} s_1 &= a_1 \sin(2\pi f_1 t + p_1) \\ s_2 &= a_2 \sin(2\pi f_2 t + p_2) \end{aligned}$$

The band outputs are:

(5.18)
$$\begin{aligned} x_1 &= w_{11}(f_1)\, s_1 + w_{12}(f_2)\, s_2 \\ x_2 &= w_{21}(f_1)\, s_1 + w_{22}(f_2)\, s_2 \end{aligned}$$

where w_{ij} = weight of filter *i* on sinusoid *j*. (We assume zero-phase filters, so that weights can be represented with real coefficients and need not be complex.)

Let us assume that the filters have known characteristics including CF's, bandwidths and gains. (We use slightly different notation here for the filter characteristics w_{ij} (weight) instead of $H_i(f_j)$ to conform to more standard notation for systems of linear equations.)

We now rewrite in a more compact matrix formulation. We use Matlab® colon notation for an index which increments in steps of 1.

Let $s_j \quad j = 1:n$ be a set of sines with arbitrary frequencies, amplitudes and phases.

Let $x_i \quad i = 1:m$ be the set of band outputs.

We can write as $\mathbf{X} = \mathbf{WS}$ where $\mathbf{W(f)}$ is an $n \times m$ mixing matrix, and we drop the *f*-dependence for brevity.

We are given $\mathbf{X}$. If we knew $\mathbf{W}$, we could simply invert in a least-squares sense, and determine $\mathbf{S}$. But $\mathbf{W}$ depends on input frequencies which are initially unknown. We thus need to find 2 unknown matrices, $\mathbf{W}$ and $\mathbf{S}$, given only $\mathbf{X}$.

We use the peaks of the waveform within each band to determine the frequency of the signal in that band. We then calculate the filter weight for that frequency and determine the elements of $\mathbf{W}$. The problem is that peaks are only reliable estimators for monochromatic signals. When dealing with a mixture, we have seen that they are often a compromise between the peaks of each signal alone, but in some cases, certain pairs of adjacent peaks of the mixture may be closer together than those of the highest-frequency component , or farther away than those of the lowest-frequency component, giving misleading results. Nevertheless, through successive iterations the estimates converge towards the correct frequencies.

5.7.1 Determining Number of Sources

In the equation $\mathbf{X} = \mathbf{WS}$, the dimensions of $\mathbf{X}$ are known.

1) The number of rows is the number of auditory filters.

2) The number of columns is the length of the recording.

Methods of Linear Algebra dictate that:

1) Number of rows of $\mathbf{W}$ equals number of rows of $\mathbf{X}$. (Number of Filters.)

2) Number of columns of $\mathbf{S}$ equals number of columns of $\mathbf{X}$. (Recording length.)

However, the number of columns of $\mathbf{W}$ and rows of $\mathbf{S}$ (which must be the same number) cannot be determined initially. In keeping with our notation from Chapter 4, we will denote this unknown by *r*. This number represents the number of sources. For example if we use 10 filters, and our source is sampled at 10 KHz, and we have 1 second of recording time, we would then have 10,000 samples. The size of $\mathbf{X}$ would then be $(10 \times 10{,}000)$. This would mean that the size of W has to be $(10 \times r)$ and the size of $\mathbf{S}$ must be $(r \times 10{,}000)$.

In order to determine r, we use two thresholds in the course of iterating. The first is the duplication threshold, whereby frequency estimates that differ by less than this amount are considered to be identical. This is well understood from our discussion of the Peak-Locus method, as we expect strings of identical frequency estimates, unless an additional component has entered into the picture. Conceptually, this corresponds to eliminating duplicate bands in Gaussian elimination.

The second is an amplitude threshold, whereby bands with amplitude estimates less than this value are eliminated. This corresponds to eliminating a row of zeros in Gaussian Elimination. In other words, when a row of a matrix can be expressed as a linear combination of 2 other bands, a row of zeros is produced, and we are being told in effect that no new information is present in that row. Similarly in our problem, if an amplitude estimate of a band falls so low in a particular iteration that it is negligible, we are being told that there is no new information on any frequency components that are not already accounted for in some other bands.

5.7.2 Algorithm

Based on the above formulation, we propose the following iterative algorithm.

1) Pass input signal through exponential filter bank. This will cause peaks of redundant bands to line up, as in the Peak-Locus method.

2) Set rank *r* initially to be the number of channels.

3) Use peaks of x_i waveforms to form initial estimates of amplitude $\hat{a}_j$, frequency $\hat{f}_j$, and phase $\hat{\phi}_j$, ignoring non-monochromaticity.

4) Consolidate those bands whose frequencies $\hat{f}_j$ differ by less than a predetermined duplication threshold, and eliminate those bands whose amplitudes $\hat{a}_j$ are less than a predetermined amplitude threshold. Adjust rank r and sizes of $\hat{\mathbf{S}}$ and $\hat{\mathbf{W}}$ accordingly.

5) From frequency estimates $\hat{f}_j$, calculate $\hat{w}_{ij}$ terms using $\hat{w}_{ij} = H_i(\hat{f}_j)$.

6) Divide $\mathbf{X}$ by pseudoinverse of $\hat{\mathbf{W}}$ to obtain $\hat{\mathbf{S}}$ whose rows are current estimates $\hat{s}_j$ of input sinusoids.

7) Use peaks of current $\hat{s}_j$ estimates to get better estimates of $\hat{f}_j$.

8) Repeat from step 4.

5.7.3 Discussion

1) The general idea has some similarity to the work of (Quatieri and Danisewicz, 1990), but whereas their method required *a priori* frequency information, this method does not. In addition, this method requires no specific window length, as do FFT frame based methods, since it computes estimates from only a single pair of peaks in each band. This potentially gives it much better time resolution. In addition, it requires no minimum frequency separation, and good results have been seen down to 0.5 Hz or better in mathematically weighted filter simulation tests, and to 1 Hz in actual filtered tests. We compare results of our various algorithms at the end of this chapter.

2) This approach is less biologically plausible, and relies on matrix methods which are designed for linear equations, but which we use for an essentially nonlinear problem. As such, the mechanism by which it works is not fully understood. Proving convergence for iterative algorithms is a notoriously difficult task. Nevertheless, the use of exponential filters to eliminate duplicate information from bands with similar estimates is similar to the Peak-Locus method which is better understood.

3) The fact that the algorithm is multi-pass allows it to refine estimates at each iteration, and is thus more immune to slight filter imperfections. We have attained valid results even with other filter types than exponential, as the idea of mapping band estimates to sources in a strict sense requires only the assumption of linear independence of the transformation matrix $\mathbf{W}$.

4) Additional *a priori* information can be incorporated, if available, such as a known number of sources, or a known harmonic relationship among frequency components, with modifications, but won't be pursued here.

5) This algorithm has the possibility of incorporating an internal consistency check, as the matrix inversion operation (on the matrix $\mathbf{W}$) depends only on frequency estimates. Hence, band amplitudes can be calculated either by local maxima methods, as usual, or by additionally calculating $H(\hat{f})$ for the frequency estimate of each band. These should match. We have not pursued this further.

6) One drawback is that the algorithm eliminates unneeded bands in the course of iterations, but has no mechanism for adding back new bands. This might be useful in improving resolution where later iterations may find a need to further subdivide a frequency estimate into separate components.

7) Overall, results with this approach even in the filtered case have been very encouraging and we have therefore adopted this algorithm as our workhorse, and will use it exclusively for tests on time-varying signals in Chapter 6.

5.8 Filter Design Considerations

As we have mentioned a number of times previously, there are various tradeoffs to be considered when designing a suitable filter for use with our algorithms. The general step-by-step descriptions we have given for each algorithm primarily focused on the numerical calculations that need to be done with band data, and did not for the most part specify how that data was obtained, whether through simulated mathematical weightings or via actual filters. We now discuss some considerations for actual filter design in more detail, since if the algorithms are to have actual utility rather than being mere mathematical curiosities, they must be able to work with actual signals and filters.

5.8.1 Rectangular Filter Design: A Motivating Case

As we have noted earlier, there are two requirements for an accurate analysis of modulation, which has been the primary motivation of all of our work. The first consideration is that filters

should have fast response times, as we want the output envelopes to reflect the amplitude modulation patterns of the source, rather than the rise times of the filters. The second is that we need accurate frequency resolution, as we want to track frequency modulation patterns. It is well known that these considerations work against each other, and that compactness in one domain causes dilation in the other domain. It therefore begs the question whether a filter bank can be designed with both properties.

At the outset of our thesis work, and a major thrust of all that we have done was to pursue the dogging question of whether using multiple filters could improve the resolution obtainable by a single filter alone. It initially occurred to us that the use of two very wide square filters which overlap everywhere except for a small region on either side might be a way to obtain both advantages. The filters themselves would have good time resolution due to their short impulse response lengths, but by subtracting two wide frequency responses one from another, possibly we could attain good frequency resolution, as well, as only components in the narrow non-overlapping frequency region would be passed through. The situation is as illustrated in Figure 65.

We constructed some square filters using common polynomial designs, but found that performance was poor. Rise times were extremely long, and due to this prolonged response, amplitude modulation was difficult to detect. We initially concluded that the reason for this was that the filter had a nonzero phase response. The response of square filters centered on two different frequencies to a signal falling within the common passband is not identical. The phase of one may be different than the phase of the other, and hence when one subtracts, one may in fact be adding, as there could be a sign change in the waveform.

However, the question then arises that what if perfect zero-phase filters can be constructed? Would this then yield the improved time response that we seek? We must again answer no. The reason is that Fourier transforms are linear. The time response to a difference signal formed by subtracting the output of one filter from the output of another filter is the same as the time response of the output signal of a filter whose frequency response is the difference of the frequency responses of the two filters. That difference is simply the exact equivalent of a very narrow square filter which perforce must have the poor time-domain characteristics we are trying to avoid.

Mathematically, the situation is

$$
\begin{aligned}
h_{combined}(f) &= h_1(f) - h_2(f) \\
&= \Im^{-1}[H_1(f)] - \Im^{-1}[(H_2(f)] \\
&= \Im^{-1}[H_1(f) - H_2(f)] \\
&= \Im^{-1}[H_{narrow}(f)]
\end{aligned}
\tag{5.19}
$$

by linearity. The time response of the overlapped portion of the combined filter is no better than a narrow filter to begin with.

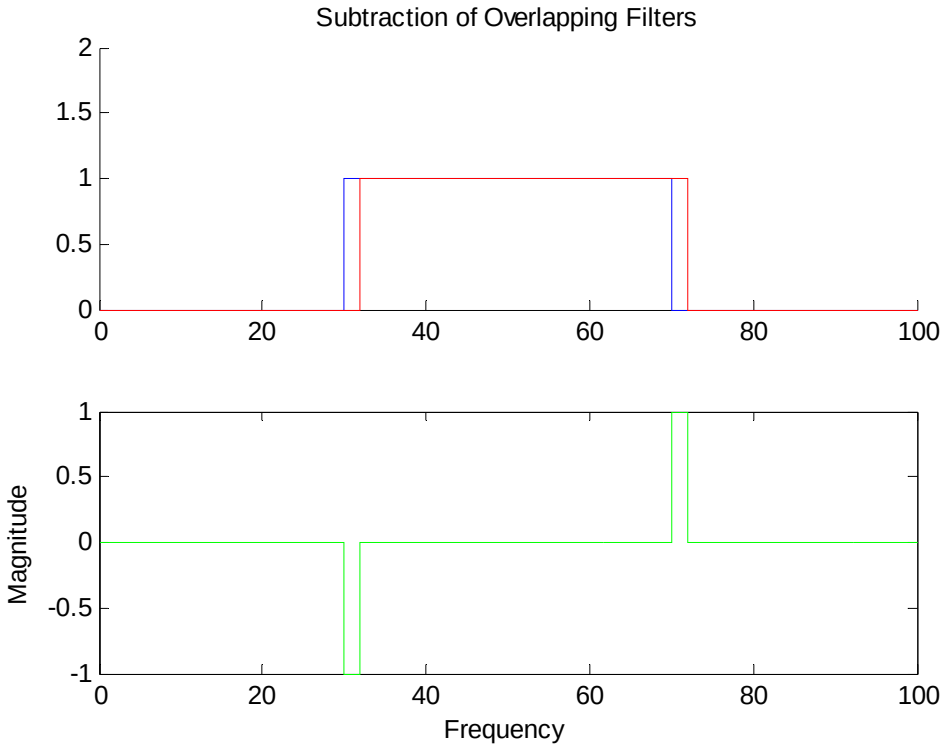

Figure 65. Subtraction of response of blue filter from red filter in top plot is equivalent to response of green area in bottom plot which is equivalent to two very narrow filters with opposite phase.

We present an example of an actual attempt to use such a filter subtraction scheme on the speech mixture of Figure 27. In Figure 66, we passed the mixture through an exponential filter bank, and plotted the resulting waveforms in 3-D. (We note that these waveforms are the basis of the Peak-Locus plot of Figure 64, but are rotated 90 degrees here for better clarity along the time axis.) In Figure 67, for each filter, we subtract the output waveform of the next lowest neighboring filter from its own output waveform, and again plot in 3-D. Comparison of Figure 66 with Figure 67 illustrates complete loss of time-variation by using such a successive filter subtraction scheme. Amplitude and frequency appear constant.

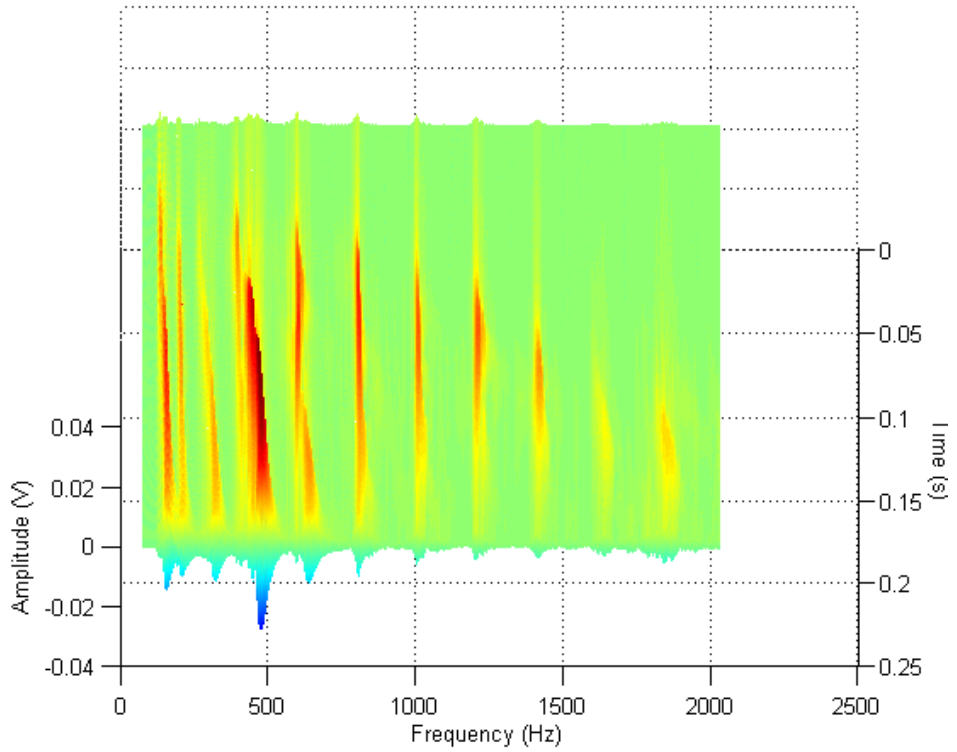

Figure 66. 3-D elevated mesh view from the right side of male and female speech mixture from Figure 27 after passing through exponential filter bank to better illustrate amplitude and frequency variations.

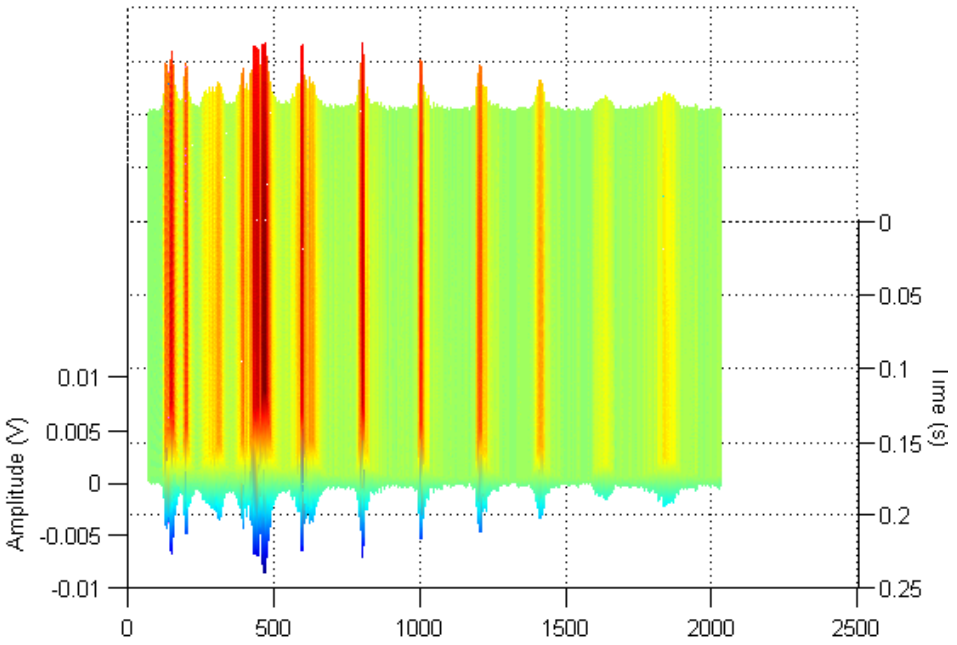

Figure 67. An attempt to sharpen frequency resolution of Figure 66 by subtracting neighboring overlapping exponential filters one from the other. All time dependence is lost for reasons explained in text.

Even if we are willing to tolerate poor time response, another problem with this scheme is the requirement that enough filters be used so that no two signals fall within the filter separation distance. Whereas in our single-pass Peak-Locus method that generally must be true, as well, however, we have successfully run trials with our other methods in which we were able to resolve signals that were spaced closer than the filter separation distance, as the estimates improve over the course of multiple iterations.

While this early attempt at using subtraction was not successful, it became the forerunner of the Iterative-Subtraction method. The major difference is that the Iterative-Subtraction algorithm forms an estimate at each step and subtracts information gleaned from each band to refine the estimate, rather than merely subtracting raw band outputs alone as done here.

An important consideration which is relevant for all of our three methods, is that they rely on the fact that the filter shapes are not flat, and hence weight signals differently with respect to each other. Flatness is a disadvantage for our methods, and would cause the equations to become singular or ill conditioned. The one-pass Peak-Locus method depends critically on the filter profile being as close as possible to a true exponential shape in order for peaks in various bands to line up. The others are somewhat more robust in being able to compensate for imperfections.

We have not obtained good results using Hann windows to taper the time response of our filters, as they tend to distort the exponential frequency response shape, and occasionally alter the characteristic concave shape to convex. In addition, they cause as much as a 10% loss of magnitude in filter frequency response from the expected unity gain at CF.

5.8.2 Filter Bank and Signal Specification

We ran an initial series of comparison tests between the algorithms on constant-amplitude, constant-frequency harmonic sets which are described in this chapter (Section 5.9). After finding that best results with actual filters were obtained with the Simultaneous-Equation method, we ran an advanced set of performance tests on more realistic signals using that algorithm alone, the results of which are described in Chapter 6. We do not imply that the other algorithms cannot achieve the same results, but that the effort to continue to tweak and search for the best combinations of parameters was becoming too time-consuming to fully explore.

	Comparison Tests (Chapter 5)	Performance Tests (Chapter 6)
Interpolation	No	Yes
Sampling Rate	100/500 KHz	4 MHz
Upper BW	1	2
Lower BW	10	50
Frequencies	Integer	Non-Integer
Signals	Unmodulated	Modulated (AM &FM), Noise, Speech
Algorithms	All	Simultaneous Eqs.
Cases	Weighted/Filtered	Filtered
Settling Margin	Half Filter-Length	Full Filter-Length

Table 9. Characteristics of the various test sets that were conducted on the algorithms.

We note that the filtered tests were all run in the steady-state region, but that the results in this chapter (Section 5.9) eliminated a half filter-width border on either side of the output response, whereas in the next chapter, in tests on modulated signals, we eliminated a stricter full filter-width border from the output response to ensure that all transients had completely subsided before any analysis was performed.

While we have avoided the transient region, we mention that it may be possible to apply the same algorithm even for shorter signal lengths that do not reach steady state. The way that could work is that any time a particular frequency value is estimated within a band, a calibrating sinusoid could be generated at the same frequency with an amplitude of unity, and it's actual output value measured at the same time point as in the signal. This gives the actual effect or weight of the filter in question on the given signal, even though it is not in the steady-state region. This value can then be used in the matrix inversion. Such a procedure is likely to give better results in the steady-state region, as well, since the current method calculates the true mathematical exponential as an approximation to $H(f)$ at each step, although we have measured deviations from ideality of a few percent. Using the value that is actually measured at

the output of the given filter to a known input is likely to improve results, but comes at an additional computational cost.

In the tests of Section 5.9, we used uninterpolated sampling rates of up to approximately 100 KHz in filtered tests, and 500 KHz in weighted tests with the memory resources we had available, with the difference being that weighted tests do not need any settling margin and hence can be shorter. We seemed to obtain best results with tighter filters, using lower BW of 10 based on initial trials. Our later performance tests of Chapter 6 were run with initial sampling rates of 10KHz, but used low-pass interpolation on a single band at a time during analysis to effectively increase sampling rate to 4 MHz. Since storage of all other channel data remained at 10KHz, the cost in memory was kept manageable. However, interpolation adds a potential additional layer of noise that we were never able to totally rid ourselves of, and would sporadically produce estimates that were completely out of line with expectations and with neighboring estimates, and were clearly due to interpolation noise. (The noise introduces false high frequency maxima.) We tried a number of interpolation filter lengths and cutoff frequencies that the Matlab® `interp` function allows, but always found occasional evidence of noise. This was not observed in weighted tests at uninterpolated sampling rates.

For the interpolated tests, trial and error determined that satisfactory performance occurred with looser filters of lower BW 50 and upper BW 2. It was found that using an upper bandwidth of 1 gave improved results in weighted case, but poorer results in filtered case apparently due to increased ripple, which probably stemmed from the lesser natural tapering of the impulse response caused by the sharper upper sidewall. However, widening the upper wall still further hurt resolution even in the weighted case, and so was unlikely to work in the filtered case.

The difference between the necessary filter bandwidths in the interpolated vs. uninterpolated cases may be due to the fact that the broader the filter, the more subtle the peak changes become from channel to channel, and hence require a concomitant greater temporal sensitivity to detect. This was seen in the comparison of Figure 58 to Figure 60. Thus, greater sampling rates may be required to precisely locate these points in time. We note that further tests should be conducted with other parameter combinations to see if better filters can possibly be found. We note that actual auditory filter data in Figure 63 seems to indicate the use of far wider filters than we have tested, and hence there is room for trials with additional sets of parameter combinations.

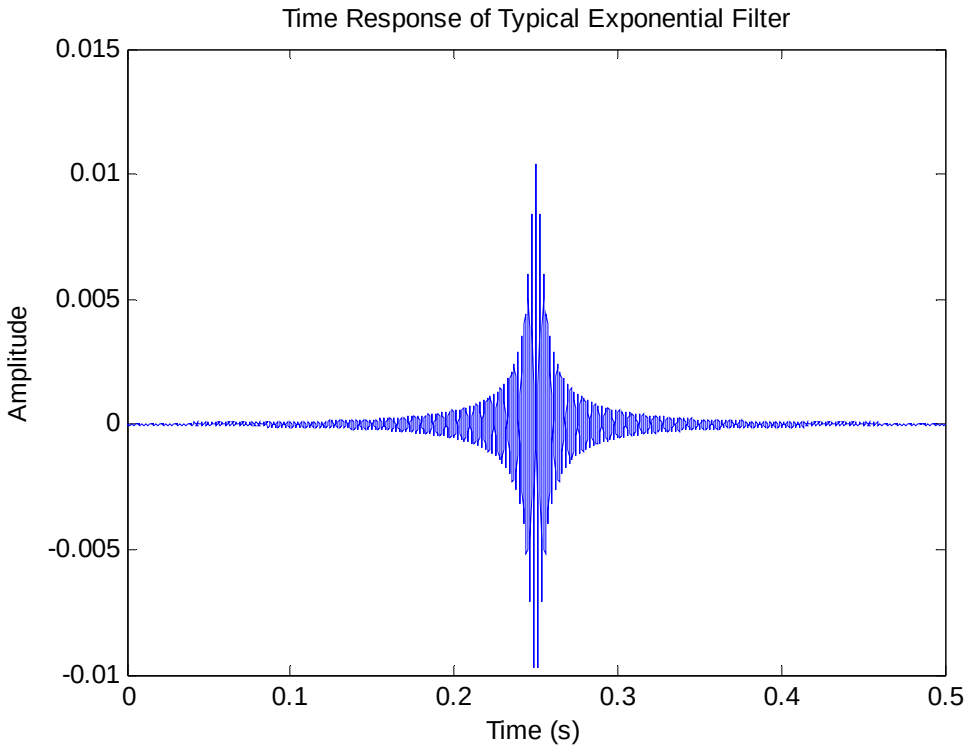

Figure 68. The impulse response of the filter design which gave best results. At 10 KHz sampling rate, duration is 0.5 seconds, but the bulk of the response in concentrated between 0.2 and 0.3 seconds. Tapered shape makes window unnecessary. CF for this filter is 475 Hz.

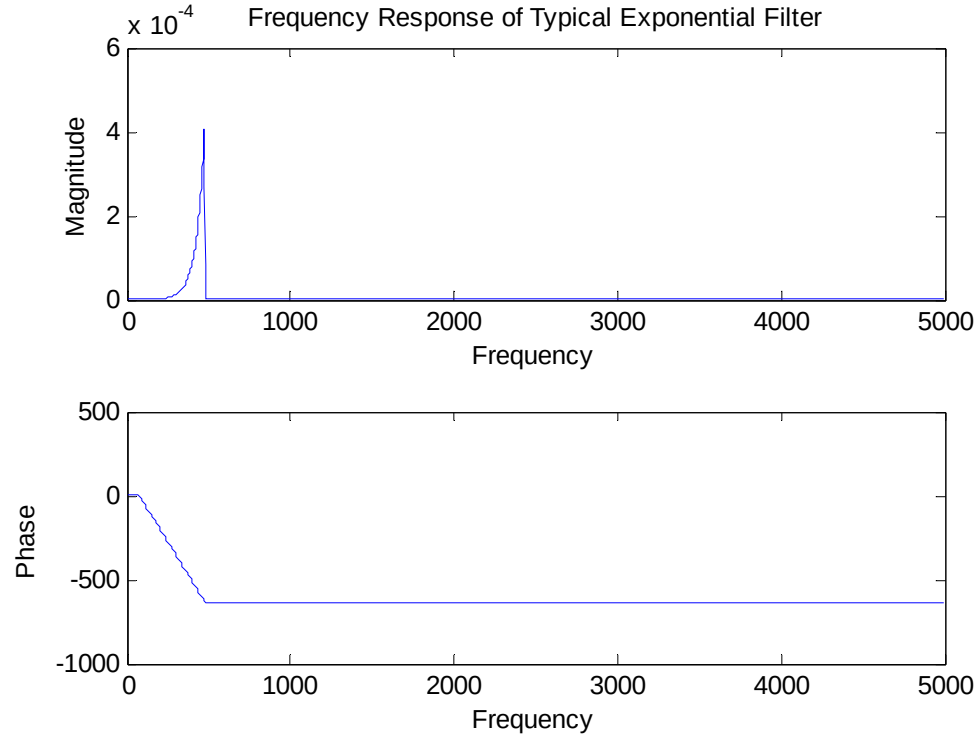

Figure 69. Frequency response of filter of Figure 68. CF was 475 Hz. Phase response was linear, and is left-shifted in practice by a half filter-length to give zero phase response.

For the interpolated tests, the time-response of the filter which gave best results is illustrated in Figure 68, and was truncated to 0.5 seconds. Frequency response is illustrated in Figure 69. The comparative list of parameters for each of the two test suites is shown in Table 9.

5.9 Comparisons and Evaluations

We now compare numerical results produced with each of the three algorithms we have developed. Due to memory limitations, we were constrained to work with the first 0.5 KHz of the spectrum for the higher sampling rates that we customarily used. Weighted tests were conducted at 500 KHz rate, while filtered tests were conducted at 100 KHz. The difference is due to the different signal lengths required for each. Amplitude thresholds were 0.1 and filter spacing was 1 Hz . Lower bandwidth was 10 and upper bandwidth was 1, as in Table 9.

We performed tests on the following types of signals with no added noise.

1) Mixtures of two constant-frequency, constant-amplitude harmonic sets using simulated filters with ideal mathematical weighting.

2) Mixtures of two constant-frequency, constant-amplitude harmonic sets using actual digital filters in the steady-state region.

The tests were repeated for three values of frequency difference (Delta) between the fundamentals of each set:

1) 10 Hz

2) 5 Hz

3) 1 Hz

The fundamental of the first harmonic set was kept at 100 Hz, while that of the second was thus tested at 110, 105 and 101 Hz. We were able to analyze the first 4 harmonics of each set. The phases were 0.5, 1.0, 1.5 and 2.0 for the first set, and -0.5, -1.0, -1.5, and -2.0 for the second set.

We note that the test sets were all conducted with integer frequencies, and hence the values were all located exactly on filter CF's. In order to determine whether nonintegral frequencies

would give correct results, at the beginning of Chapter 6 we repeated the Delta-1-Hz test at fractional frequencies, with results displayed in Table 22. (However, the filter parameters used for tests in Chapter 6 were completely different, as noted in Table 9, hence it was not a true head-to-head comparison, but rather a general gauge of accuracy.) We point out that it was important to verify performance on inter-filter frequencies, since the FFT itself is notorious for giving inaccurate results with frequencies that do not fall exactly on bin centers. One goal of our methods is to develop techniques for frequency estimation which are less dependent on particular window lengths and are not restricted to discrete frequency bins, in keeping with the spirit of instantaneous frequency analysis. We discuss these results in Chapter 6.

Following are the tabulated results for the suite of comparison tests:

5.9.1 Weighted case: Iterative-Subtraction Method

Iterative-Subtraction: Delta 10 Hz.

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
100.00	100.00	0.00	3.00	3.00	0.00	0.50	0.50	0.00
110.00	109.99	0.01	2.00	2.00	0.00	-0.50	-0.48	0.67
200.00	200.00	0.00	2.00	2.00	0.00	1.00	1.00	0.01
220.00	219.97	0.01	1.00	1.00	0.00	-1.00	-0.96	1.32
300.00	299.94	0.02	4.00	4.00	0.00	1.50	1.59	2.97
330.00	330.03	0.01	3.00	3.00	0.00	-1.50	-1.55	1.66
400.00	400.00	0.00	2.00	2.00	0.00	2.00	2.00	0.06
440.00	439.75	0.06	1.00	1.00	0.05	-2.00	-1.61	12.34

Table 10. Results for the weighted case at 10 Hz separation using Iterative-Subtraction algorithm.

Iterative-Subtraction: Delta 5 Hz

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
100.00	100.00	0.00	3.00	3.00	0.00	0.50	0.50	0.00
105.00	105.00	0.00	2.00	2.00	0.00	-0.50	-0.50	0.10
200.00	200.00	0.00	2.00	2.00	0.00	1.00	1.00	0.01
210.00	210.00	0.00	1.00	1.00	0.00	-1.00	-0.99	0.18
300.00	299.94	0.02	4.00	4.00	0.00	1.50	1.59	2.97
315.00	314.86	0.04	3.00	3.00	0.01	-1.50	-1.28	6.94
400.00	400.00	0.00	2.00	2.00	0.00	2.00	2.00	0.06
420.00	419.82	0.04	1.00	1.00	0.00	-2.00	-1.71	9.21

Table 11. Results for the weighted case at 5 Hz separation using Iterative-Subtraction algorithm.

Iterative-Subtraction: Delta 1 Hz

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
100.00	100.00	0.00	3.00	3.00	0.00	0.50	0.50	0.00
101.00	101.01	0.01	2.00	2.00	0.01	-0.50	-0.52	0.51
200.00	200.00	0.00	2.00	2.00	0.00	1.00	1.00	0.01
202.00	202.02	0.01	1.00	1.00	0.00	-1.00	-1.03	1.00
300.00	299.94	0.02	4.00	4.00	0.00	1.50	1.59	2.97
303.00	302.85	0.05	3.00	3.00	0.02	-1.50	-1.26	7.50
400.00	400.00	0.00	2.00	2.00	0.00	2.00	2.00	0.06
404.00	404.20	0.05	1.00	1.00	0.00	-2.00	-2.32	10.19

Table 12. Results for the weighted case at 1 Hz separation using Iterative-Subtraction algorithm.

5.9.2 Weighted case: Simultaneous-Equations Method

Simultaneous-Equations: Delta 10 Hz.

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
100.00	100.00	0.00	3.00	3.00	0.00	0.50	0.50	0.00
110.00	109.99	0.01	2.00	2.00	0.13	-0.50	-0.48	0.67
200.00	200.00	0.00	2.00	2.00	0.00	1.00	1.00	0.01
220.00	219.97	0.01	1.00	1.00	0.26	-1.00	-0.96	1.32
300.00	299.94	0.02	4.00	4.03	0.65	1.50	1.59	2.97
330.00	330.03	0.01	3.00	3.11	3.69	-1.50	-1.55	1.66
400.00	400.00	0.00	2.00	2.00	0.00	2.00	2.00	0.06
440.00	439.75	0.06	1.00	1.02	2.49	-2.00	-1.61	12.34

Table 13. Results for the weighted case at 10 Hz separation using Simultaneous-Equation algorithm.

Simultaneous-Equations: Delta 5 Hz

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
100.00	100.00	0.00	3.00	3.00	0.00	0.50	0.50	0.00
105.00	105.00	0.00	2.00	2.00	0.02	-0.50	-0.50	0.10
200.00	200.00	0.00	2.00	2.00	0.00	1.00	1.00	0.01
210.00	210.00	0.00	1.00	1.00	0.04	-1.00	-0.99	0.18
300.00	299.94	0.02	4.00	4.02	0.60	1.50	1.59	2.97
315.00	314.86	0.04	3.00	3.04	1.40	-1.50	-1.28	6.94
400.00	400.00	0.00	2.00	2.00	0.00	2.00	2.00	0.06
420.00	419.82	0.04	1.00	1.02	1.86	-2.00	-1.71	9.21

Table 14. Results for the weighted case at 5 Hz separation using Simultaneous-Equation algorithm.

Simultaneous-Equations: Delta 1 Hz

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
100.00	100.00	0.00	3.00	3.00	0.00	0.50	0.50	0.00
101.00	100.99	0.01	2.00	2.00	0.11	-0.50	-0.48	0.51
200.00	200.00	0.00	2.00	2.01	0.29	1.00	1.01	0.17
202.00	202.02	0.01	1.00	1.02	1.58	-1.00	-1.03	1.00
300.00	299.94	0.02	4.00	4.02	0.46	1.50	1.61	3.45
303.00	303.03	0.01	3.00	3.06	2.13	-1.50	-1.55	1.53
400.00	400.00	0.00	2.00	1.90	4.90	2.00	2.06	1.86
404.00	404.20	0.05	1.00	1.01	0.99	-2.00	-2.32	10.19

Table 15. Results for the weighted case at 1 Hz separation using Simultaneous-Equation algorithm.

5.9.3 Weighted case: Peak-Locus Method

Peak-Locus: Delta 10 Hz

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
100.00	100.00	0.00	3.00	3.00	0.00	0.50	0.50	0.00
110.00	109.99	0.01	2.00	2.00	0.12	-0.50	-0.48	0.67
200.00	200.00	0.00	2.00	2.00	0.01	1.00	1.00	0.01
220.00	219.97	0.01	1.00	1.00	0.27	-1.00	-0.96	1.32
300.00	299.94	0.02	4.00	4.03	0.66	1.50	1.59	2.97
330.00	330.03	0.01	3.00	2.99	0.33	-1.50	-1.55	1.66
400.00	400.00	0.00	2.00	2.00	0.16	2.00	2.00	0.06
440.00	439.75	0.06	1.00	1.02	2.49	-2.00	-1.61	12.34

Table 16. Results for the weighted case at 10 Hz separation using Peak-Locus algorithm.

Peak-Locus: Delta 5 Hz

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
100.00	100.00	0.00	3.00	3.00	0.00	0.50	0.50	0.00
105.00	105.00	0.00	2.00	2.00	0.02	-0.50	-0.50	0.10
200.00	200.00	0.00	2.00	2.00	0.01	1.00	1.00	0.01
210.00	210.00	0.00	1.00	1.00	0.02	-1.00	-0.99	0.18
300.00	299.94	0.02	4.00	4.02	0.60	1.50	1.59	2.97
315.00	314.86	0.04	3.00	3.04	1.43	-1.50	-1.28	6.94
400.00	400.00	0.00	2.00	2.00	0.01	2.00	2.00	0.06
420.00	419.82	0.04	1.00	1.02	1.90	-2.00	-1.71	9.21

Table 17. Results for the weighted case at 5 Hz separation using Peak-Locus algorithm.

Peak-Locus: Delta 1 Hz

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
100.00	100.00	0.00	3.00	3.00	0.00	0.50	0.50	1.50
101.00	101.01	0.01	2.00	2.00	0.11	-0.50	-0.52	0.51
200.00	200.00	0.00	2.00	2.00	0.00	1.00	1.00	0.01
202.00	202.02	0.01	1.00	1.00	0.24	-1.00	-1.03	1.00
300.00	299.94	0.02	4.00	4.02	0.60	1.50	1.59	2.97
303.00	303.03	0.01	3.00	2.99	0.19	-1.50	-1.55	1.52
400.00	400.00	0.00	2.00	2.00	0.00	2.00	2.00	0.06
404.00	403.88	0.03	1.00	1.01	1.43	-2.00	-1.81	6.07

Table 18. Results for the weighted case at 1 Hz separation using Peak-Locus algorithm.

5.9.4 Filtered Case: Simultaneous-Equations Method

Simultaneous-Equations: Delta 10 Hz

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
100.00	100.00	0.00	3.00	3.00	0.11	0.50	0.50	0.12
110.00	110.01	0.01	2.00	2.00	0.18	-0.50	-0.52	0.75
200.00	200.00	0.00	2.00	2.00	0.18	1.00	0.99	0.23
220.00	219.78	0.10	1.00	1.00	0.43	-1.00	-0.67	10.62
300.00	299.40	0.20	4.00	4.00	0.09	1.50	2.42	29.20
330.00	330.03	0.01	3.00	3.03	1.05	-1.50	-1.58	2.58
400.00	400.00	0.00	2.00	2.00	0.19	2.00	1.97	0.86
440.00	440.53	0.12	1.00	1.00	0.06	-2.00	-2.86	27.31

Table 19. Results for the filtered case at 10 Hz separation using Simultaneous-Equation algorithm.

Simultaneous-Equations: Delta 5 Hz

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
100.00	100.00	0.00	3.00	3.01	0.26	0.50	0.50	0.12
105.00	105.04	0.04	2.00	1.99	0.47	-0.50	-0.57	2.36
200.00	200.00	0.00	2.00	2.00	0.08	1.00	0.99	0.23
210.00	210.08	0.04	1.00	1.00	0.38	-1.00	-1.15	4.72
300.00	300.30	0.10	4.00	4.04	1.08	1.50	0.99	16.37
315.00	314.47	0.17	3.00	3.07	2.20	-1.50	-0.66	26.68
400.00	400.00	0.00	2.00	2.00	0.22	2.00	1.97	0.86
420.00	420.17	0.04	1.00	1.00	0.15	-2.00	-2.28	9.03

Table 20. Results for the filtered case at 5 Hz separation using Simultaneous-Equation algorithm.

Simultaneous-Equations: Delta 1 Hz

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
100.00	100.00	0.00	3.00	2.45	18.37	0.50	0.50	0.08
101.00	100.81	0.19	2.00	2.46	22.85	-0.50	-0.33	5.43
200.00	200.00	0.00	2.00	1.62	19.10	1.00	0.83	5.43
202.00	201.21	0.39	1.00	1.39	38.67	-1.00	0.07	33.94
300.00	300.30	0.10	4.00	4.16	4.08	1.50	1.04	14.56
303.00	303.03	0.01	3.00	2.92	2.70	-1.50	-1.49	0.17
400.00	400.00	0.00	2.00	1.74	13.20	2.00	1.97	0.86
404.00	403.23	0.19	1.00	1.33	32.69	-2.00	-0.71	41.08
	401.61			0.61			-0.85	

Table 21. Results for the filtered case at 1 Hz separation using Simultaneous-Equation algorithm.

5.9.5 Analysis of Results

As discussed previously, the matrix approach is the most robust, being a multi-pass algorithm which is better able to tolerate filter nonidealities. The iterated subtraction method is prone to error buildup from the repeated subtraction of bands with almost identical frequencies. The Peak-Locus method is a single-pass algorithm, and depends critically on perfect exponential properties, which are not easy to achieve in practice, as we discussed in Section 5.8. We therefore present results of the latter two only for the weighted situation.

We note that for the weighted tests, in many cases the numerical results obtained with all three methods for the same test case were similar. This seems to indicate that any of the three algorithms can potentially find the optimal solution, and are limited mainly by external factors such as roundoff errors or inaccuracies due to insufficient sampling rate.

For the filtered case, the proper number of sines was resolved in the Delta-10-Hz and Delta-5-Hz tests, but an extra non-existent component appeared in the Delta-1-Hz test at 401.61 Hz of amplitude 0.6 in between the actual 400 and 404 Hz components of the mixture. This would then affect the accuracy of the other estimates, as well, as it "steals" energy from where it rightfully belongs. It is somewhat paradoxical that the closer 100 and 101 Hz pair were resolved fairly well, while the 400 and 404 Hz pair were not. The explanation is probably due in part to the reduced sampling rate that we used in the filtered tests compared to the weighted tests, thus leaving insufficient sampling points available to correctly locate the more closely spaced peaks of higher frequency signals.

The filtered case results shown here were later improved upon with the use of low-pass interpolation to increase the effective sampling rate to 4 MHz, and with the widening of the filters discussed in Section 5.8 from 10-Hz lower and 1-Hz upper bandwidth to 50-Hz lower and 2-Hz upper bandwidth. The correct number of sines was attained, and error was decreased. In Chapter 6 we describe these refinements in conjunction with tests on time-varying signals.

5.10 Summary

We have examined the use of local maxima for providing additional temporal information on parameters of mixtures to improve parameter estimation. We motivated this need by looking at

interference within individual bands of a spectrogram and within filter channels caused by competing speakers. We also noted the need to sort out modulation from interference, and that the use of multiple harmonics could assist. We then turned to the general question of whether consolidation of information from multiple overlapping channels could provide more accuracy than from a single channel alone. We found that this appeared to be the case in initial tests of an Iterative-Subtraction algorithm on both single and multiply weighted versions of a signal. Due to some difficulties with convergence in our initial approach when extending to simulations of actual filters, we developed two additional algorithms based on properties of overlapping exponential filter banks. We introduced the use of Peak-Locus plots which pinpoint discontinuities in the patterns of local maxima between channels. We showed that these points indicate the presence of spectral components. We then developed two algorithms to harness this, with the first being a single-pass algorithm based directly on the Peak-Locus diagram which subtracts estimates one at a time until no discontinuities remain. This algorithm had some difficulties when used with actual filters due to the fact that in practice these discontinuities are more gradual than in pure mathematical simulations, making it difficult to pinpoint the exact location, especially when dealing with closely spaced components. We then introduced a more accurate, iterative approach based on a Simultaneous-Equation formulation which was more adept at demarcating these points, and recovering the parameters of the underlying components. All of the foregoing approaches represent effective methods of dimensionality reduction, in that they efficiently collapse multiple varied channel outputs to those which contain novel information on component parameters.

We performed comparative evaluations on the three algorithms using mixtures of constant four-harmonic test sets with different levels of separation in frequency, and found that in their present forms, the Simultaneous-Equation method produced the best results. We continue with further tests of this algorithm in Chapter 6 on more realistic test sets. Nevertheless, the other algorithms were important in the step-by-step development of our thinking, and are useful as pedagogical tools. In particular, the Peak-Locus method is a logical prerequisite for understanding the Simultaneous-Equation formulation.

Chapter 6

Combined-Channel Analysis of Modulation, Noise and Speech

6.1 Introduction

In this chapter we conduct tests on more complex signals than the mixtures of harmonic sets we used previously. Our goals are to test the performance of the matrix method of Chapter 5 in more realistic situations, and to gain further insight into the source separation problem based on examination of results obtained with this method. We test on modulated sinusoids, harmonic sets in noise, and actual speech recordings.

We will find that the approach we have taken in using information from multiple bands yields frequency resolution which improves upon the resolution of the FFT in certain cases, while simultaneously producing parameter estimates that could be characterized as local or instantaneous in nature. We will discuss the implications and arguments both for and against this claim, and possible application to source separation.

6.2 Filter Parameters

As discussed in Section 5.8, the use of interpolation permitted an increase in effective sampling rate to 4 MHz from the 500 KHz that we had used previously without exceeding system resources. This allowed for wider filters without sacrificing frequency resolution, and as a result, better time resolution could be achieved. Interpolation was performed via the Matlab® command `interp` with interpolation filter of length 4 and cutoff frequency 0.5. Occasional evidence of high frequency, low-amplitude interpolation noise plagued some tests, and may be a possible source of error, but was found to be less of a problem with these settings than with

longer interpolation filters. Upper bandwidth of the exponential filters was set at 2.0 and lower bandwidth was 50.0 for all filters and all tests. As previously, upper bandwidth needs to be sharper than lower bandwidth because of the dominance of higher frequencies upon the locations of the peaks. Higher frequencies, therefore, need to be more heavily limited.

Because the Simultaneous-Equations approach is a multipass algorithm, we have found it to be the most reliable, and best tolerates filter nonidealities. We have therefore conducted all tests in this chapter using that method. The complete list of all parameters used was presented in Table 9, previously.

6.3 Harmonic Mixture-Delta 1 Hz

6.3.1 Results with Simultaneous-Equations Algorithm

To illustrate performance of the wider filter, we repeat tests on two sets of four harmonics with fundamental frequencies separated by 1 Hz. They were deliberately chosen to be at nonintegral values for reasons we explained in 5.8.

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
97.25	97.19	0.07	3.00	2.60	13.43	0.50	0.78	8.82
98.25	98.36	0.11	2.00	1.60	19.99	-0.50	-0.96	14.65
194.50	194.42	0.04	2.00	1.90	4.94	1.00	1.27	8.75
196.50	196.48	0.01	1.00	1.02	2.32	-1.00	-0.85	4.90
291.75	291.74	0.00	4.00	3.97	0.79	1.50	1.54	1.35
294.75	294.75	0.00	3.00	3.03	1.04	-1.50	-1.49	0.19
389.00	388.95	0.01	2.00	1.95	2.70	2.00	2.15	4.93
393.00	393.16	0.04	1.00	0.95	5.35	-2.00	-2.54	17.06

Table 22. Results on nonintegral harmonic sets with fundamentals separated by 1 Hz.

Frequency errors were less than 0.12%, but amplitude errors for the first two components, the most closely spaced, were as much as 20%. Possibly, better filter design and choice of parameters might improve upon this, as well as possible enhancements mentioned in Section 5.7. Nevertheless, for most situations, we did not encounter frequencies that close, and relied upon the current filter design which is based upon a simple inverse FFT of an ideal exponential, with great care taken in properly indexing the time axis. We did not use windows, as we found they negatively impacted upon desired exponential filter shape.

6.3.2 Results Using FFT

It is now appropriate to make a direct comparison with the performance of the FFT which is shown in Figure 70. Using the same 1-second length of the above signal with 10-KHz sampling rate, the FFT was unable to resolve either of the first two harmonic pairs. The third pair was resolved, but with frequency and amplitude errors. The fourth pair was correctly resolved and estimated.

The reason for this is that the FFT can only give accurate estimates at bin centers. Bin centers are calculated by realizing that the number of points on the time axis N get translated into the number of points on the frequency axis, after the transform is applied. The sampling rate F_s determines the highest frequency on the frequency axis, which according to the Nyquist theorem is $F_s/2$. Because there are positive and negative frequencies along the frequency axis, the actual number of unique frequency points is reduced to $N/2$. Therefore, the bin centers lie at multiples of

$$\frac{F_s/2}{N/2} = \frac{F_s}{N} \tag{6.1}$$

However, the number of points N can further be broken down into the product of the signal duration T and the sampling rate F_s, or $N = TF_s$. We therefore have that the distance between bin centers is

$$\Delta f = \frac{F_s}{N} = \frac{F_s}{TF_s} = \frac{1}{T} \tag{6.2}$$

This is a simple derivation of the useful rule of thumb we noted in Section 5.2 that the resolution of conventional methods is the reciprocal of the signal duration. But more than that, it tells us that amplitudes of components whose frequencies are not themselves exact multiples of $1/T$ will not be correctly estimated, even under standalone conditions, and certainly in mixtures. An equivalent way to think about this is that the signal duration should enclose an integral number of cycles. For our case we deliberately chose frequencies which lie in-between filter locations, or denominations of 0.25 Hz. For such frequencies, we would need a recording length of 4 seconds. Since we limited our signal to a duration of 1 second, only integral

frequency components can be correctly estimated. From Table 22, only the last harmonic pair has integral frequencies, and hence it is the only one correctly resolved and estimated.

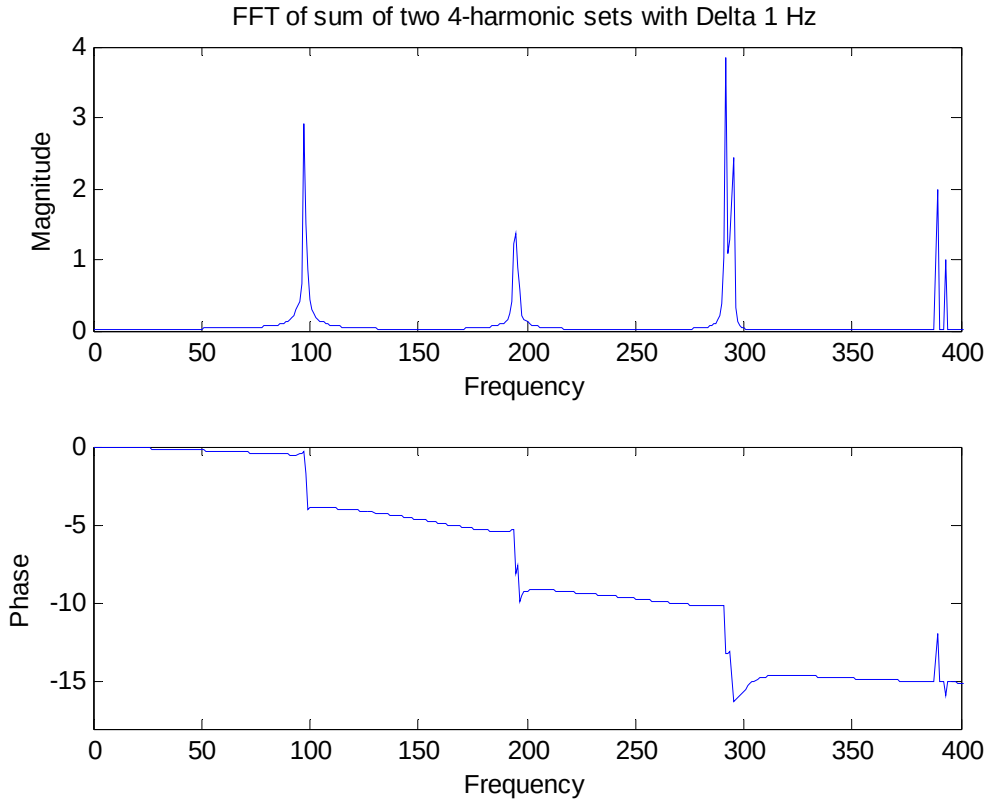

Figure 70. FFT computed from the two 4-harmonic sets of Delta 1 Hz whose parameters are listed in Table 22. First two harmonics pairs are not resolved separately. Third pair is resolved, but with frequency and amplitude errors. Fourth is correctly resolved and estimated.

It is clear that the performance of our algorithm exceeds the performance of the FFT both in overall resolution and accuracy, except for the last pair, where the FFT is slightly more accurate. We note further that because our filter duration is 0.5 seconds as in Figure 68, the extra 0.5 seconds by which the signal exceeds the filter duration is actually irrelevant and superfluous for our analysis, hence the effective length used by our algorithm is really only 0.5 seconds, making the actual resolution limit of the FFT for a similar length signal 2 Hz, rather than the 1 Hz stated earlier. We furthermore believe that amplitude estimates with our method can possibly be improved by adding an additional calibrating step to avoid reliance on the filter acting as a true mathematical exponential, as suggested in Section 5.8, but we have not implemented this due to the high computational burden.

6.3.3 Source of Errors in FFT

For the sake of completeness, we further digress to better understand the difficult and confusing issue regarding the nature of the errors that occur when the frequency to be estimated does not lie on a bin center. Our discussion follows that of (Schiff, 1997) which is noteworthy for its concise and clear linkage of digital-signal-processing convention with its continuous-time counterpart.

The starting point is the continuous-time Fourier transform pair.

$$H(f) = \int_{-\infty}^{\infty} h(t) e^{2\pi jft} dt$$
$$h(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(f) e^{-2\pi jft} df \tag{6.3}$$

Since in the real world, we can only observe over a finite time T we rewrite $H(f)$ as

$$H(f) = \int_{0}^{T} h(t) e^{2\pi jft} dt \tag{6.4}$$

The next step is to convert into a form that the digital computer can use, which is as a series of discrete time steps, rather than as a continuum. We numerically approximate the integral as a series of rectangles.

$$H_T(f) = \Delta t \sum_{k=0}^{N-1} h_k e^{2\pi jfk\Delta t}$$
$$h_k = h(k\Delta t) \tag{6.5}$$

where $\Delta t = 1/F_s$ is the time spacing between samples.

For computational purposes, we must similarly limit the frequency spectrum to a set of finite frequencies separated by Δf , so we rewrite Eq. 6.4 as

$$H_T(m\Delta f) = H_m = \Delta t \sum_{k=0}^{N-1} h_k e^{2\pi jmk\Delta t\Delta f} \tag{6.6}$$

Because $\Delta t = 1/F_s$ as before, and because $\Delta f = F_s/N$ as before, we have

$$\Delta t \Delta f = 1/N \tag{6.7}$$

Making this substitution, the DFT transform pair becomes

$$H_m = \sum_{k=0}^{N-1} h_k e^{2\pi jkm/N}$$
$$h_k = \frac{1}{N}\sum_{m=0}^{N-1} H_m e^{-2\pi jkm/N} \tag{6.8}$$

Let us now apply these to a sine wave where

$$\begin{aligned} h(t) &= \cos(2\pi f_0 t) \\ h_k &= \cos(2\pi f_0 k\Delta t) \\ &= \cos(2\pi n\Delta f k\Delta t) \\ &= \frac{e^{2\pi nk\Delta f\Delta t} + e^{-2\pi nk\Delta f\Delta t}}{2} \end{aligned} \tag{6.9}$$

in which we have chosen to represent the frequency $f_0 = n\Delta f$ where n is not necessarily an integer.

For simplicity, considering only one of the exponential terms, and substituting into the transform equation we have

$$H_m = \sum_{k=0}^{N-1} e^{2\pi jk(m-n)/N} \tag{6.10}$$

Being a geometric series, the expression can be evaluated in closed form as

$$H_m = \frac{e^{j2\pi(m-n)} - 1}{e^{j2\pi(m-n)/N} - 1} \tag{6.11}$$

In keeping with Schiff's preference to use the power spectrum $P_m = H_m H_m^*$, rather than the magnitude, because of its analogy to an analog spectrum analyzer, we compute

$$P_m = \frac{\sin^2[\pi(m-n)]}{\sin^2[\pi(m-n)/N]} \tag{6.12}$$

We consider three cases. The first, when n is an integer, and $m \neq n$, then $P_m = 0$, since m is also an integer. The second, when n is an integer, and $m = n$, the numerator and denominator are both zero, and the limiting result is that $P_m = N^2$. From these two cases, when n is an integer we have that $P_m = 0$ everywhere, except at the input frequency $f_0 = n\Delta f$, which is intuitively as we would expect. However, when n is not an integer, then the digital representation and the

continuous representation as would be seen on a spectrum analyzer differ the most, and this is a major source of confusion which we will shortly address.

As an example, Figure 71 shows the power spectrum of a sinusoid of 32.5 Hz. Recording length T is 1 second, hence frequency resolution $\Delta f = 1/T$ is 1 Hz. Because the frequency is nonintegral, there is noticeable spectral spreading which appears to indicate the presence of modulation, even though the signal is a pure tone. The response appears to roll off slowly as $1/f^2$. When the frequency is changed to 32 Hz, as in Figure 72, spreading disappears, and, as expected, frequency response becomes an impulse over many orders of magnitude all the way down to the hardware noise floor of the computer.

To understand this behavior we must go back to the original transform pair of Equation 6.3, and compute the transform for the same sinusoid in continuous time. After converting to the continuous power spectrum, as we did for the discrete time case, the result is

$$P(f) = \left\{ \frac{\sin[\pi(f - f_0)T]}{\pi(f - f_0)T} \right\}^2 \tag{6.13}$$

In the following two figures, we overlay the continuous power spectrum with the digital power spectrum for the same sinusoid. In the nonintegral case of Figure 73, the digital sampling frequencies become out of sync with the underlying continuous sinc function, and the result has the appearance of spreading. In the integral case of Figure 74, the sampling frequencies correspond with the centerlobe maximum and the sidelobe minima, producing the behavior of a digital impulse function, as expected.

Counterintuitively, A) increasing sampling rate will not increase resolution, but only allow for additional higher frequencies to be represented without aliasing, which will not be of assistance if the signal of interest is a lower frequency signal; B) increasing recording time may actually decrease digital resolution if it causes frequency of signal to become mismatched with digital sampling frequencies, i.e., if $f_0 \neq n/T \;\; \exists n$. The recording time needs to be exactly matched with the frequency of the signal to avoid spreading. The implications for frequency-varying cases are that it will be impossible to match the frequency everywhere within the recording time, and this will inadvertently lead to some degree of spreading.

Our methods are not as vulnerable to this type of phenomenon, since they do not compute frequency directly from the discrete transform. Instead, they use digital filters only as a means of differentially weighting signals from different channels. The individual channel responses are then all concurrently fed into the algorithm, and a set of parameter estimates is produced from the combined information.

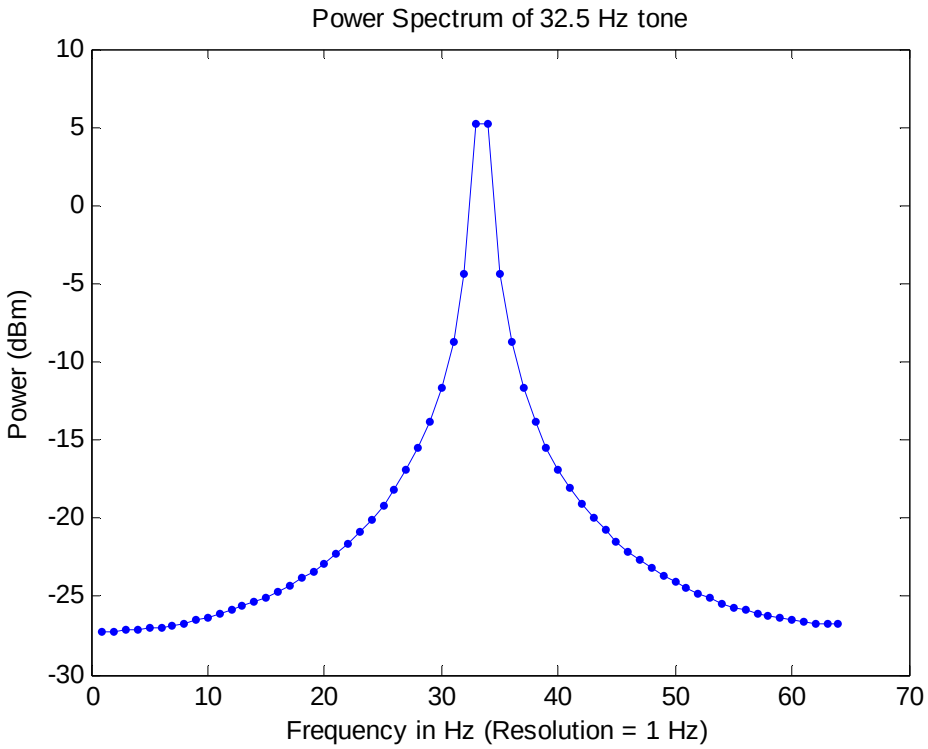

Figure 71. Power Spectrum of a 32.5 Hz tone of duration *T*=1 second and sampling rate 128 Hz. Resolution=1/T or 1 Hz. Because frequency is nonintegral, spreading appears to occur as explained in text.

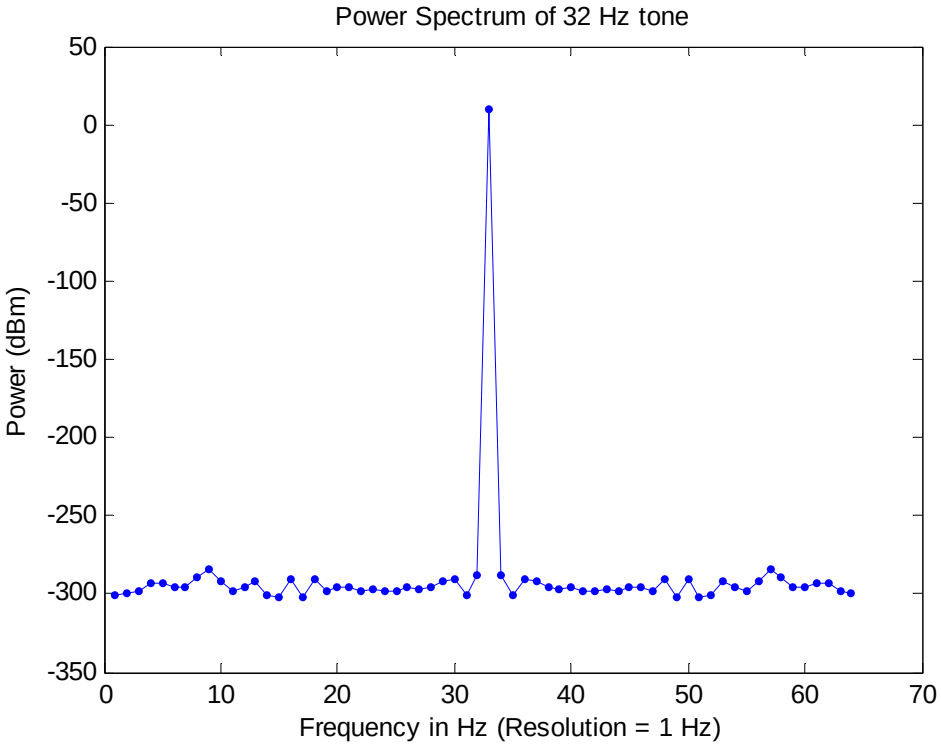

Figure 72. Power Spectrum of a 32 Hz tone of duration *T*=1 second and sampling rate 128 Hz. Resolution=1/T or 1 Hz. Because frequency is now an integer, impulse-like behavior occurs over many orders of magnitude.

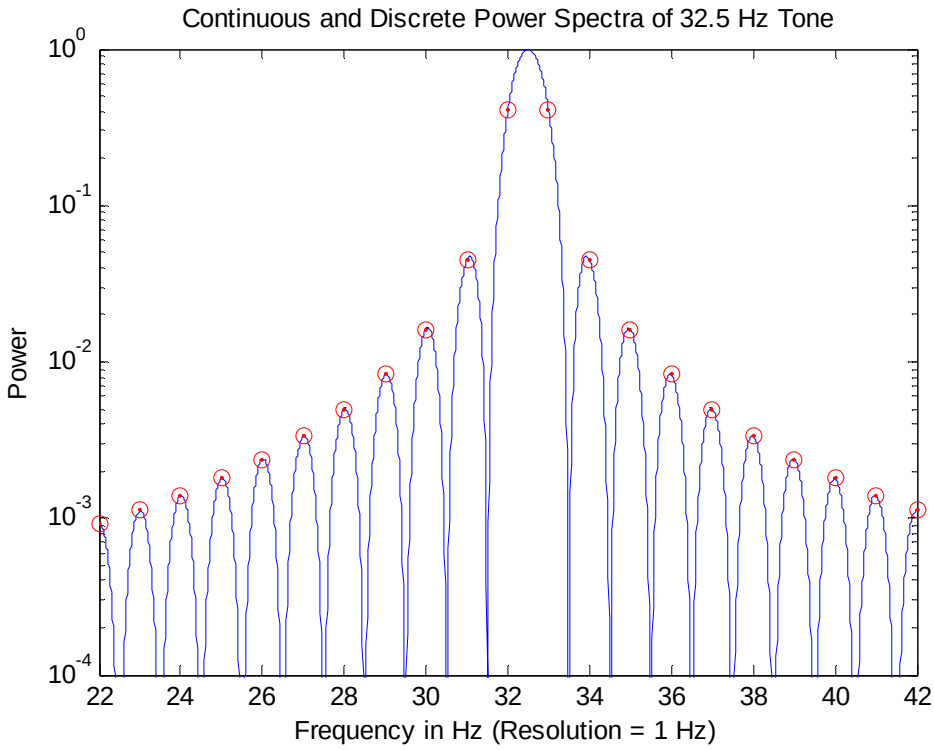

Figure 73. In the nonintegral frequency case, the digital frequency sampling points do not correspond with the center frequency of the underlying continuous distribution, thus causing errors in the digital power spectrum and the appearance of spreading.

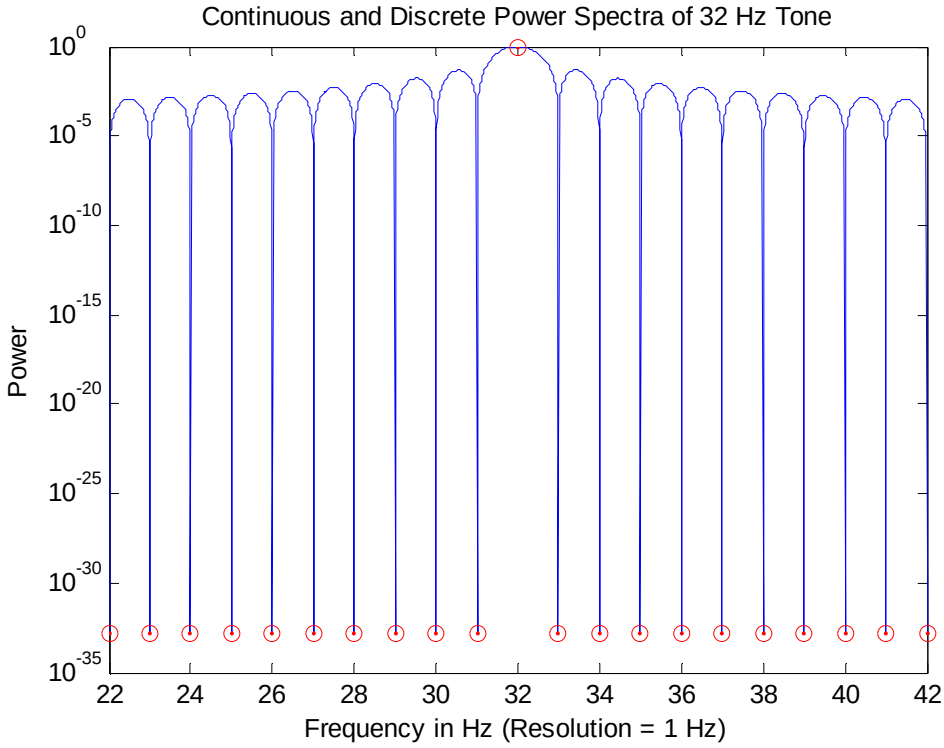

Figure 74. In the integral frequency case, the digital frequency sampling points correspond with the center maximum and sidelobe minima of the underlying continuous distribution, giving a digital impulse function.

6.4 Modulated Sinusoids

In addressing signal modulation, we used the following signals: AM and FM ramp-modulated sinusoids, and AM and FM sine-modulated sinusoids. Each of the tests in this section was performed with a single time-varying sinusoid. As our methods are designed to analyze instantaneous parameters, we occasionally repeated tests at various times to see whether recovered parameters match expected time-dependent behavior of modulated signals. In addition, we occasionally repeated tests at the same time point using different threshold parameters for reasons which will become clearer later, and these produced interesting effects on the results. We remind reader that there are two thresholds that can be adjusted, a frequency duplication threshold which controls how far apart in frequency two components need to be before being counted as separate, and an amplitude threshold which eliminates bands with amplitudes below this value, which can possibly be of use in eliminating low level noise.

6.4.1 AM Ramp-Modulated sine

A test signal was generated as follows.

(6.14) $$x = t[\sin(2\pi 100t)]$$

T=0.65 sec

Algorithm finds a single sine of expected instantaneous amplitude corresponding to time 0.65 seconds in ramp.

Recov Freq	Recov Amp	Recov Phase
100.00	0.65	-0.00

Results are consistent with expected instantaneous behavior of an AM signal. (At time 0.65 seconds, the recovered amplitude of the ramp is 0.65.)

6.4.2 FM Ramp-Modulated Sine

A test signal was generated as follows:

(6.15) $$x = \sin(2\pi \int [95 + 10t]dt)$$

T=0.35 sec

Recov Freq	Recov Amp	Recov Phase
96.94	0.32	-1.31
98.51	0.82	-3.11
100.05	0.32	-1.86

Table 23. Results of FM ramp at time 0.35 seconds. Three components were found. Center component corresponds to instantaneous frequency expected at that time, but with slightly lower amplitude, as is common in FM. Symmetric frequency components were found above and below, each of which apparently is a consolidation of many closely spaced sidebands.

T=0.5 sec

Recov Freq	Recov Amp	Recov Phase
98.43	0.31	2.45
100.02	0.84	-0.89
101.58	0.32	-1.19

Table 24. Results of FM ramp at time 0.50 seconds. Three components were found, as before. Center component corresponds to instantaneous frequency expected at that time, but with slightly lower amplitude, as is common in FM. Symmetric frequency components were found above and below, each of which apparently is a consolidation of many closely spaced sidebands, as before.

T=0.65 sec

Recov Freq	Recov Amp	Recov Phase
99.94	0.32	-1.51
101.50	0.82	0.02
103.06	0.32	-1.68

Table 25. Results of FM ramp at time 0.65 seconds. Three components were again found. Center component corresponds to instantaneous frequency expected at that time, but again with slightly lower amplitude, as is common in FM. Symmetric frequency components were again found above and below, each of which apparently is a consolidation of many closely spaced sidebands.

At each time, the central frequency exactly corresponds to the instantaneous value expected from Equation (6.15) on the basis of 10 Hz/sec sweep. In each case, there are two sidebands about 1.55 Hz above and below the central carrier. The amplitude of the carriers is similar at each time, and is below unity, as is common in FM signals. The upper and lower sideband amplitudes are of comparable magnitude, and are similar at each of the three times.

On the basis of these tests, we conclude that results are consistent with expected instantaneous behavior of FM signals. We note that, to our knowledge, there are no tabulated results of the exact sidebands of ramped FM signals, as there are for sine-modulated FM signals that we will examine later. The reason is probably because of the mathematically complex and dense distribution of these sidebands. The algorithm seems to consolidate these into three

components, with the center component exactly obeying the relationship $f_{center} = 95 + 10t$, as expected.

In Section 6.7 we will discuss the implications of the finding of three separate components, rather than one single component.

6.4.3 AM Sine-Modulated Sine.

A test signal was generated as follows:

(6.16) $$x = [1 + \sin(2\pi 5t)][\sin(2\pi 100t)]$$

The waveform is shown in Figure 75. Tests were repeated at 3 different times, to study time course of behavior.

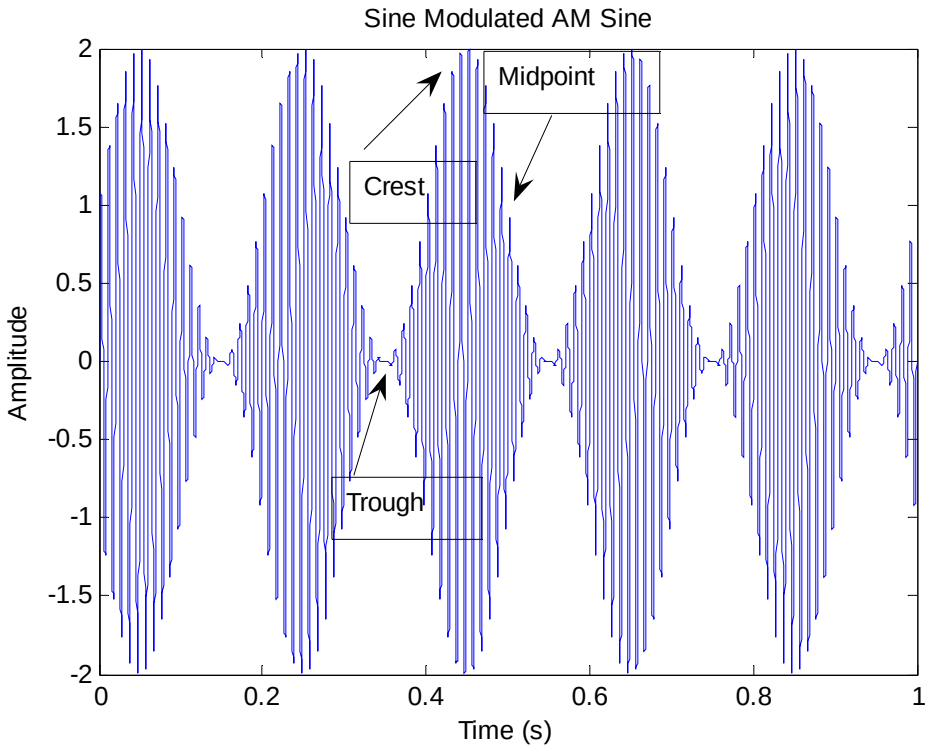

Figure 75. AM sinusoidally modulated sine waveform discussed in text. The three sets of peak pairs at which analysis was run are labeled. The corresponding times are 0.35 seconds (trough), 0.45 seconds (crest) and 0.5 seconds (midpoint).

Results were as follows:

T=0.5 sec (Midpoint of Envelope)

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
95.00	95.01	0.01	0.50	0.50	0.28	1.57	1.55	0.52
100.00	100.00	0.00	1.00	0.99	0.84	0.00	0.00	0.05
105.00	104.99	0.01	0.50	0.50	0.26	-1.57	-1.54	0.98

Table 26. Results for AM sine-modulated sine at time 0.5 seconds, corresponding to midpoint of envelope.

T=0.45 sec (Crest of Envelope)

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
95.00	95.19	0.20	0.50	0.52	4.05	1.57	1.04	16.94
100.00	100.00	0.00	1.00	0.95	4.79	0.00	0.00	0.00
105.00	104.73	0.26	0.50	0.53	5.25	-1.57	-0.82	23.92

Table 27. Results for AM sine-modulated sine at time 0.45 seconds, corresponding to crest of envelope.

T=0.35 sec (Trough of Envelope)

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
95.00	95.15	0.16	0.50	0.51	1.58	1.57	1.23	10.98
100.00	100.00	0.00	1.00	1.01	0.90	0.00	0.00	0.00
105.00	104.92	0.08	0.50	0.50	0.03	-1.57	-1.39	5.72

Table 28. Results for AM sine-modulated sine at time 0.35 seconds, corresponding to trough of envelope.

Figure 76 show results for time *t*=0.5 seconds. At that time, the value of the signal envelope is 1.0, midway between the extremes of 0 and 2 through which it oscillates. Computed results for this run listed above show that the values of all components are virtually identical to the theoretical values of 95, 100 and 105 Hz for the lower sideband, carrier and upper sideband, respectively. Amplitudes are expected to be 0.5, 1.0 and 0.5; while the phases are expected to be $\pi/2$, 0, and $-\pi/2$, respectively.

A question which might be raised is whether the values of the signal components would change at times near the crest or trough of the modulating envelope. We therefore repeated at *t*=0.35 seconds, and at *t*=0.45 seconds where the envelope is minimum and maximum, respectively. The differences in above results do not appear to be significant, although there are some slight variations in the parameter estimates at these last two trials compared to the first.

While the values of the estimated parameters do not differ significantly at the three times, we note that the raw band data is completely different for the runs at the maximum, midpoint and minimum of the envelope. Figure 76, Figure 77 and Figure 78 display in graphical form the initial unprocessed frequency and amplitude band data superimposed on the final processed frequency and amplitude data listed above for each of the three runs. The continuous green line

represents the frequency estimate of each band (left ordinate) vs. the filter CF of the band (abscissa) for the first (unprocessed) iteration. This data is obtained from the local maxima of that band at the given time point before consolidation with data from other bands. The continuous blue line represents the amplitude estimate of each band (right ordinate) vs. the filter CF of the band (abscissa) for the first (unprocessed) iteration. This data is again obtained from the local maxima of that band at the given time point before consolidation with data from other bands. The strength of the algorithm is in its ability to uncover the hidden components within the band data that are the cause of the behavior observed in the separate channels. The most prominent example of this is at time t=0.35 seconds which is at the minimum of the swing. In that run, the unprocessed amplitude values of almost all bands are seen to be extremely low. The maximum amplitude for any band is only about 0.5 in the 101 Hz band which has a corresponding frequency estimate of 103.29 Hz. Successively higher bands have somewhat decreasing frequency estimates until the 105.5 Hz band, at which there is a sudden jump to very high frequency estimates. This behavior seems almost bizarre and hard to explain. Recall that the Peak-Locus method predicts that there will be discontinuities at values at which a new frequency component enters. It does not say what the estimates will be at those points, they can be greater than previous estimates as one might expect, but can also be less, as we will see in Chapter 7. There are sudden and clear changes near, but not exactly at, the 100 Hz and the 105 Hz bands, as theory predicts, but the picture overall is quite obfuscated. The fact that the matrix algorithm is able to decipher this extremely mystifying situation into 3 simple components of amplitude 0.5, 1.0 and 0.5 at 95, 100 and 105 Hz is quite a feat, considering that there is no visible evidence of any component with amplitude greater than 0.5. As before, we will discuss the implications of the finding of 3 components, rather than one single component in Section 6.7

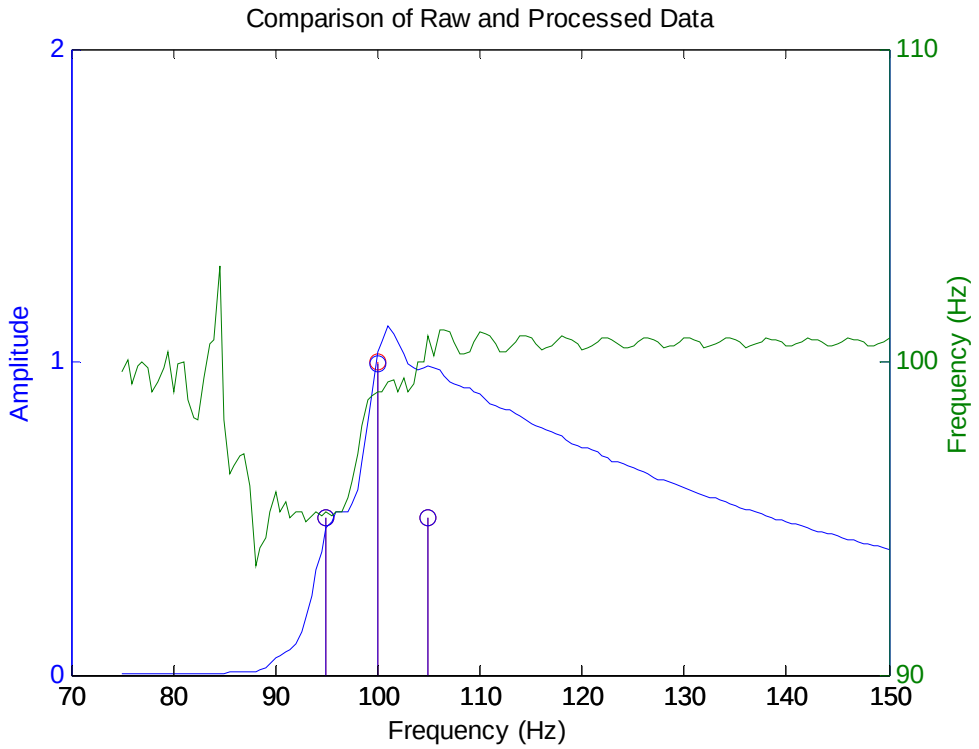

Figure 76. Comparison of raw vs. processed band data at t=0.50 seconds for a sinusoidally amplitude-modulated sine. The continuous plots show band CF in Hz (X axis) vs. amplitude (blue, left Y axis), and frequency (green, right Y axis) of initial unprocessed data as determined by the pair of local maxima which brackets time *t*. The blue stems depict final processed values of frequency vs. amplitude for each component found using the iterative Simultaneous-Equation algorithm. (X axis for stem plots is no longer filter CF, as before, rather is processed frequency.) The red stems indicate theoretical values of frequency vs. amplitude for the carrier and sidebands of this AM signal.

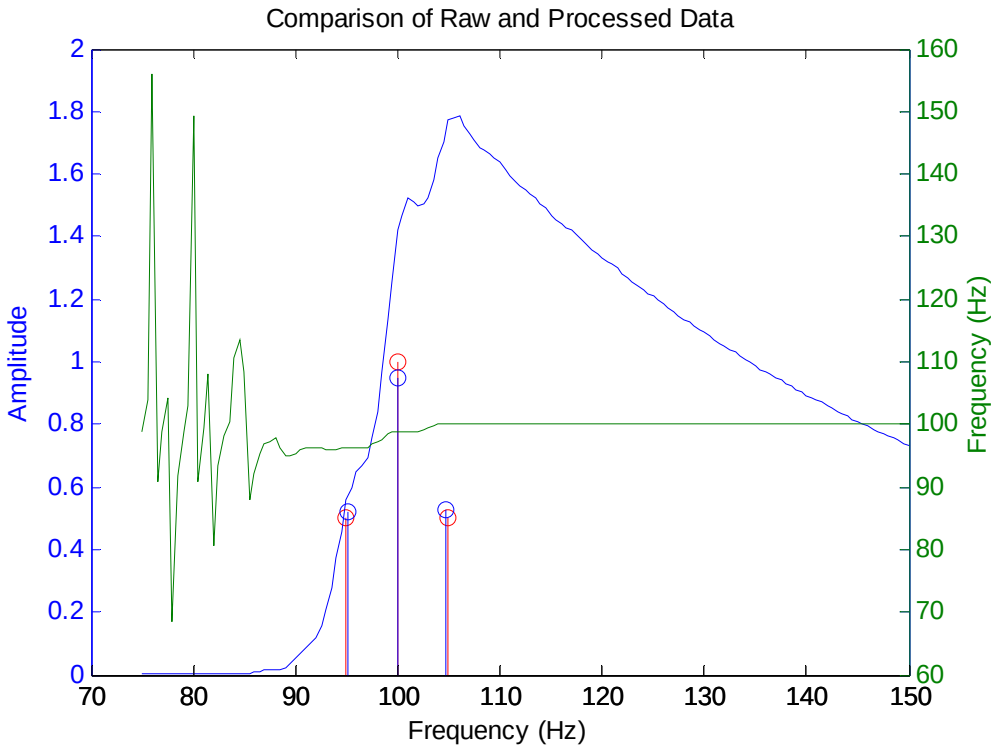

Figure 77. Comparison of raw vs. processed band data at t=0.45 seconds for a sinusoidally amplitude-modulated sine. The continuous plots show band CF in Hz (X axis) vs. amplitude (blue, left Y axis), and frequency (green, right Y axis) of initial unprocessed data as determined by the pair of local maxima which brackets time *t*. The blue stems depict final processed values of frequency vs. amplitude for each component found using the iterative Simultaneous-Equation algorithm. (X axis for stem plots is no longer filter CF, as before, rather is processed frequency.) The red stems indicate theoretical values of frequency vs. amplitude for the carrier and sidebands of this AM signal. Note that band data indicates amplitudes values as high as 1.8 due to choice of analysis time at crest of modulation waveform, although actual component values do not exceed 1.0.

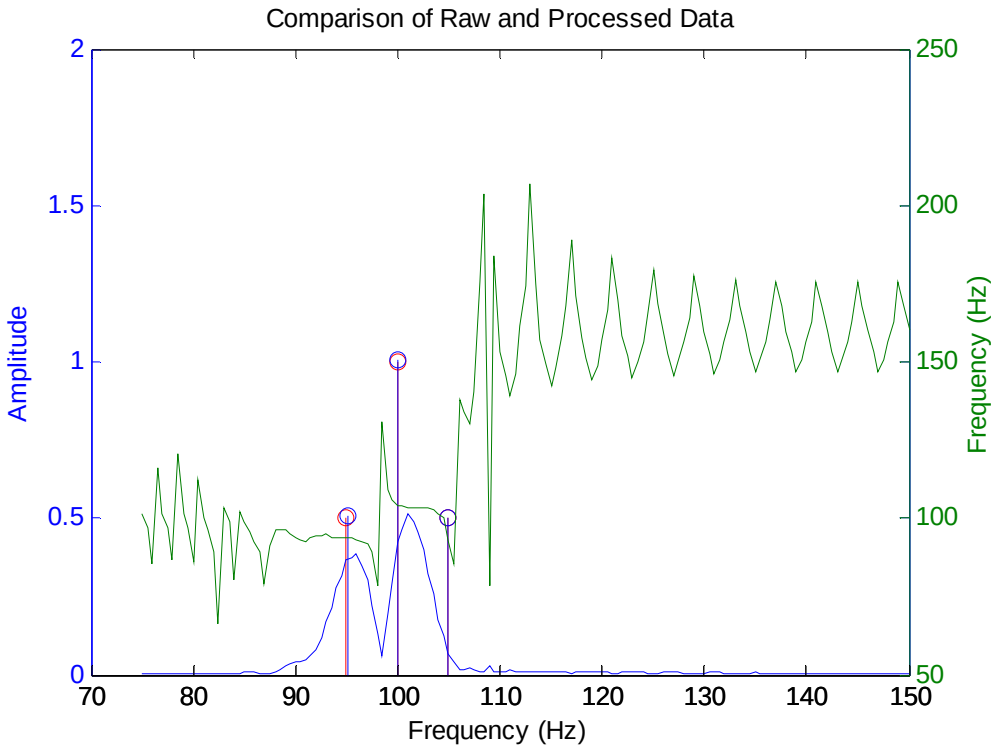

Figure 78. Comparison of raw vs. processed band data at t=0.35 seconds for a sinusoidally amplitude-modulated sine. The continuous plots show band CF in Hz (X axis) vs. amplitude (blue, left Y axis), and frequency (green, right Y axis) of initial unprocessed data as determined by the pair of local maxima which brackets time *t*. The blue stems depict final processed values of frequency vs. amplitude for each component found using the iterative Simultaneous-Equation algorithm. (X axis for stem plots is no longer filter CF, as before, rather is processed frequency.) The red stems indicate theoretical values of frequency vs. amplitude for the carrier and sidebands of this AM signal. Note that there is no indication in unprocessed amplitude data of anything significant at the frequency 105 Hz. The component only becomes apparent after processing.

For comparison, Figure 79 shows the FFT of the same AM signal. Duration was chosen to be 0.5 seconds to exactly match the filter length of 0.5 seconds used in all tests of this chapter. Lesser accuracy in both frequency and amplitude is apparent. This can be improved by adjusting the FFT duration to better match the signal characteristics, but this sensitivity represents a weakness in FFT based methods, as slight adjustments in FFT duration drastically alter estimates. The problem was worse in speech tests in Section 6.6, where the waveform is only quasi-periodic, and was worse still in tests of mixed speech in which it was impossible to match FFT duration to an integral number of waveform cycles since the mixture was completely aperiodic by nature. Our methods do not require such a match.

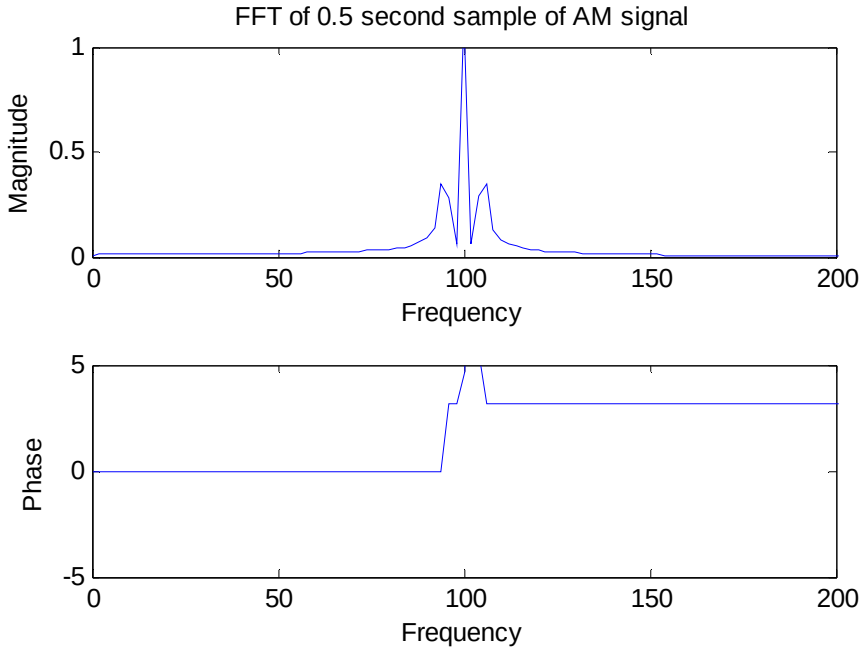

Figure 79. The FFT of the AM sine-modulated sine used in tests. Duration was 0.5 seconds. Three spectral components are visible, but exhibit spreading rather than the expected line structure. Peak values are 94 Hz, 100 Hz and 106 Hz, with amplitudes of 0.35, 1.0 and 0.35, respectively. These differ from the expected values discussed in text. Better matching of FFT length to signal characteristics does improve accuracy, but nevertheless represents a practical disadvantage.

6.4.4 FM Sine-Modulated Sine

The signal was generated as follows:

(6.17) $$x = \sin\left\{2\pi \int [100 + 5\cos(2\pi 5t)dt]\right\}$$

This represents a 100 Hz sinusoidal carrier which is frequency-modulated by a 5-Hz cosine with a carrier deviation of 5 Hz, thus oscillating in a range of 95-105 Hz.

The test was repeated 4 times. Two time points of analysis were used, and two amplitude thresholds were tried for each.

T=0.40 seconds, Amplitude threshold=0.01

Analysis was performed at time 0.40 seconds using duplication threshold of 0.1 Hz. At this time, the instantaneous frequency would be expected to be at its peak of 105 Hz. Results were as in Table 29.

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error
85.00	X	X	0.02	X	X
90.00	90.35	0.39	0.11	0.09	17.68
95.00	95.04	0.05	0.44	0.43	2.13
100.00	100.00	0.00	0.77	0.74	3.88
105.00	104.48	0.50	0.44	0.41	6.81
110.00	108.61	1.26	0.11	0.15	26.65
115.00	113.64	1.18	0.02	0.05	129.95

Table 29. Results for FM sine-modulated sine at time 0.40 seconds.

There is no indication of anything special about the freq 105 Hz. Rather, the picture seems to be that of a 100 Hz carrier surrounded by sidebands of approximately +/- 5 Hz, +/- 10 Hz, and +/- 15 Hz, although the lower 3rd sideband expected at 85 Hz is missing, and the upper sidebands of 108.61 and 113.64 differ by a bit more than 1 Hz from the theoretical values of 110 and 115, respectively.

Table 30 below is from (Libbey, 2006).

Modulation m_f	Carrier J_0	Sideband 1 J_1	Sideband 2 J_2
0.25	0.98	0.12	0.01
0.50	0.94	0.24	0.03
1.00	0.77	0.44	0.11
1.50	0.51	0.56	0.23
2.00	0.22	0.56	0.35
2.40	0.00	0.52	0.43
3.00	-0.25	0.34	0.49
4.00	-0.40	-0.07	0.36
5.00	-0.18	-0.33	0.05
5.50	0.00	-0.34	-0.12
6.00	0.15	-0.28	-0.24
7.00	0.30	0.00	-0.30
8.00	0.17	0.23	-0.11
8.65	0.00	0.27	0.06

Table 30. Relative amplitudes of FM sidebands for various values of the modulation index m_f defined in text.

$$m_f = \frac{Carrier\,Frequency\,Deviation}{Modulation\,Frequency} = \frac{\Delta f_c}{f_m}$$

The amplitudes of FM sidebands are shown in communications texts to be Bessel functions $J_0, J_1, J_2 \ldots$ of the modulation index m_f, and are tabulated in Table 30.

For our signal, the ratio of frequency deviation to modulation frequency is 1.0. Hence, line 3 of the table indicates theoretical amplitude values of 0.77, 0.44 and 0.11 for the carrier and 1st and 2nd sideband pairs, respectively. The values we have found in most cases are within a few percent of the theoretical values. Note that the carrier amplitude, as shown in the table, is often less than unity for FM signals. Figure 80 graphically displays the initial unprocessed raw data superimposed on the final processed results, and compares with the theoretically expected values.

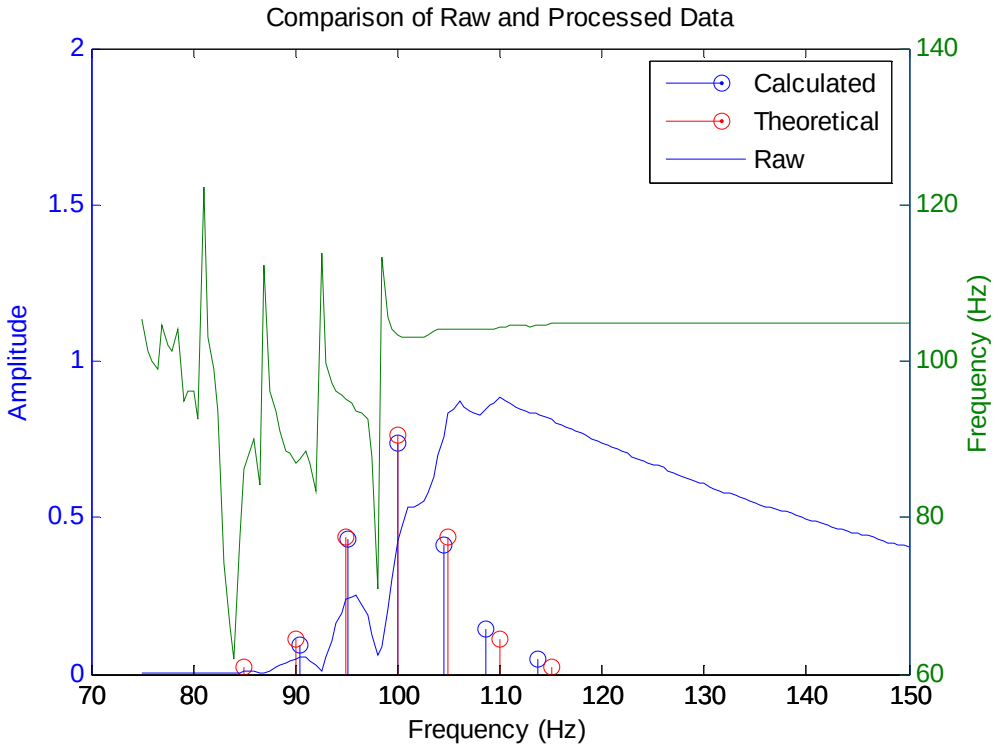

Figure 80. Comparison of raw vs. processed band data at t=0.40 seconds for a sinusoidally frequency-modulated sine. The continuous plots show band CF in Hz (X axis) vs. amplitude (blue, left Y axis), and frequency (green, right Y axis) of initial unprocessed data as determined by the pair of local maxima which brackets time *t*. The blue stems depict final processed values of frequency vs. amplitude for each component found using the iterative Simultaneous-Equation algorithm. (X axis for stem plots is no longer filter CF, as before, rather is processed frequency.) The red stems are the theoretical values of the sideband frequencies and amplitudes as determined by Table 30.

Note that the higher CF filters show frequencies close to 105 Hz, as expected from the properties of the Peak-Locus of exponential filter banks introduced in Chapter 5, which are expected to show strings of similar values in regions in which no new components exist. Note also, that there is no clear evidence of the 110 Hz sideband in the raw data, except for a very slight discontinuity in the unprocessed frequency estimates at 110 Hz CF. The algorithm is able to

recover the latent frequencies by solving equations to obtain the set of sines that would explain the band data.

T=0.30 seconds, Amplitude threshold=0.01

With the generated signal the same as before, and analysis at time 0.30 seconds, at which point instantaneous frequency is expected to be at its minimum value of 95 Hz, results were as in Table 31.

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error
85.00	83.35	1.94	0.02	0.02	2.45
90.00	90.27	0.30	0.11	0.14	17.97
95.00	95.24	0.25	0.44	0.44	0.30
100.00	100.00	0.00	0.77	0.73	4.23
105.00	105.26	0.25	0.44	0.46	4.74
110.00	108.70	1.18	0.11	0.14	17.68
115.00	X	X	0.02	X	X

Table 31. Results for FM sine-modulated sine at time 0.30 seconds.

Figure 81 again displays the initial unprocessed raw data superimposed on the final processed results, and compares with the theoretically expected values. As before, results do not indicate anything special about 95 Hz, but rather appear to show sets of sidebands around a 100 Hz carrier in close agreement with previous results and with Table 30. Amplitude and frequency values are within a few percent of expected values as in previous trial. The upper 3rd sideband is missing, and instead there appears a value of 83.35 Hz which seems to correspond to the expected lower 3rd sideband at 85 Hz. The situation is thus reversed from the previous trial in which the upper 3rd sideband was present, and the lower one was missing. We do not place any significance on the presence or absence of this 3rd sideband, as it is relatively weaker than the lower-order sidebands and carrier. What is interesting about the results in the two cases we have examined, is that the raw band data is completely different at the two times.

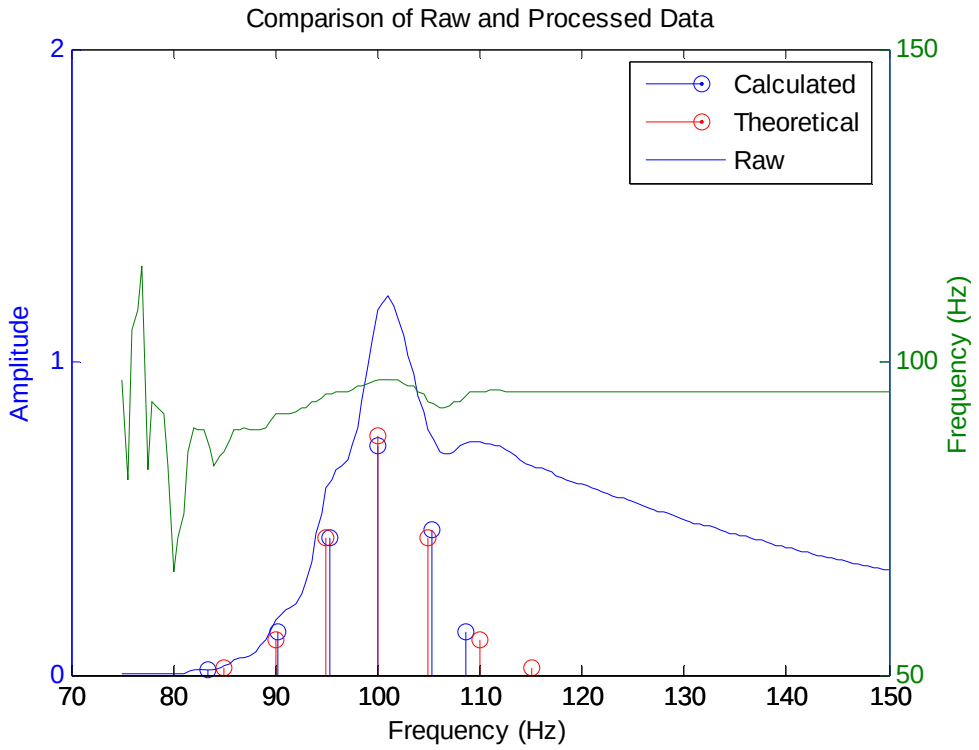

Figure 81. Comparison of raw vs. processed band data at t=0.30 seconds for a sinusoidally frequency-modulated sine. The continuous plots show band CF in Hz (X axis) vs. amplitude (blue, left Y axis), and frequency (green, right Y axis) of initial unprocessed data as determined by the pair of local maxima which brackets time *t*. The blue stems depict final processed values of frequency vs. amplitude for each component found using the iterative Simultaneous-Equation algorithm. (X axis for stem plots is no longer filter CF, as before, rather is processed frequency.) The red stems are the theoretical values of the sideband frequencies and amplitudes as determined by Table 30.

Note that at t=0.30 seconds, where the instantaneous frequency is expected to be at it minimum value of 95 Hz, the data in the higher frequency channels is close to 95 Hz. This is again consistent with the properties of the Peak-Locus of exponential filters. There is no direct indication of a relatively higher amplitude component at 100 Hz, other than a slight rise in the frequency estimates of the bands near 100 Hz CF; nor of a comparable sideband at 105 Hz, other than a slight dip in the frequency estimates of the bands near 105 Hz CF. This information is almost completely hidden. Only after processing the raw data via the algorithm does the existence of these components become apparent.

T=0.40 seconds, Amplitude threshold=0.6

The previous FM tests produced virtually identical carrier/sideband patterns at the two different times 0.30 seconds and 0.40 seconds, with the exception of some errors in high-order low-amplitude sidebands. A question of concern is why such similar behavior should be

observed at two separate times, despite the fact that the instantaneous frequency is expected to be different, i.e., 95 and 105 Hz, respectively, obtained by plugging each of these time points time into Equation 6.17. We wished to determine whether the results of these algorithms display truly local behavior, or whether they provide improved frequency resolution only at the expense of time resolution, and hence the results would not represent actual instantaneous values of parameters. We saw a similar situation in Section 5.8 where sharpening of the frequency response by naïve subtraction of adjacent filter channels caused loss of time dependency as shown in Figure 67. It was necessary to verify that such a phenomenon is not occurring here, as well. We therefore retested the same raw data using an amplitude threshold of 0.6, on the basis of results of the original test which indicated that the only component with amplitude above this level was the carrier, all sidebands were below. This would force the algorithm to consolidate all data into a single estimate. The results were:

Recov Freq	Recov Amp	Recov Phase
104.42	0.98	1.53

Table 32. Results for a sine-modulated FM signal at time 0.40 seconds using amplitude threshold of 0.6, which consolidates low-amplitude sidebands with the center carrier. Frequency is close to expected value of 105 Hz, and amplitude is close to expected value of 1.0.

This corresponds closely to the expected instantaneous frequency of 105 Hz, and amplitude of 1.0. Frequency is underestimated by a bit more than 0.5%, and amplitude by about 1.5%. Possibly, some of the error can be attributed to filter nonidealities, and some to the effect of an artificially high threshold on proper convergence. Nevertheless, the results appear to be consistent with local behavior, which is what we set out to determine.

T=0.30 seconds, Amplitude threshold=0.6

Repeating the higher amplitude threshold test at time 0.3 seconds, results were:

Recov Freq	Recov Amp	Recov Phase
95.24	0.99	2.77

Table 33. Results for a sine-modulated FM signal at time 0.30 seconds using amplitude threshold of 0.6, which consolidates low-amplitude sidebands with the center carrier. Frequency is close to expected value of 95 Hz, and amplitude is close to expected value of 1.0.

Frequency is within 0.25% of expected value 95 Hz; and amplitude is within 1.5% of expected value 1.0 for this time point. Results are thus again consistent with local behavior. In Section 6.7 we will consider the implications of these results in more detail.

6.5 Noise tests

Tests in additive white Gaussian noise were performed on a mixture of two sets of four harmonics with separation of 10 Hz in fundamental frequencies, at three values of signal to noise ratio: 10 dB, 0 dB and -10 dB. Noise levels were specified with respect to RMS value of sum of all 4 harmonics of both sets (8 components total). Results for each case follow.

6.5.1 10 dB SNR

With duplication threshold and amplitude threshold both set at 0.1, at time 0.5 seconds results were as in Table 34.

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
97.25	97.35	0.10	3.00	2.99	0.47	0.50	0.19	9.91
107.25	107.24	0.01	2.00	2.02	1.16	-0.50	-0.51	0.32
194.50	194.67	0.09	2.00	2.01	0.73	1.00	0.52	15.21
214.50	214.26	0.11	1.00	0.92	8.41	-1.00	-0.21	25.26
291.75	291.31	0.15	4.00	3.92	2.09	1.50	2.86	43.43
321.75	322.71	0.30	3.00	2.90	3.21	-1.50	1.77	95.81
389.00	389.18	0.05	2.00	2.01	0.48	2.00	1.40	19.15
429.00	430.34	0.31	1.00	1.04	3.62	-2.00	0.04	64.99

Table 34. Results for 4-harmonic sets at frequency separation 10 Hz at 10dB SNR.

Amplitude error was generally less than 0.1, and frequency error was up to 1.5 Hz (0.31%) at frequencies of signal components. Phase errors were higher, possibly due to strong frequency dependence of phase estimation procedure described in Chapter 5, in which phase is referred back to origin of time axis. Slight errors in frequency, therefore, compound into major errors in phase. Noise probably contributes directly to some of the phase error, as well.

In between the signal components are various noise components found by the algorithm. While noise is actually distributed across frequencies, the consolidation process leads to the appearance of discrete noise values spaced at various intervals between actual signal components, or a line-spectrum representation, as seen in Figure 82. The amplitude of these noise components at this SNR is generally between about 0.1 and 0.2, and thus makes them readily distinguishable from the stronger signal components. These values of noise magnitude may account for the amplitude error of signal components, as they are of roughly similar magnitude, but error is less than the noise level.

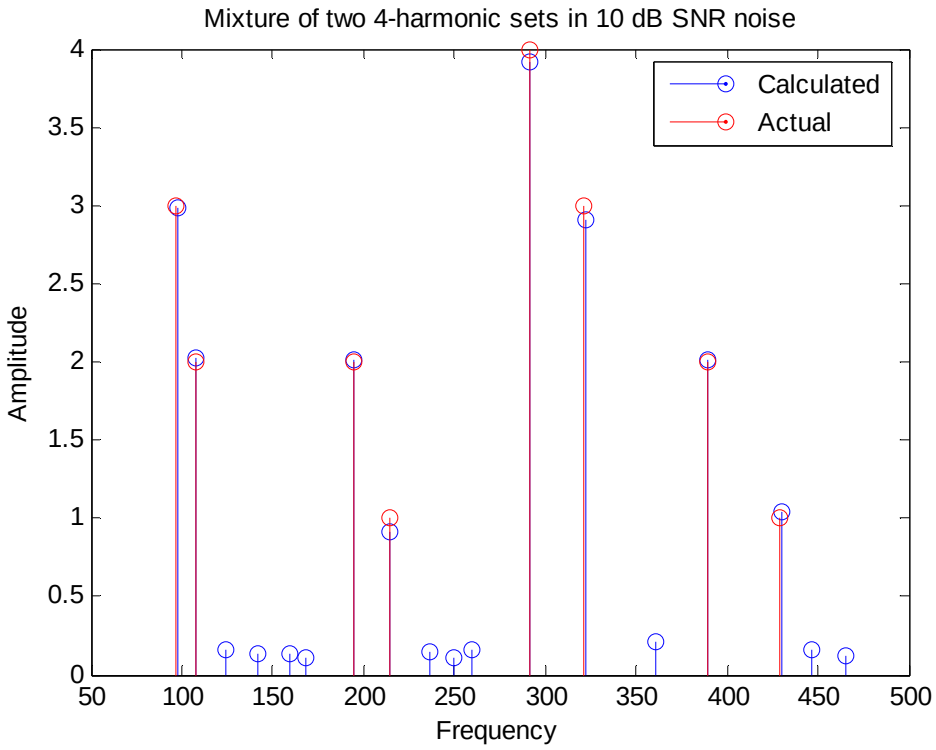

Figure 82. The actual (red) and calculated (blue) values of amplitude and frequency of all components found by the iterative Simultaneous-Equation algorithm for SNR of 10 dB.

6.5.2 0 dB SNR

To increase rate of convergence for this higher noise level the duplication threshold was adjusted to 2.5 with amplitude threshold held at 0.1. Numerical results are shown in Table 35.

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
97.25	97.93	0.70	3.00	3.14	4.77	0.50	-1.67	69.13
107.25	107.24	0.01	2.00	1.97	1.32	-0.50	-0.70	6.49
194.50	195.59	0.56	2.00	2.05	2.38	1.00	-2.27	95.94
214.50	212.34	1.01	1.00	0.75	25.02	-1.00	-0.36	20.47
291.75	291.33	0.14	4.00	3.67	8.20	1.50	2.75	39.88
321.75	322.40	0.20	3.00	2.78	7.18	-1.50	2.75	64.63
389.00	389.75	0.19	2.00	2.08	4.17	2.00	-0.55	81.32
429.00	432.15	0.73	1.00	1.08	7.81	-2.00	0.56	81.53

Table 35. Results for 4-harmonic sets at frequency separation 10 Hz at 0dB SNR.

Results are graphically displayed in Figure 83. In this case, the noise components appear at higher amplitudes, generally between approximately 0.3 and 0.5, with amplitude errors of signal components in a corresponding 0.25-0.35 range. Frequency errors are also increased and can range up to approximately 3 Hz. The error is most severe for the lower amplitude

components, and less than 1 Hz for components of amplitude 3 or greater. As we have seen in Section 5.4.3, peaks of greater amplitude components dominate those of lower amplitude components, and hence are more immune to being shifted by noise within any given band.

In an attempt to improve performance for this case, trials were rerun with amplitude threshold increased to 0.4, 0.5, and 0.6 in an effort to remove noise below these levels, thus possibly improving accuracy overall. These were not successful, and loss of actual signal components resulted. A possible explanation is that, as we have seen in the AM modulation examples, raw band data may initially be weak in certain bands before iterative processing. If these levels do not exceed threshold, they will be dropped, and prevent later processing from incorporating information from those bands. An alternate solution might be to manually set rank to the number of apparent signal components as determined from an earlier run, and force the algorithm to find best fit to this number of sinusoids. Instead of the 20 or so components that are currently generated, we could request only 8 at the outset. We have not explored this, although it may have potential to effectively average out some of the noise contribution, and seems worthwhile to pursue in the future.

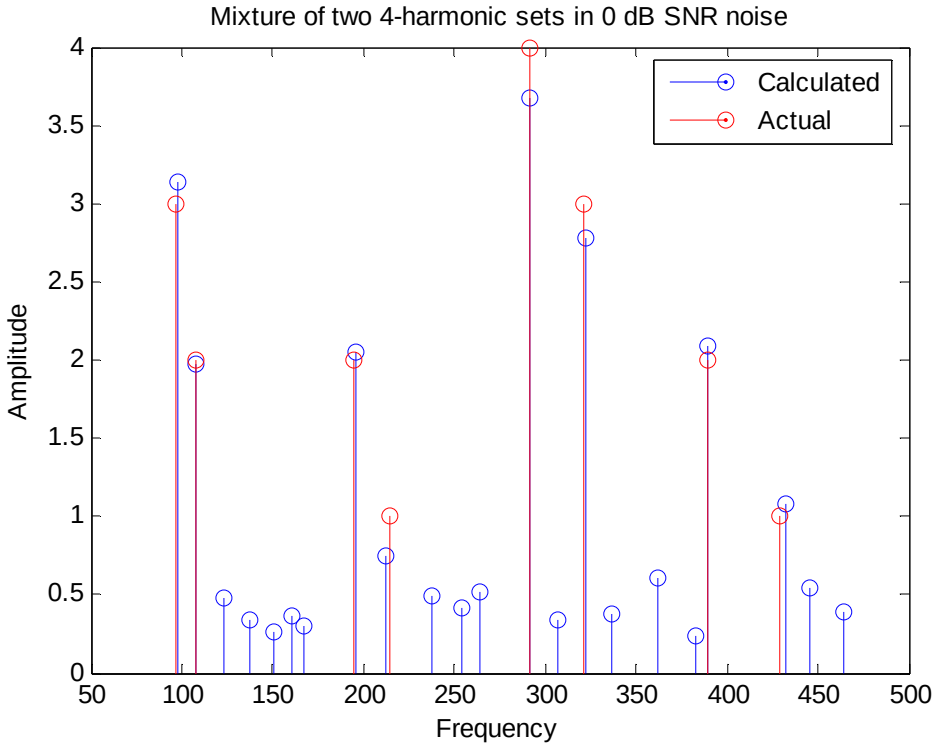

Figure 83. The actual (red) and calculated (blue) values of amplitude and frequency of all components found by the iterative Simultaneous-Equation algorithm for SNR of 0 dB.

6.5.3 -10 dB SNR

With amplitude threshold at 0.1 and duplication threshold at 2.5 , results were as in Table 36.

Orig Freq	Recov Freq	Pct Error	Orig Amp	Recov Amp	Pct Error	Orig Phase	Recov Phase	Pct Error
97.25	98.12	0.89	3.00	3.28	9.33	0.50	-2.30	89.17
107.25	107.18	0.07	2.00	2.27	13.42	-0.50	-0.83	10.53
194.50	194.62	0.06	2.00	1.67	16.31	1.00	0.92	2.54
214.50	203.01	5.36	1.00	0.86	14.49	-1.00	-0.46	17.28
291.75	290.38	0.47	4.00	3.03	24.14	1.50	-0.72	70.60
321.75	324.60	0.89	3.00	2.27	24.40	-1.50	2.16	83.44
389.00	389.18	0.05	2.00	2.40	19.95	2.00	1.03	30.79
429.00	430.76	0.41	1.00	1.23	23.44	-2.00	-1.63	11.90

Table 36. Results for 4-harmonic sets at frequency separation 10 Hz at -10dB SNR.

Results are graphically displayed in Figure 84. In this case the algorithm faired much worse. Noise components were of comparable magnitude to the signal components, and could not be readily distinguished. Even the higher level signal components were significantly affected in both amplitude and frequency. Errors in amplitude reached levels of approximately 1.0, and in frequency up to a few Hz. Worse was the fact that the signal component of 214.50 Hz seems to be missing altogether, and closest values were 203.01 and 238.99 Hz. Amplitudes of noise components reached levels of up to about 2.0, making recognition of signal components impossible without *a priori* knowledge of their existence. Clearly, thresholding would be of no assistance in this case, as noise components could exceed signal components in magnitude. However, if *a priori* information was available as to the number of signal components, possibly the approach suggested in the previous case could be attempted, where rank is set manually, and best solution for the given number of signal components is computed. As before, this has not been implemented in practice, and results are unknown.

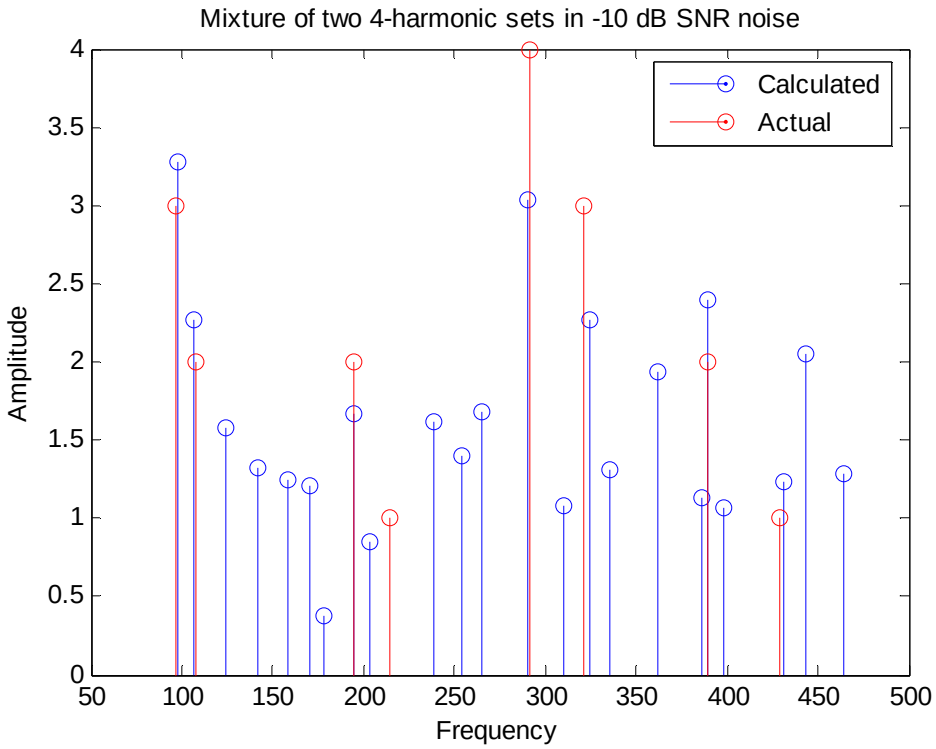

Figure 84. The actual (red) and calculated (blue) values of amplitude and frequency of all components found by the iterative Simultaneous-Equation algorithm for SNR of -10 dB.

6.5.4 Comparison of Noise Test Results with FFT

For comparison, we include the graphical FFT results for each of the above three noise levels along with the clean, no-noise case in Figure 85. As before, the two fundamentals were separated by 10 Hz.

For reasons described in Section 6.3.2, the FFT does not produce correct estimates on the non-integral frequencies used in these tests, even with no added noise. Frequencies are estimated as nearby integral values, rather than as the correct, fractional values. Amplitude estimates are even worse, with errors of up to 35%. However, at the highest noise level of -10 dB, using the Simultaneous-Equations algorithm it is difficult to distinguish noise components from signal components, and signal components themselves are greatly impacted. As this was only an initial run, and with the foreknowledge that signal components were separated by more than 10 Hz, we selected a wide duplication threshold of 2.5 Hz, to speed convergence. However, it is possible that this consolidated noise energy into a few large-amplitude components which were on a par with signal components, rather than representing noise as large numbers of small-amplitude components which could more readily be distinguished from the signal components.

Runs should be repeated at several more combinations of threshold values to study this possibility.

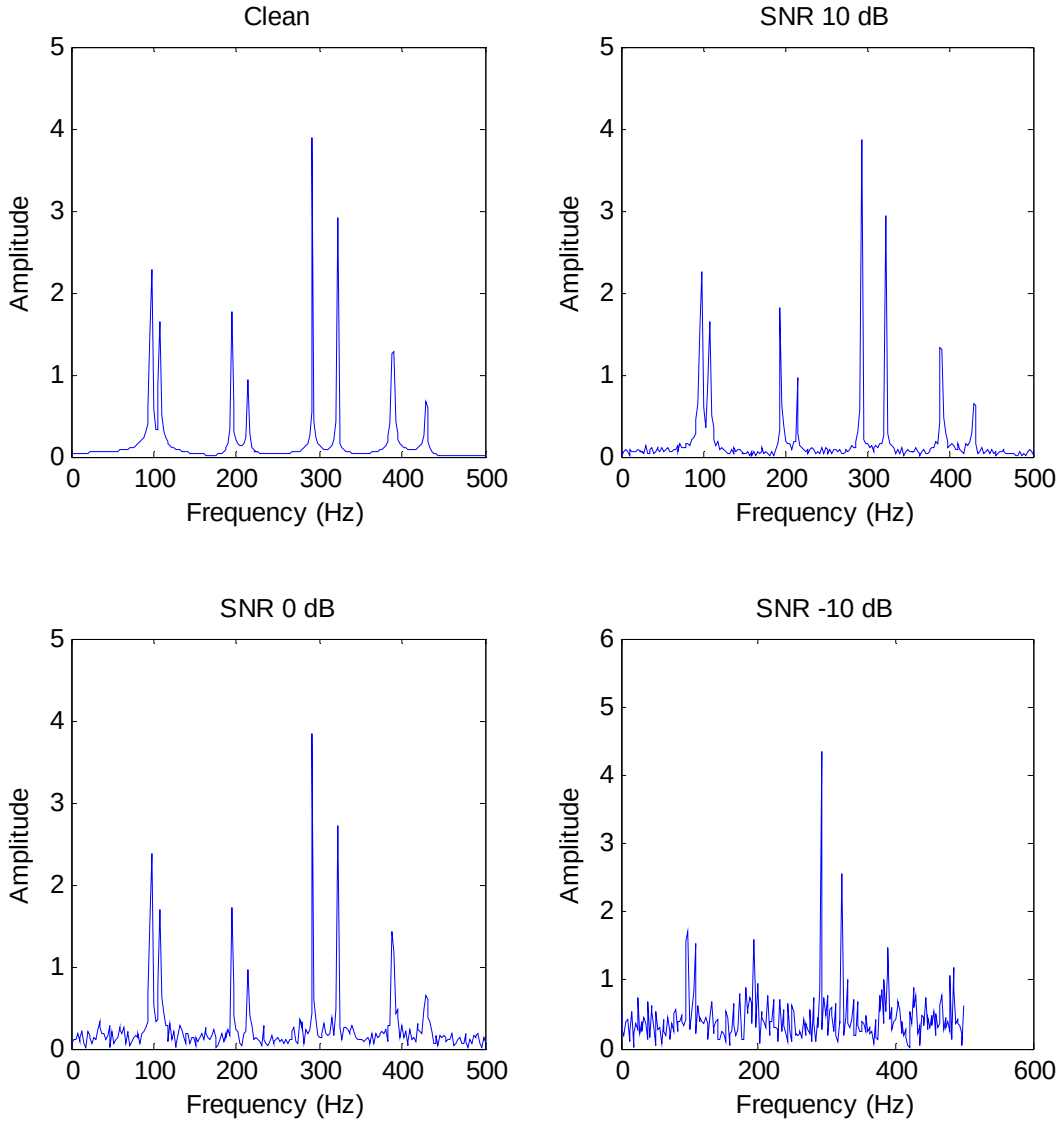

Figure 85. FFT results for the two 4-harmonic sets at each of 3 noise levels and with no noise for Delta 10 Hz. At lower noise levels, the Simultaneous-Equation algorithm yields better accuracy, as FFT exhibits errors in amplitudes by up to 35%, and errors in frequency of up to 0.77%. At higher noise levels, the FFT seems to be more robust. However, further tests are necessary with the Simultaneous-Equation algorithm to determine if results could be improved by varying the amplitude and duplication thresholds.

6.6 Speech Tests

6.6.1 Male Speech

We analyzed the recording of male speech from Chapter 5 continuing with same test parameters as previously, except where noted. Analysis at time 0.35 seconds with duplication threshold 0.25 and amplitude threshold 0.0005 produced the following results shown in Figure 86:

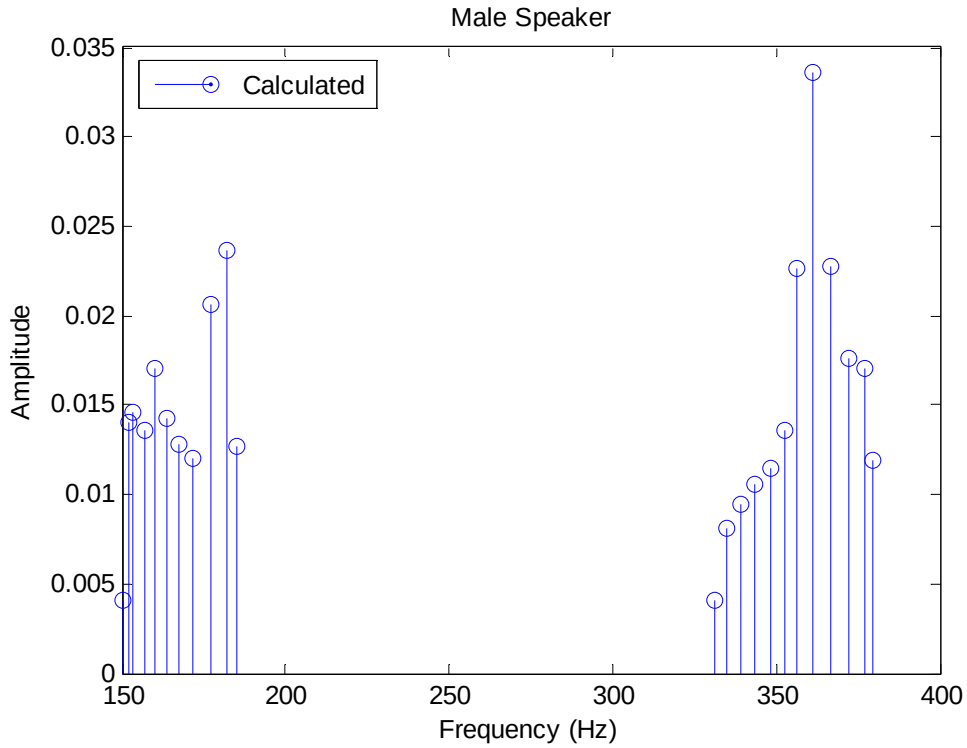

Figure 86. Frequency and amplitude of spectral components found by iterative Simultaneous-Equation algorithm in analysis of male speaker at time 0.35 seconds. Components are believed to represent the first and second harmonics and associated sidebands.

There appears to be a series of components extending from about 155 Hz to about 185 Hz, some of which are regularly spaced, and others which are not. At 182.1 Hz, there is a component of greater amplitude which appears to be the fundamental, as it has a counterpart 2nd harmonic. Possibly some of the surrounding components are sidebands due to AM and/or FM modulation. There is a component at 361.47 Hz which appears to be the second harmonic of the 182 Hz component, and is also surrounded by lower valued components which may possibly be sidebands. The 2nd harmonic is not exactly an integral multiple of the fundamental, but this has been observed by other authors, as well (Naylor and Boll, 1987).

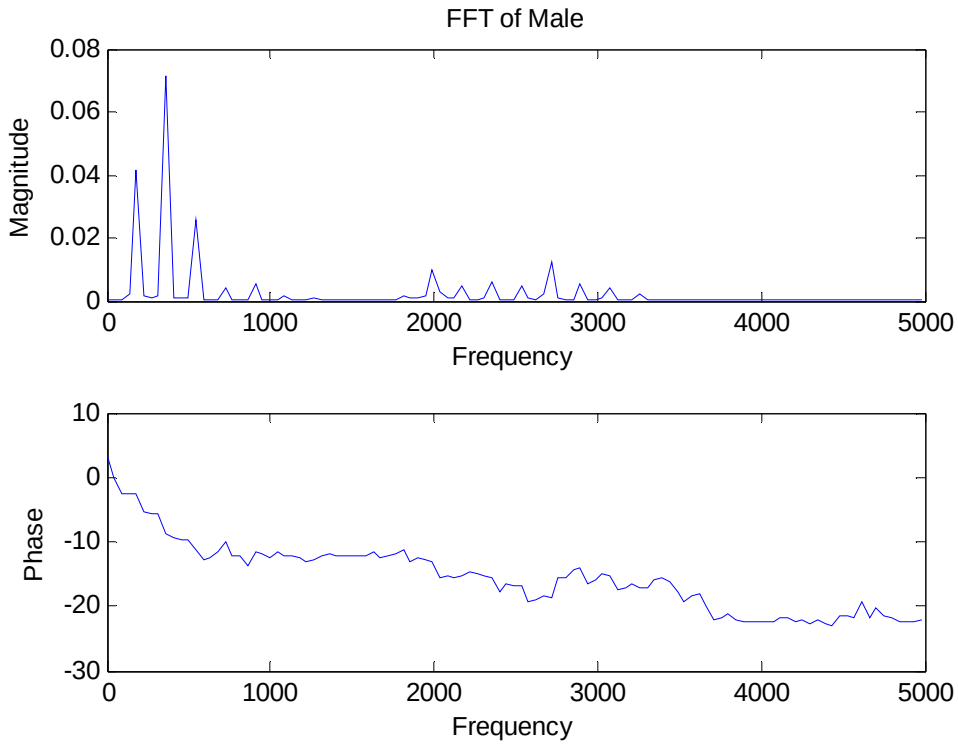

Figure 87. FFT of 4-cycle segment from the male speaker. Estimates for first two harmonics are 180.96 Hz and 361.99 Hz, respectively.

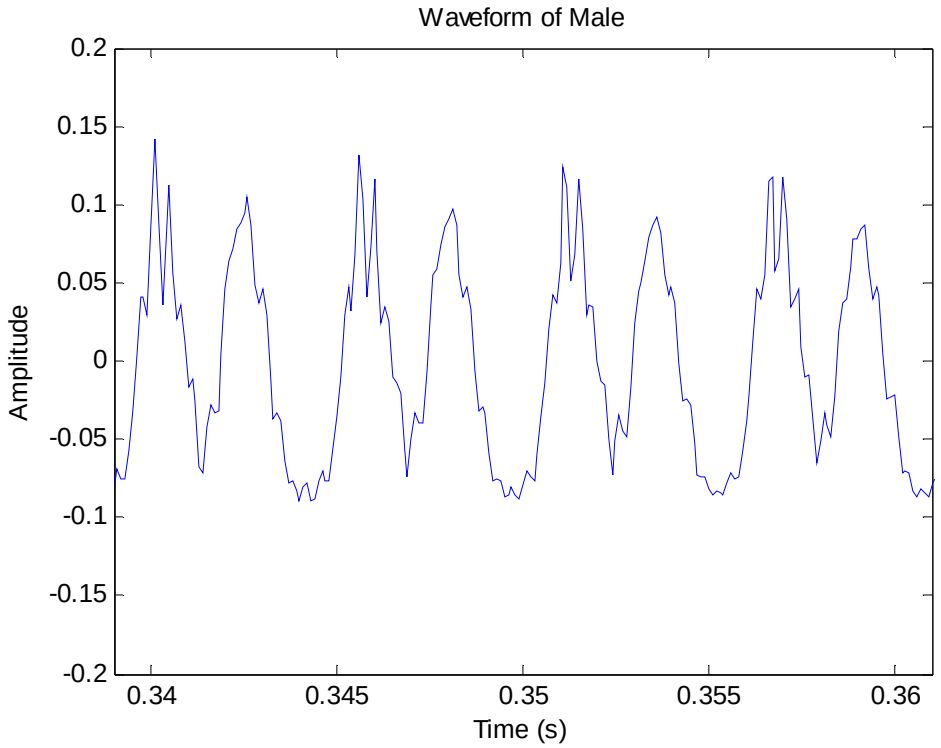

Figure 88. The time waveform of male speaker from which FFT was computed. Duration was 0.022 seconds, enclosing 4 cycles of the waveform.

Examination of the FFT shown in Figure 87 of a 4-period segment of the waveform shown in Figure 88 and centered on the point of analysis at 0.35 seconds into the signal, produces estimates of 180.96 Hz and 361.99 Hz for the first and second harmonics, which differ with our estimates by 1.15 Hz and 0.52 Hz respectively. We note that varying the time duration of the segment from the 0.220 seconds we used, even slightly (by +/- 0.0001 seconds), alters the FFT estimate noticeably. This points out an advantage of our methods, as they do not require a particular window length in relation to the signal, however, there may be some variability in our methods depending on choice of thresholds.

There is a possibility that some of the spread in the estimated frequency components surrounding the fundamental may be due to pseudo-periods in which energy from a large-amplitude higher order harmonic (possibly close to the second formant) is great enough to shift the peak location of the fundamental so that it becomes difficult to distinguish the true period from the pseudo-period. As can be seen in Figure 88, the proximity of the two sharp peaks at 0.3456 and 0.3460 seconds and again at 0.3511 and 0.3515 seconds make it difficult to determine which represents the true reference point of the cycle. If the outer peaks are used, the estimate becomes 169.5 Hz, while if the inner peaks are used, the estimate becomes 196.1 Hz. The average of these estimates is 182.8 which is very close to our estimated fundamental, and the larger and smaller values may possibly account for some of the spread of energy.

In general, the spectral lines produced by the algorithm are difficult to interpret, and require further tests and analysis to determine their origin, whether due to noise, or to processing artifacts or whether they are sidebands due to modulation of the speech signal. The finding of asymmetric sidebands of uneven height would be consistent with discussion in (Hartmann, 1998) regarding mixed modulation. The author demonstrates that because of phase interactions among AM and FM sidebands, one side may be boosted while the other is attenuated.

6.6.2 Female Speech

Tests were run on a corresponding segment of the female speaker from Chapter 5, with all settings identical to the test on the male speaker of 6.6.1.

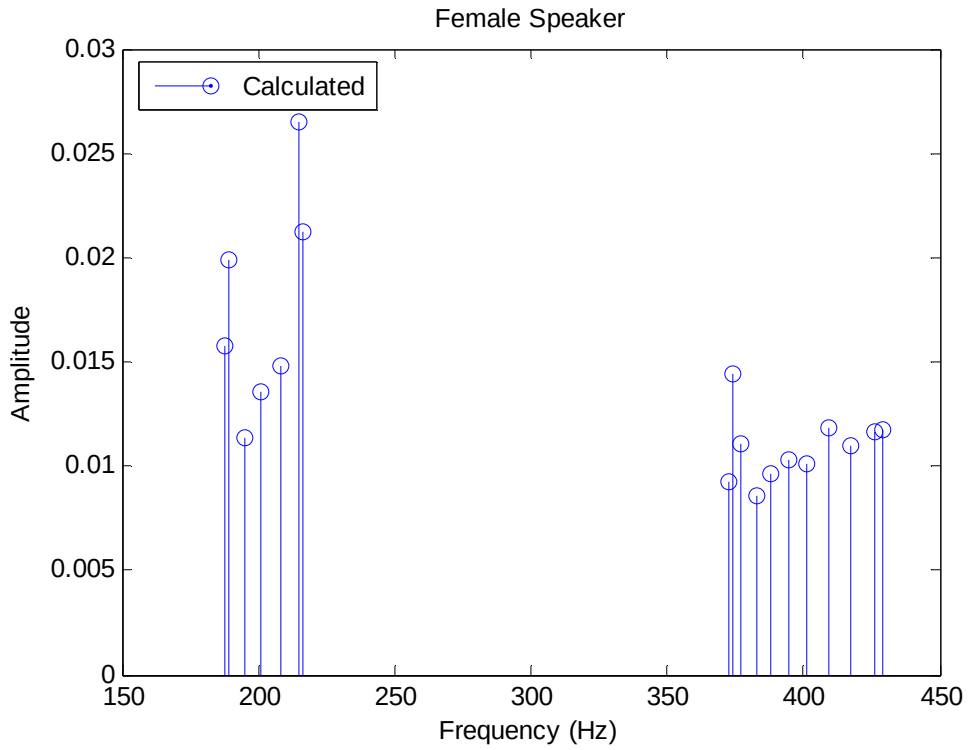

Figure 89. Frequency and amplitude of spectral components found by iterative Simultaneous-Equation algorithm in analysis of female speaker at time 0.35 seconds. Components are believed to represent the first and second harmonics and associated sidebands.

Figure 89 illustrates graphically. There appears to be a fundamental of 208.19 Hz surrounded by a series of components of both regular and irregular spacing and variable height. Although the component we have identified as the fundamental appears to be of lower amplitude than the higher frequency component immediately adjacent to it, nevertheless this can happen with certain types of FM modulation, as we saw in Table 30.

There similarly appears to be a second harmonic of frequency 417.36 Hz surrounded by an assortment of spectral lines. The second harmonic is close to but not exactly twice the fundamental. We again note that other methods have similarly found that in actual speech waveforms, relationships between component frequencies are not perfectly integral (Naylor and Boll, 1987).

The FFT of a 4-cycle segment of the female speaker is shown in Figure 90. The waveform, again centered at 0.35 seconds, the point of analysis, is shown in Figure 91. The duration chosen is now 0.0191 seconds to fit the shorter pitch periods of the female voice. The FFT estimates the fundamental to be 208.33 Hz, and the second harmonic to be 416.66 Hz. These differ from our

estimates by 0.14 Hz and 0.70 Hz, respectively. This is in very close agreement to our methods, and appears to confirm the accuracy. However, because the FFT does not resolve surrounding spectral lines, we cannot make a definitive comparison with our method.

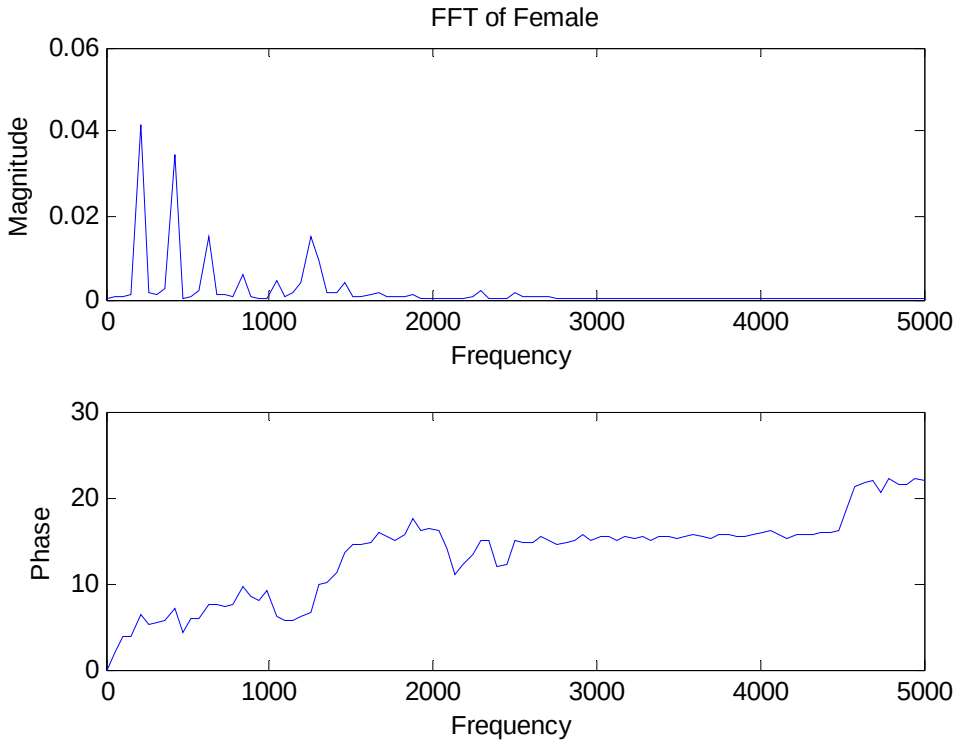

Figure 90. FFT of 4-cycle segment from female speaker. Estimates for first two harmonics are 208.33 Hz and 416.66 Hz, respectively.

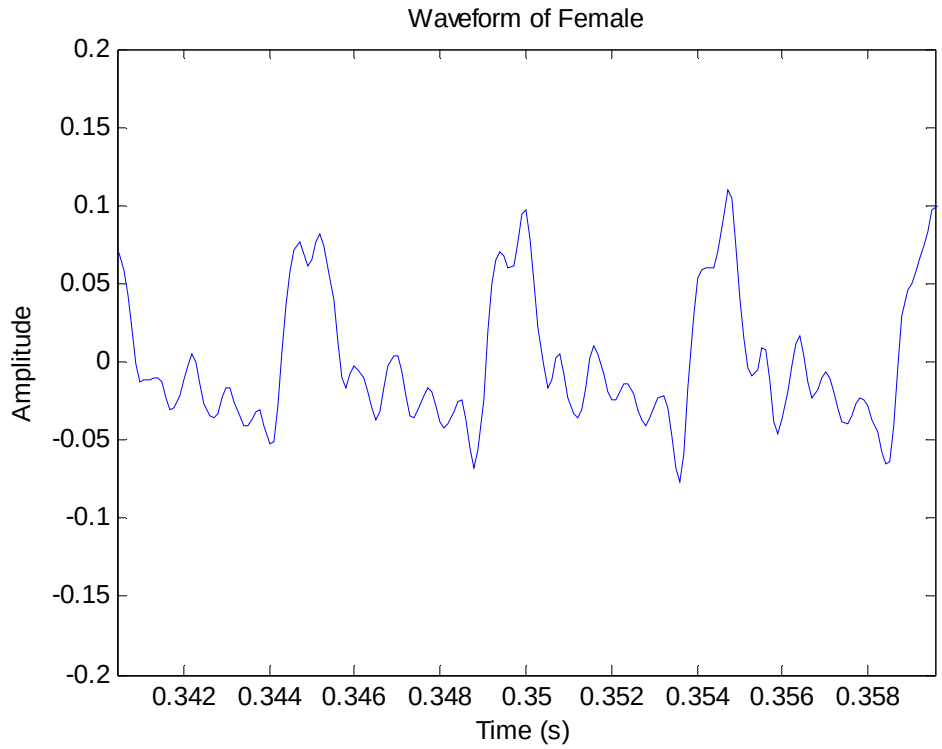

Figure 91. The time waveform of female speaker from which FFT was computed. Duration was 0.0191 seconds, enclosing 4 cycles of the waveform.

We again note that there are possibly pseudo-periods caused by strong higher harmonic components which may produce a spread of energy. Using the inner pair of peaks at 0.3494 and 0.3452 seconds gives a frequency estimate of 238.10 Hz, while using the outer pair of peaks at 0.3447 and 0.3500 produces an estimate of 188.68 Hz. Using the first peak of both cycles yields an estimate of 212.77 Hz, while using the second peak of both pairs yields 208.33, which matches the FFT estimate. Regardless of whether this observation has significance, it is clear that even on short time scales of less than a cycle there is frequency variation, which may also account for the spread of spectral lines. We remind the reader that the algorithm does not actually look at the peaks of the summed waveform, as shown here, but rather at the peaks of the individual channel waveforms. These are bandpass-filtered so that interference from a distant harmonic is less likely.

6.6.3 Mixed Speech

Using the same parameters and time point of analysis as in the male and female cases of Sections 6.6.1 and 6.6.2, results were as shown in Figure 92.

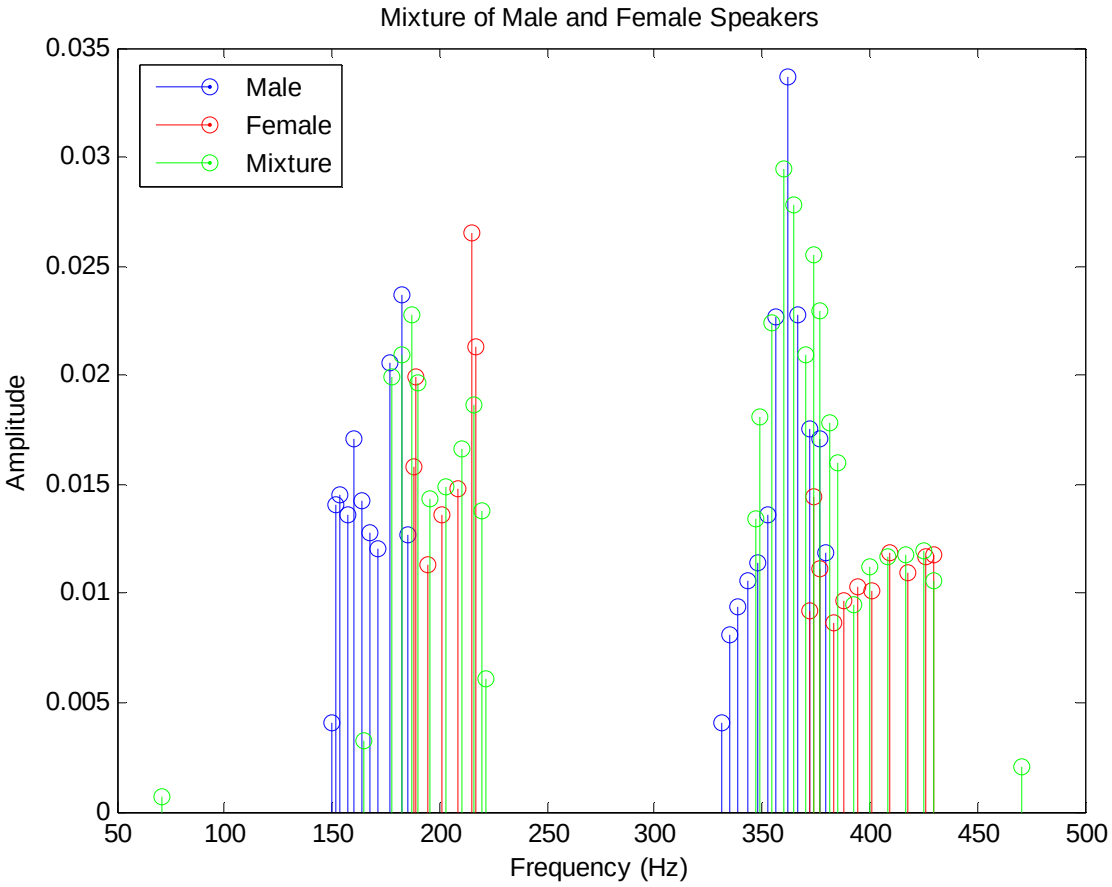

Figure 92. Frequency and amplitude of spectral components found by iterative Simultaneous-Equation algorithm in analysis of mixture of male and female speakers at time 0.35 seconds. Components are believed to represent the first and second harmonics and associated sidebands. The extracted components of the male speaker from Figure 86 and the female speaker from Figure 89 are replotted here in blue and red, respectively, for comparison with components extracted from mixture which is shown in green.

The mixed speech case is difficult to interpret, as the spectral line pattern does not appear to be a simple union of the male and female spectral lines, as one might initially expect. Some of the lines are recognizable from the plots of the individual speakers, but others have shifted or disappeared. There is a component at 182.1 Hz which appears to be the fundamental of the male speaker, with frequency exactly as measured in the single speaker case, although amplitude is now slightly less. Proximal frequency components appear to have shifted slightly.

There is a component at 210.8 Hz which appears to be the fundamental of the female speaker, but has shifted from the value of 208.19 Hz that was measured in the individual case. Possible explanation may be due to interference from a nearby male component. However, in this region the energy of the male appears to be relatively low, and would seem less likely to interfere with

measurements of female speech parameters. Further investigation is needed to understand this shift.

There is a strong component of 359.71 Hz which appears to be the second harmonic of the male. However, this second harmonic and its proximal components have now shifted down approximately 2 Hz from the values measured in the individual male case. The explanation is unclear.

There is a strong component at 416.41 which appears to be the second harmonic of the female speaker, and is consistent with only a nominal shift from the value of 417.36 Hz with similar neighboring components as measured in the individual female case.

In general, there appears to be some loss of spectral components in the analysis of the mixture which appeared to be present in the analysis of the individual recordings. There is also some shifting of values. However, for other spectral lines, there is close correspondence between the individual and mixture components.

The FFT does not resolve separate peaks for the two speakers, but merges first harmonics of both into a single spectral peak, and second harmonics of both into a single spectral peak. The maximum values of these peaks occur at 156.25 Hz and 364.58 Hz, respectively. These lie closer to the corresponding values of the male, than the female, but are in error by 24.71 Hz and 2.59 Hz, respectively. The duration used of the mixed segment was 0.0191 seconds, as in the female recording. However, being that the mixture is aperiodic, it no longer encloses an integral number of cycles. As before, modification of the segment duration drastically alters these estimates. For this reason, it is very difficult to consider these estimates produced by the FFT to be meaningful, in any way. The individual spectral lines found with our method did not suffer such large shifts, in a sense being more robust under mixed conditions. However, there were nontrivial losses of some lines, and one cannot clearly see from the line clusters of Figure 92 a clear demarcation between the two speakers that identifies it as being from a mixture as opposed to a single source. However, one characteristic that is noticeable is that in both the first and second harmonic clusters, the sidebands of the female are more widely spaced, whereas the male sidebands are more tightly spaced. The fact that this is true for both the first and second harmonics, may be a useful manifestation of comodulation, that commonly modulated

harmonics have common sideband patterns. We will discuss this further in Section 6.7. While we can only speculate without further trials and analysis, we note that some of the character of the overall shape of the line spectra in the individual cases seems to be preserved in the mixture, and that this may be useful in separation. This is in contrast with the mixed FFT results of Figure 93, where no clear trace of the individual FFT results is evident.

These tests represent only initial work on actual speech recordings, and more refinement is needed to find optimal test parameters and thresholds, and to determine whether accuracy can thus be improved. It is very difficult to interpret these preliminary findings, and we can do no more than speculate. It is unclear at this time whether the components we have found have any physical meaning, or are merely due to peculiarities of the algorithm. Further testing is needed on mixtures of doubly-modulated AM and FM signals where the behavior of each signal is known independently.

We also note that due to computational limitations, we are forced to use a filter separation of 0.5 Hz. Possibly, finer separation will produce better results. We did not repeat these tests with other settings due to the heavy computational burden involved. In the ideal situation, one would seek to find a combination of settings that would consolidate the sidebands of each harmonic into a single value, as we were able to do in the earlier modulation tests in this chapter. In the case of a mixture, one would then hope to find two distinct values each of the fundamental and second harmonic frequencies, one for each speaker, thus clearly identifying the waveform as a mixture.

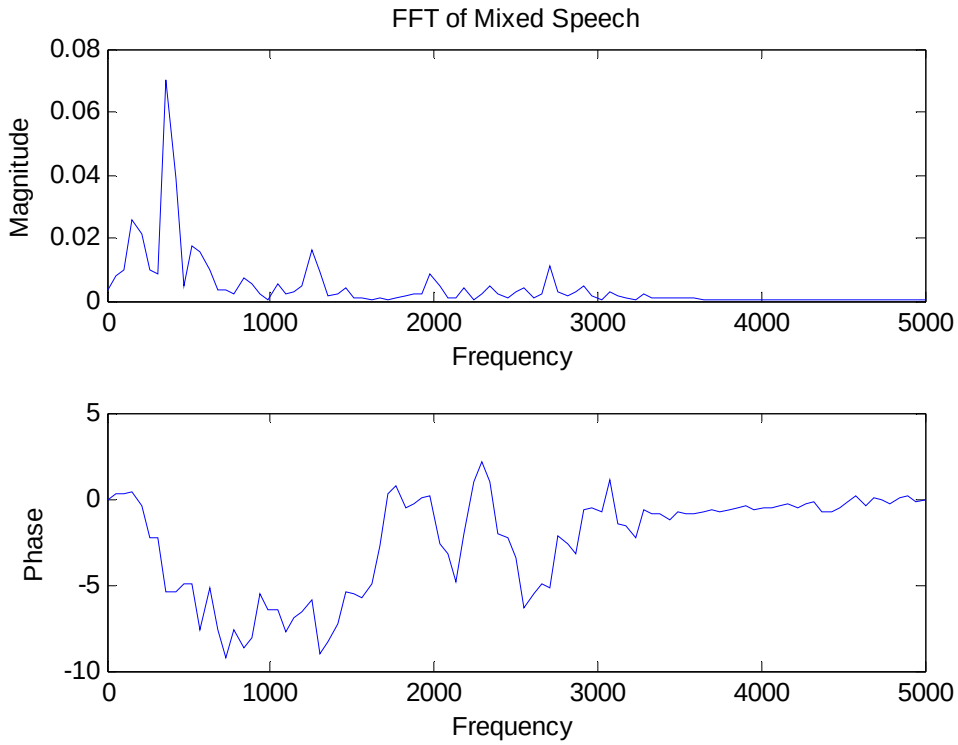

Figure 93. FFT of mixed speech segment. Estimates for first two spectral peaks are 156.25 Hz and 364.58 Hz, respectively. These do not correspond to those of either the male or female speaker.

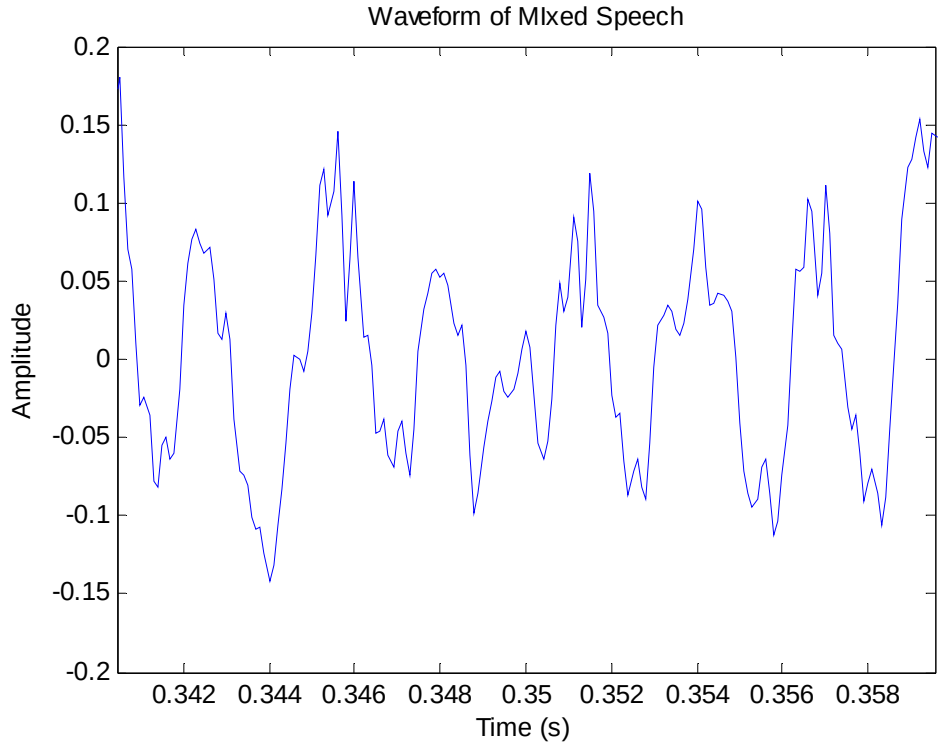

Figure 94. The time waveform of mixed speech from which FFT was computed. Duration was 0.0191 seconds, as in the case of the female recording, but due to aperiodicity of mixture, does not enclose an integral number of pitch periods.

6.7 Analysis

6.7.1 Instantaneous Frequency

The picture that emerges from the tests in this chapter is complex. On one hand, we have seen behavior that is consistent with the notion of instantaneous parameters, such as in the ramped modulation tests, where the values of amplitudes at points in the AM ramp, or values of frequency at points in the FM ramp were exactly on target for the respective time points, based on the formula used to generate the original signals.

However, for the case of sine-modulated AM and FM signals, the algorithm found sidebands that appeared to conform nicely to theoretical expectations, but were almost identical at different time points, seeming to lack time dependency. We did show that by consolidating data from multiple channels together by the use of higher thresholds, it was possible to obtain a single value that had local properties. Nevertheless, we need to address why the behavior in the case of ramps appeared to be instantaneous in nature, whereas the behavior in the case of sine modulation appeared to be time-independent

It appears that there is a duality in describing signals which is not due to the uncertainty principle. It is rather due to a simple trigonometric identity.

As an example, the sinusoidally modulated AM signal had the form:

$$\begin{aligned} x &= [1 + \sin(a)][\sin(b)] \\ &= \sin(b) + \sin(a)\sin(b) \end{aligned}$$

Using the trigonometric identity

$$\sin(a)\sin(b) = 1/2\cos(a - b) - 1/2\cos(a + b)$$

we have:

$$\begin{aligned} x &= \sin(b) + 1/2\cos(a - b) - 1/2\cos(a + b) \\ &= \sin(b) + 1/2\cos(b - a) - 1/2\cos(b + a) \end{aligned}$$

since $\cos(\xi) = \cos(-\xi)$.

Therefore, a modulated sinusoid is always equivalent to a sum of carrier and sidebands. This is true at all times, and at any one instant of time. The question of what we mean by instantaneous frequency can never be answered definitively, as it is a matter of preference whether we want to think in terms of the sum of constant components, or one time-varying component. In the case of the sine-modulated signals, the sidebands are sufficiently far from the carrier that they are separately resolved. The algorithm's preference is to treat everything as separate, unless a component's separation is below the duplication threshold or beyond the algorithm's limits of resolution. (That limit is hard to quantify, as it depends on the shapes and spacing of the filters and the sampling rate, and we do not have a full understanding of all the factors that affect its ability to converge. Nevertheless, in our current version, we have seen that it can resolve frequencies of about 1 Hz separation, although there is nontrivial amplitude error at such close spacing.) However, for the case of the ramp, the fundamental modulation rate is lower, and sidebands are more closely spaced. In addition, a ramp has multiple harmonics in its own Fourier transform and these each contribute sidebands above and below the carrier. This is true in both AM and FM cases. For this reason, most components are not resolved, and are consolidated into a single value. They thus behave as a single signal with time-varying properties. This might contribute to the different behavior in the ramp case.

We note that there have been at least two recent papers that have examined the use of the traditional definition of instantaneous frequency in the case of mixtures (Nho and Loughlin, 1999), (Oliveira and Barroso, 1999). Both came to the same conclusion that errors will generally occur when signals are not monocomponent, but that there is an exception. We cite the following from Oliveira and Barroso:

> "The concept of instantaneous frequency IF_t, (where the subscript emphasizes a possible time dependency) is of paramount importance in fields such as sonar, radar, communications, and medicine. During the last few decades, this concept has become intimately associated with the derivative of the phase function. For a general complex signal $z(t) = a(t)e^{j\phi(t)}$, this traditional definition states that $IF_t = \phi'(t)$; for real signals, it defines IF_t as the derivative of the phase function of the associated analytic signal (L. Cohen, 1995), (Ville, 1948).

> "It has been known that this traditional definition does, in some cases, provide physically unacceptable results (L. Cohen, 1995), (Loughlin and Tacer, 1997), (Mandel, 1974). In fact, as we will show, this definition will provide unacceptable results in all but a few special cases. In the general case, $\phi'(t)$ is void of any physical significance and should not be identified with IF_t."

The "few special cases" that the authors found are where there are certain types of symmetry among the amplitudes and frequencies of pairs of components. The frequencies need to be spaced equally about a central frequency, and the amplitudes of each member of a pair need to be equal. From what we have seen in both AM and FM cases, this is satisfied by the sidebands. In other words, the spectral components or harmonics of a signal can be further broken down into microcomponents consisting of a carrier and sidebands for each. The microcomponents of each harmonic will each behave properly if grouped together and the phase derivative is calculated. One therefore has a choice whether to describe each component as one time-varying component, or a set of microcomponents. We conjecture that for the purpose of separation, it may be more effective to break down into microcomponents. One isolates the microcomponents of each harmonic, and then separates from the mixture. To compute the pitch track, one then takes the flip side perspective, and adds these microcomponents together. One can then take the phase derivative to obtain a physically meaningful value of the pitch at any point.

We further note a very interesting point. The essential reason why the authors consider the phase derivative definition of IF to be "unacceptable" and "void of any physical significance" in most cases of mixtures is the fact that it produces estimates which may be "outside the known spectral range of the signal." In other words, estimates can be lower than the lowest or higher than the highest frequency of the mixture. As we will show in Chapter 7, the peak-based estimation that we have been using has exactly the same property. For this reason we use it only for an initial estimate, but iterate until we purify each component into a monocomponent state. Based on the similar behavior of the two definitions, we can ask whether they are not one and the same. In analogy with the definition of instantaneous velocity, we note that the goal is to obtain a number which describes the rate of change of position at a single instant. However, at any actual instant of true zero duration, the position will not change, as one cannot go anywhere in zero time. The solution is to use a limiting process and calculate how far one could

go in shorter and shorter intervals of time and to define the ratio of distance covered to time elapsed in the limit as elapsed time approaches zero, as the instantaneous velocity. In the case of signals whose components are known explicitly as in the algebraic examples in the paper, then the phase derivative may represent such a limiting process. However, in cases where they cannot be described mathematically, but only by observation, the shortest interval over which one may be able to reliably measure the rate of change of phase may be the interval between two peaks, which although strictly is analogous only to an average velocity, but nevertheless, might be the finest-grained approximation that is directly measurable. This may account for the similarity in behavior.

Based on this, we can point to a weakness in the reassigned spectrogram which we reviewed in Chapter 2 in that it uses the phase derivative to reposition components along the frequency axis. In the case of mixtures or overlapping components, this repositioning may cause components to overshoot the mark, and give physically meaningless results. For this reason, it does not handle interference properly, as the authors note.

Upon further consideration, one might indeed ask what, exactly, does one expect to obtain by consolidating the frequencies of two disparate components into one number, and why would one ever consider that a reliable or useful representation of the behavior of such a mixture. In our work, we consider estimates of instantaneous frequencies of mixtures obtained from local maxima to be A) only starting points for analysis, and B) only to be used in conjunction with similar estimates from other bands in order to calculate the parameters of the underlying components. We would never consider a single such estimate as a reliable indicator for any of the parameters. It seems quite logical to us that the phase derivative should be treated no differently in this respect.

6.7.2 Comodulation and Sidebands

Based on some of the results we have seen in the speech tests, we now have a new perspective on comodulation. In the frequency domain, components that are comodulated will share similar sideband patterns. One could then group clusters of components with similar sideband distributions together, and separate from components with different distributions. Since there are multiple sidebands, even if some are overlapped by an interfering microcomponent from another source, they may be recoverable on the basis of their neighbors, as they should share

similar spacing. We can only speculate, as we have not gone this far in our work, and leave the reconstruction problem for the future. We again emphasize that the distribution of sidebands is not due to uncertainty or blurriness from modulation. It is due to the trigonometric identity at the beginning of this section which allows a dual perspective on signals. Any blurriness occurs from using techniques that cannot definitively resolve the sidebands from the carrier, and hence the carrier appears as a single, broadened peak.

A final question is how to distinguish between sidebands and noise components. From what we have seen, sidebands generally appear to have a more regular spacing and amplitude progression, while noise components are spaced more randomly, and do not have an amplitude progression. This needs to be confirmed with further tests.

We conclude this section with an observation regarding the difference between modulation and beating caused by constructive and destructive interference, that we first discussed in Chapter 3. In addition to the comodulation approach for distinguishing the two we suggested there, we might also note that analysis of both AM and FM signals seems to lead to symmetric spectral components, as we have described, whereas analysis of interference from unrelated signals will generally have no such relationship, other than by coincidence. However, our limited speech tests did not always find this symmetry in actual signals, and it is not clear whether it is due to imperfect resolution of the algorithm, or whether it is due to noise, or whether it is simply not necessarily an inherent property of all modulated signals to begin with. Tests on doubly-modulated AM-FM signals might shed more light on this.

6.8 Transient Response

With the understanding we have gained from the trials in this chapter as to how the algorithm functions in various test environments, we are in a position to offer a few thoughts on the effect of the transient response of individual exponential filters on Peak-Locus plots. It is well-known that the output $y(t)$ of a linear, time invariant (LTI) filter $h(t)$ to an input $x(t)$ is represented by the convolution integral

$$y(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau \tag{6.18}$$

The interpretation is that we resolve the continuous signal $x(t)$ into successive, infinitesimally short impulses each of whose area is the amplitude of the signal at that instant, and then for each impulse, the output is found by multiplying $h(t)$ times the area of that impulse. To obtain the response to the entire signal, by linearity, we simply add the responses to each impulse, staggering them in time appropriately. It is also well-known that the set of complex exponentials $A_i e^{j\omega_i t}$ (where we have subsumed the phase term into the complex amplitude A_i for convenience) are eigenfunctions of linear filters, meaning that the response to one of these signals is simply the original signal multiplied by a complex constant. This may be shown using the commutative property of convolution which follows from (6.18) by a simple change of time variables

$$y(t) = \int_{-\infty}^{\infty} x(t-\tau)h(\tau)d\tau \tag{6.19}$$

Applying $Ae^{j\omega t}$ as the input to such a filter, we obtain

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} Ae^{j\omega(t-\tau)}h(\tau)d\tau \\ &= Ae^{j\omega t}\int_{-\infty}^{\infty} h(\tau)e^{-j\omega\tau}d\tau \\ &= Ae^{j\omega t}H(\omega) \end{aligned} \tag{6.20}$$

where $H(\omega)$ is the Fourier Transform of the filter impulse response $h(t)$, and is known as the frequency response of the filter. Since integration is in the dummy variable τ, the t-dependent terms are effectively constant and can be moved out of the integral.

The frequency response $H(\omega)$ (which is the corresponding eigenvalue of the system) may alter the amplitude and phase of the input signal, respectively, with its magnitude multiplying the input amplitude, and its angle adding to the input phase, but the frequency of the input cannot be altered. By linearity, the response to a sum of such inputs is the sum of the individual responses to each input. This is the basis for the application of Fourier analysis to linear systems.

It is commonly believed that the preceding frequency domain analysis only holds true when the system is in steady state, i.e., after the settling of all transients. Indeed, if one tries convolution

in the time domain between two sinusoids of differing frequency, one would expect to get zero output response according to Fourier analysis. The reason is that real sinusoids are representable as a pair of impulses in the frequency domain at the corresponding points on the frequency axis, one each for the positive and negative frequencies. Since convolution in the time domain is equivalent to multiplication in the frequency domain, the result should be zero everywhere, being that the positive and negative impulses of these two signals do not coincide, and everywhere else along the frequency axis the value is zero, hence one would expect that each signal would multiplicatively zero out the impulses of the other. However, in practice, this does not happen, and there is a finite-length non-zero response at the beginning and end of the output signal. Figure 95 shows a graphical demonstration of the situation for a 4-Hz filter and 5-Hz signal.

It seems that the usual way of viewing such behavior within the signal processing community is to reason that during the sliding operation of convolution, the filter output cannot reach steady state until its entire width is filled with the input signal. The logic would seem to be that in steady state, both halves of the filter should be balanced, so that any contribution to the time integral from one side is properly offset by an equal and opposite contribution from the other side. If this does not occur, and only the right hand side is within the non-zero part of the signal while the left hand side is not, then there may temporarily be a much greater (or lesser) output than is warranted, due to interaction between the algebraic signs of the contributing regions within the integral. This viewpoint views the problem as lying within the filter itself (although it is unavoidable). Common terminology such as settling time, or the need to "warm up" the system reflects this viewpoint. While this is certainly correct, for our purposes it gives little insight to the behavior we seek to understand.

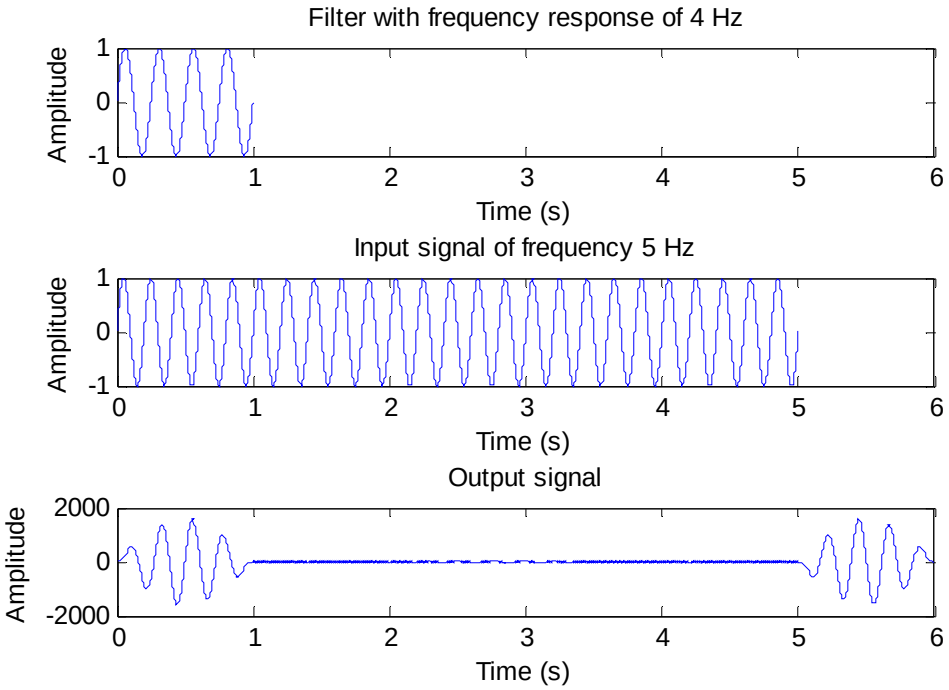

Figure 95. A graphical illustration of the effect of transient response on the convolution of two sinusoids of differing frequencies. The top signal, viewed as filter, has response of 4 Hz. The middle signal, viewed as input, has response of 5 Hz. The bottom signal is the output. Note non-zero transient response at beginning and end.

We therefore propose that the transient response of a linear, time invariant filter can be viewed in an alternate manner. The filter responds immediately to the frequency components within its purview. However, referring to our graphical example, these components are initially smeared, due to the fact that the input presented to the filter is not pure 5-Hz energy, but is zero for part of the duration. The filter picks up whatever 4-Hz energy is present (which actually comes from spreading of the 5-Hz input signal). This level varies as the filter slides along, becoming zero when the entire filter becomes filled with 5-Hz energy at the end of the transient region. This may explain the seeming presence within the transient response of a 4-Hz waveform (as can be seen by counting peaks) although an LTI filter can never alter the frequency content of the input.

The above discussion relates to understanding modulated signals, as well. When we began this chapter, our intuition was that a signal such as an AM sine-modulated sine could never be considered steady state at any point in time, since due to its continually changing amplitude, the product of the signal with the left hand side of the convolution kernel would never balance

with the product of the right hand side. The same would hold true for FM signals, as well, because a changing frequency will also produce an asymmetric product. It followed that for signals which are not in steady state, the Peak-Locus concept might not hold, as one had no way to predict what the effect of transient signals would be on the Peak-Locus. However, the results of this chapter showed that known, modulated signals, gave accurate, predictable results and were resolved exactly into their constituent components as predicted from theory. Furthermore, the same results were found at multiple points within the same signal. This led to rethinking the concept of steady state by analyzing in the frequency domain, rather than in the time domain alone, and it seemed intuitive that the sum of steady-state components must be considered steady state, as well, even though the resultant waveform may be asymmetric. Within the current viewpoint, modulated signals can be considered steady state, as they are composed of the sum of constant microcomponents (carrier and sidebands), hence the Peak-Locus framework is valid for analysis.

The underlying philosophy behind this argument is based on reexamining the limits in the integrals at the beginning of this section. While we formally write that we integrate from minus infinity to plus infinity, in practice we usually eliminate the zero portions of the signal, as they make no contribution to the integral. But, we can do even better by considering that, for example, modulation on a signal which occurred before we were born is unlikely to affect our percept of that signal. We actually need consider only those portions of past history which fit inside the purview of the filter. Those can be incorporated into the output signal through the overlap-add framework common in digital speech processing (Quatieri, 2002).

Based on this way of thinking, we can extend the reasoning to explain the effect of the transient response on a Peak-Locus plot. In the initial region before the filter has moved a filter-length across the signal, there will be smearing of the input components. This will lead to a continuous type spectrum. However, since, as we have shown in Chapter 5 that any time there is a component that lies in between the filter CF's of two adjacent filters there must be a curvature or discontinuity in the Peak-Locus lines, and because every continuous spectrum must have components that lie in between any two fixed points, the effect of transients must be to cause curvature of the lines. Similarly, in the case of modulation types in which there are likely to be very close microcomponents, there must be curvature in between each such microcomponent,

as well. This may possibly explain the curvature of the Peak-Locus lines seen in the unprocessed speech plot of Figure 64. The curvature of the lower harmonics is especially visible and prominent. We note however, that at the outset of our investigation we expected a single point of discontinuity between each interfering harmonic, and rigidly straight lines everywhere else. We viewed the gradual curvature in the Peak-Locus lines of Figure 64 near harmonic frequencies as a major disappointment, and we attributed it to poor filter design, and to an overly long transient response. Our view of transients was that they were inaccuracies due to unpredictable behavior of filter outputs before steady state is reached, and that nothing more could be done, other than trying to find an acceptable filter with shorter time response. We assumed that modulation of harmonics meant steady state could never be reached, and hence the Peak-Locus method might never be useful in practice with time-varying signals. After the tests in this chapter, we realized that the viewpoint of a single frequency component per harmonic was incomplete, and that one must account for each individual sideband. The disadvantage of this finding is that in practice, it is much more difficult to work with gradually curved Peak-Locus lines, as opposed to sharp single point discontinuities, as we have discussed previously. However, it turned out to be very fortunate that the Simultaneous-Equations algorithm seems to identify these very closely spaced individual microcomponents with high accuracy and reliability.

We may now be in a position to offer another explanation for the discrepancy in our earlier FM tests in 5.5.5 with our results in Section 6.4.2. In the first set of tests, we treated FM as being composed of a single time-varying component. The elimination of sidebands from analysis obscures the true trajectory of the FM signal, and gives it only a piecewise approximation in between peak pairs. To correctly account for the continuous nature of the modulation, it is necessary to analyze and incorporate the sidebands, as well. This may explain some of the inaccuracy in the estimated parameters of even the constant component in the earlier tests.

In closing, we note that it makes intuitive sense to tie the two concepts of transient response and modulation in the same framework for another reason, as well, since turn-on transients are nothing more than a special case of amplitude modulation, namely, a step-function envelope. In addition, the stop consonants abruptly interrupt the course of speech numerous times within an utterance. Hence one should not expect that a separate theory would be necessary or ideal. The

difference might be essentially a matter of degree. In cases of modulation, the sidebands may be far enough apart to resolve separately; while in the case of a sharp onset, the sidebands may have a more continuous or tightly-clustered distribution, and cannot readily be resolved into discrete components. In either case, curvature of Peak-Locus lines would be observed.

We further note that the issue of the appropriate limits on the integrals in the beginning of this section still needs further thought, as we are troubled by the uneven heights of the peaks within the transient regions of Figure 95. The appearance suggests some type of modulation, which in turn implies that there must be some spreading of spectral energy, and hence the filter does not appear to be functioning as a pure 4-Hz filter during the transient region. Although we managed to account for most of the features in the plot with this model, this last aspect requires further thought.

6.9 Summary

We have looked at AM and FM modulation, noise and speech signals. We have found that our method of analysis appears to process raw band data so that the underlying sources that gave rise to this data are revealed. We view this as being extremely useful in the sense of dimensionality reduction. While, in general, unprocessed band data from multiple channels appears chaotic and unstructured even in simple cases, we have shown that it can be neatly consolidated.

We have seen that we can resolve sidebands of both AM and FM signals, and that these give useful information on the trajectory of these signals. We discussed the dichotomy within the concept of instantaneous frequency, and tried to elucidate the shortcomings inherent in describing the behavior of modulated signals and mixtures of sinusoids on the basis of a single value of instantaneous frequency. We suggest that using one number to describe a mixture of sinusoids of differing frequency is like adding apples and oranges. We have shown that other authors have found problems, as well. We believe that this shortcoming has impeded the ability to analyze closely-spaced signals, and is why so many authors make the assumption that their algorithms are only reliable when the minimum signal separation is about 25 Hz. We have shown accurate separation at frequency differences of less than 5 Hz. However, further tests on time-varying mixtures are warranted before drawing conclusions.

We conducted tests in additive noise and found that the algorithm consolidates noise into scattered components of random frequency and amplitude. When these are low, the original signal values can be extracted. As noise increases, it causes amplitude errors in the estimated parameters. As it increases still further, it causes frequency errors, as well, and in addition, causes the original signal components to become more and more unrecognizable from the noiselike components.

Speech tests seemed to resolve individual sidebands of each harmonic. However, in tests on a speech mixture, the results were not entirely consistent with the results on separate speakers. Further work on filter design, and search for optimal threshold parameters should be conducted.

We concluded the chapter with a discussion of the effect of filter transient response on the performance of our algorithms, and concluded that while it causes curving of Peak-Locus lines from their ideal linearity, the behavior can possibly be understood via analysis in the frequency domain, and the viewing of transients as a spectral spreading of that part of the input signal which is located within the purview of the filter.

We further concluded that the Simultaneous-Equation approach which is initially based on the Peak-Locus concept, has the advantage that the iterative process seem to overcome the curving of the lines, at least in cases where it is caused by a set of discrete sidebands, and is able to identify components which appear hidden in a simple visual analysis.

Chapter 7

Analytical Approaches

7.1 Introduction

While we have examined signals of various types using the computational approaches described in the last two chapters, there is much to gain by pursuing an exact, closed-form solution to the central problem of parameter estimation in mixtures of sinusoids by means of local maxima. The motivations are the following:

1. The computational and memory requirements of repeated iterations of highly sampled signals are overwhelming. The reason for the high sampling rates is that precise identification of coordinates of local maxima is absolutely required. If a local maximum happens to be located in between two sampling points, its exact coordinates cannot be known. For example, if one is using a conventional sampling rate of 10 KHz, and is working with a mixture of signals containing components near 2.5 KHz, one has only 4 samples per cycle. There is no guarantee that a local maxima will fall on or near a sampling point. The use of interpolation eases the burden somewhat, but brings along with it possible inaccuracies, as mentioned previously.

2. The time required to compute results for even a single pair of time points is very long, making the use on a full-length recording impractical with the current computing platform.

3. We wish to avoid numerical errors inherent in the repeated iterations.

4. While we used intuitive arguments to motivate much of the work, there was no rigorous development of the mathematics. This left unanswered questions regarding the existence and uniqueness of solutions.

We therefore seek analytical approaches to gain additional insight and for the possibility that they may yield solutions without need for iterations. We restrict ourselves to the simplest case, 2 sinusoids in 2 bands. We note that while it is still necessary to use very high sampling rates, this does not provide a high computational burden, as we must only generate one copy of each of the channel outputs. After we extract the coordinates of the local maxima, we never again need to return to the original signals or to generate new signals; all subsequent calculations are made with the 8 extracted values (2 time coordinates and 2 amplitude coordinates in 2 bands each).

We outline the approach as follows. We first show that if one knows the frequencies and the phases of the two sinusoids, then computing the amplitudes from the local maxima of the two mixtures is trivial, and just a solution of an over-determined linear system. We then show that given the frequencies, a unique solution for the phases exists. Finally, we show that the frequencies must obey certain constraints, and additionally that they are bounded within a region. The constraints are that a certain 4×4 determinant in functions of frequency alone must vanish, and that the product of two 3×3 such determinants must equal the product of two other 3×3 such determinants. The separate bounding condition will be shown graphically.

Our approach is to assume that the sinusoids are constant, although as we have seen in Chapter 6, time-varying signals can often be expressed as the sum of a small number of constant signals.

7.2 Derivation

We begin by writing one equation for each member of the pair of local maxima closest to the chosen time point of analysis in each of the two channels.

$$
\begin{aligned}
x_1(t_{11}) &= H_1(\omega_1)a_1\sin(\omega_1 t_{11}+\phi_1)+H_1(\omega_2)a_2\sin(\omega_2 t_{11}+\phi_2)\\
x_1(t_{12}) &= H_1(\omega_1)a_1\sin(\omega_1 t_{12}+\phi_1)+H_1(\omega_2)a_2\sin(\omega_2 t_{12}+\phi_2)\\
x_2(t_{21}) &= H_2(\omega_1)a_1\sin(\omega_1 t_{21}+\phi_1)+H_2(\omega_2)a_2\sin(\omega_2 t_{21}+\phi_2)\\
x_2(t_{22}) &= H_2(\omega_1)a_1\sin(\omega_1 t_{22}+\phi_1)+H_2(\omega_2)a_2\sin(\omega_2 t_{22}+\phi_2)
\end{aligned}
\tag{7.1}
$$

where

$x_1(t_{11})$ represents the output of the first filter at the time of the first peak in that channel.

$x_1(t_{12})$ represents the output of the first filter at the time of the second peak in that channel.

$x_2(t_{21})$ represents the output of the second filter at the time of the first peak in that channel.

$x_2(t_{22})$ represents the output of the second filter at the time of the second peak in that channel.

$H_1(\omega)$ and $H_2(\omega)$ represents the frequency responses of the first and second filters, respectively, to an applied frequency ω.

a_1 and a_2 are the amplitudes of the first and second sinusoids, respectively.

ω_1 and ω_2 are the radian frequencies of the first and second sinusoids, respectively.

ϕ_1 and ϕ_2 are the phases of the first and second sinusoids, respectively.

We emphasize that t_{11}, t_{12}, t_{21}, and t_{22} are measured and known parameters, corresponding to the times of the local maxima, as are $x_1(t_{11})$, $x_1(t_{12})$, $x_2(t_{21})$, and $x_2(t_{22})$ which are the heights of the local maxima.

As noted earlier, if we are given the frequencies ω_1 and ω_2, and phases ϕ_1 and ϕ_2, then it is trivial to solve the set for the amplitudes a_1 and a_2, since everything else depends on the known frequencies and phases. We assume that the forms of $H_1(\omega)$ and $H_2(\omega)$ are known *a priori*, as they represent the characteristics of the filters we are using.

In addition, from elementary calculus, we know that the derivatives at the times of the local maxima must vanish. Thus we have 4 derivative equations:

$$
\begin{aligned}
\frac{dx_1(t_{11})}{dt} &= H_1(\omega_1)a_1\omega_1\cos(\omega_1 t_{11}+\phi_1)+H_1(\omega_2)a_2\omega_2\cos(\omega_2 t_{11}+\phi_2)=0\\
\frac{dx_1(t_{12})}{dt} &= H_1(\omega_1)a_1\omega_1\cos(\omega_1 t_{12}+\phi_1)+H_1(\omega_2)a_2\omega_2\cos(\omega_2 t_{12}+\phi_2)=0\\
\frac{dx_2(t_{21})}{dt} &= H_2(\omega_1)a_1\omega_1\cos(\omega_1 t_{21}+\phi_1)+H_2(\omega_2)a_2\omega_2\cos(\omega_2 t_{21}+\phi_2)=0\\
\frac{dx_2(t_{22})}{dt} &= H_2(\omega_1)a_1\omega_1\cos(\omega_1 t_{22}+\phi_1)+H_2(\omega_2)a_2\omega_2\cos(\omega_2 t_{22}+\phi_2)=0
\end{aligned}
\tag{7.2}
$$

where we use the shorthand $\frac{dx_1(t_{11})}{dt}$ to mean $\left.\frac{dx_1}{dt}\right|_{t=t_{11}}$

Rearranging, we get the following:

$$\begin{aligned} H_1(\omega_1)a_1\omega_1\cos(\omega_1 t_{11}+\phi_1) &= -H_1(\omega_2)a_2\omega_2\cos(\omega_2 t_{11}+\phi_2) \\ H_1(\omega_1)a_1\omega_1\cos(\omega_1 t_{12}+\phi_1) &= -H_1(\omega_2)a_2\omega_2\cos(\omega_2 t_{12}+\phi_2) \\ H_2(\omega_1)a_1\omega_1\cos(\omega_1 t_{21}+\phi_1) &= -H_2(\omega_2)a_2\omega_2\cos(\omega_2 t_{21}+\phi_2) \\ H_2(\omega_1)a_1\omega_1\cos(\omega_1 t_{22}+\phi_1) &= -H_2(\omega_2)a_2\omega_2\cos(\omega_2 t_{22}+\phi_2) \end{aligned} \tag{7.3}$$

Dividing the first equation of set (7.3) by the second equation of set (7.3) yields

$$\frac{\cos(\omega_1 t_{11}+\phi_1)}{\cos(\omega_1 t_{12}+\phi_1)} = \frac{\cos(\omega_2 t_{11}+\phi_2)}{\cos(\omega_2 t_{12}+\phi_2)} \tag{7.4}$$

Dividing the third of (7.3) by the fourth of (7.3) yields

$$\frac{\cos(\omega_1 t_{21}+\phi_1)}{\cos(\omega_1 t_{22}+\phi_1)} = \frac{\cos(\omega_2 t_{21}+\phi_2)}{\cos(\omega_2 t_{22}+\phi_2)} \tag{7.5}$$

Note that these last two (7.4) and (7.5) are independent of amplitudes and filter characteristics.

We can derive two more relations by dividing the first of set (7.3) by the third of set (7.3) and the second of set (7.3) by the fourth of set (7.3).

$$\frac{H_1(\omega_1)\cos(\omega_1 t_{11}+\phi_1)}{H_2(\omega_1)\cos(\omega_1 t_{21}+\phi_1)} = \frac{H_1(\omega_2)\cos(\omega_2 t_{11}+\phi_2)}{H_2(\omega_2)\cos(\omega_2 t_{21}+\phi_2)} \tag{7.6}$$

$$\frac{H_1(\omega_1)\cos(\omega_1 t_{12}+\phi_1)}{H_2(\omega_1)\cos(\omega_1 t_{22}+\phi_1)} = \frac{H_1(\omega_2)\cos(\omega_2 t_{12}+\phi_2)}{H_2(\omega_2)\cos(\omega_2 t_{22}+\phi_2)} \tag{7.7}$$

Equations (7.6) and (7.7) are independent of amplitude, but not of filter characteristics.

We now apply trigonometric addition formulas to each of equations (7.4)-(7.7). This step, although greatly expanding and complicating the algebra, is necessary to separate the contributions of frequency from phase. The addition formula for cosines is

$$\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b) \tag{7.8}$$

We now have a product of functions of phase and frequency, rather than a sum, but have also picked up a second term of sine functions, as well. We nevertheless proceed.

Equation (7.4) becomes

$$\frac{\cos(\omega_1 t_{11})\cos(\phi_1)-\sin(\omega_1 t_{11})\sin(\phi_1)}{\cos(\omega_1 t_{12})\cos(\phi_1)-\sin(\omega_1 t_{12})\sin(\phi_1)}=\frac{\cos(\omega_2 t_{11})\cos(\phi_2)-\sin(\omega_2 t_{11})\sin(\phi_2)}{\cos(\omega_2 t_{12})\cos(\phi_2)-\sin(\omega_2 t_{12})\sin(\phi_2)} \tag{7.9}$$

Equation (7.5) becomes

$$\frac{\cos(\omega_1 t_{21})\cos(\phi_1)-\sin(\omega_1 t_{21})\sin(\phi_1)}{\cos(\omega_1 t_{22})\cos(\phi_1)-\sin(\omega_1 t_{22})\sin(\phi_1)}=\frac{\cos(\omega_2 t_{21})\cos(\phi_2)-\sin(\omega_2 t_{21})\sin(\phi_2)}{\cos(\omega_2 t_{22})\cos(\phi_2)-\sin(\omega_2 t_{22})\sin(\phi_2)} \tag{7.10}$$

Equation (7.6) becomes

$$\frac{H_1(\omega_1)[\cos(\omega_1 t_{11})\cos(\phi_1)-\sin(\omega_1 t_{11})\sin(\phi_1)]}{H_2(\omega_1)[\cos(\omega_1 t_{21})\cos(\phi_1)-\sin(\omega_1 t_{21})\sin(\phi_1)]}=\frac{H_1(\omega_2)[\cos(\omega_2 t_{11})\cos(\phi_2)-\sin(\omega_2 t_{11})\sin(\phi_2)]}{H_2(\omega_2)[\cos(\omega_2 t_{21})\cos(\phi_2)-\sin(\omega_2 t_{21})\sin(\phi_2)]} \tag{7.11}$$

Equation (7.7) becomes

$$\frac{H_1(\omega_1)[\cos(\omega_1 t_{12})\cos(\phi_1)-\sin(\omega_1 t_{12})\sin(\phi_1)]}{H_2(\omega_1)[\cos(\omega_1 t_{22})\cos(\phi_1)-\sin(\omega_1 t_{22})\sin(\phi_1)]}=\frac{H_1(\omega_2)[\cos(\omega_2 t_{12})\cos(\phi_2)-\sin(\omega_2 t_{12})\sin(\phi_2)]}{H_2(\omega_2)[\cos(\omega_2 t_{22})\cos(\phi_2)-\sin(\omega_2 t_{22})\sin(\phi_2)]} \tag{7.12}$$

We now cross-multiply the fractions in equations (7.9)-(7.12). The algebra is extremely tedious, but simplicity will eventually emerge.

The top of the left side of (7.9) multiplied by the bottom of the right side of (7.9) is

$$\begin{aligned}&\cos(\omega_1 t_{11})\cos(\phi_1)\cos(\omega_2 t_{12})\cos(\phi_2)\\&-\sin(\omega_1 t_{11})\sin(\phi_1)\cos(\omega_2 t_{12})\cos(\phi_2)\\&-\cos(\omega_1 t_{11})\cos(\phi_1)\sin(\omega_2 t_{12})\sin(\phi_2)\\&+\sin(\omega_1 t_{11})\sin(\phi_1)\sin(\omega_2 t_{12})\sin(\phi_2)\end{aligned} \tag{7.13}$$

The bottom of the left side of (7.9) multiplied by the top of the right side of (7.9) is

$$\begin{aligned}&\cos(\omega_1 t_{12})\cos(\phi_1)\cos(\omega_2 t_{11})\cos(\phi_2)\\&-\sin(\omega_1 t_{12})\sin(\phi_1)\cos(\omega_2 t_{11})\cos(\phi_2)\\&-\cos(\omega_1 t_{12})\cos(\phi_1)\sin(\omega_2 t_{11})\sin(\phi_2)\\&+\sin(\omega_1 t_{12})\sin(\phi_1)\sin(\omega_2 t_{11})\sin(\phi_2)\end{aligned} \tag{7.14}$$

The top of the left side of (7.10) multiplied by the bottom of the right side of (7.10) is

$$
\begin{aligned}
&\cos(\omega_1 t_{21})\cos(\phi_1)\cos(\omega_2 t_{22})\cos(\phi_2)\\
&-\sin(\omega_1 t_{21})\sin(\phi_1)\cos(\omega_2 t_{22})\cos(\phi_2)\\
&-\cos(\omega_1 t_{21})\cos(\phi_1)\sin(\omega_2 t_{22})\sin(\phi_2)\\
&+\sin(\omega_1 t_{21})\sin(\phi_1)\sin(\omega_2 t_{22})\sin(\phi_2)
\end{aligned}
\tag{7.15}
$$

The bottom of the left side of (7.10) multiplied by the top of the right side of (7.10) is

$$
\begin{aligned}
&\cos(\omega_1 t_{22})\cos(\phi_1)\cos(\omega_2 t_{21})\cos(\phi_2)\\
&-\sin(\omega_1 t_{22})\sin(\phi_1)\cos(\omega_2 t_{21})\cos(\phi_2)\\
&-\cos(\omega_1 t_{22})\cos(\phi_1)\sin(\omega_2 t_{21})\sin(\phi_2)\\
&+\sin(\omega_1 t_{22})\sin(\phi_1)\sin(\omega_2 t_{21})\sin(\phi_2)
\end{aligned}
\tag{7.16}
$$

The top of the left side of (7.11) multiplied by the bottom of the right side of (7.11) is

$$
\begin{aligned}
&H_1(\omega_1)H_2(\omega_2)\cos(\omega_1 t_{11})\cos(\phi_1)\cos(\omega_2 t_{21})\cos(\phi_2)\\
&-H_1(\omega_1)H_2(\omega_2)\sin(\omega_1 t_{11})\sin(\phi_1)\cos(\omega_2 t_{21})\cos(\phi_2)\\
&-H_1(\omega_1)H_2(\omega_2)\cos(\omega_1 t_{11})\cos(\phi_1)\sin(\omega_2 t_{21})\sin(\phi_2)\\
&+H_1(\omega_1)H_2(\omega_2)\sin(\omega_1 t_{11})\sin(\phi_1)\sin(\omega_2 t_{21})\sin(\phi_2)
\end{aligned}
\tag{7.17}
$$

The bottom of the left side of (7.11) multiplied by the top of the right side of (7.11) is

$$
\begin{aligned}
&H_2(\omega_1)H_1(\omega_2)\cos(\omega_1 t_{21})\cos(\phi_1)\cos(\omega_2 t_{11})\cos(\phi_2)\\
&-H_2(\omega_1)H_1(\omega_2)\sin(\omega_1 t_{21})\sin(\phi_1)\cos(\omega_2 t_{11})\cos(\phi_2)\\
&-H_2(\omega_1)H_1(\omega_2)\cos(\omega_1 t_{21})\cos(\phi_1)\sin(\omega_2 t_{11})\sin(\phi_2)\\
&+H_2(\omega_1)H_1(\omega_2)\sin(\omega_1 t_{21})\sin(\phi_1)\sin(\omega_2 t_{11})\sin(\phi_2)
\end{aligned}
\tag{7.18}
$$

The top of the left side of (7.12) multiplied by the bottom of the right side of (7.12) is

$$
\begin{aligned}
&H_1(\omega_1)H_2(\omega_2)\cos(\omega_1 t_{12})\cos(\phi_1)\cos(\omega_2 t_{22})\cos(\phi_2)\\
&-H_1(\omega_1)H_2(\omega_2)\sin(\omega_1 t_{12})\sin(\phi_1)\cos(\omega_2 t_{22})\cos(\phi_2)\\
&-H_1(\omega_1)H_2(\omega_2)\cos(\omega_1 t_{12})\cos(\phi_1)\sin(\omega_2 t_{22})\sin(\phi_2)\\
&+H_1(\omega_1)H_2(\omega_2)\sin(\omega_1 t_{12})\sin(\phi_1)\sin(\omega_2 t_{22})\sin(\phi_2)
\end{aligned}
\tag{7.19}
$$

The bottom of the left side of (7.12) multiplied by the top of the right side of (7.12) is

$$
\begin{aligned}
&H_2(\omega_1)H_1(\omega_2)\cos(\omega_1 t_{22})\cos(\phi_1)\cos(\omega_2 t_{12})\cos(\phi_2)\\
&-H_2(\omega_1)H_1(\omega_2)\sin(\omega_1 t_{21})\sin(\phi_1)\cos(\omega_2 t_{12})\cos(\phi_2)\\
&-H_2(\omega_1)H_1(\omega_2)\cos(\omega_1 t_{21})\cos(\phi_1)\sin(\omega_2 t_{12})\sin(\phi_2)\\
&+H_2(\omega_1)H_1(\omega_2)\sin(\omega_1 t_{21})\sin(\phi_1)\sin(\omega_2 t_{12})\sin(\phi_2)
\end{aligned}
\tag{7.20}
$$

We now note that it is possible to consolidate expressions (7.13)-(7.20) using matrices. This will be of great assistance in our future analysis. Let us examine the structure of these expressions.

Each pair, (7.13) & (7.14), (7.15) & (7.16), (7.17) & (7.18), and (7.19) & (7.20) was obtained by cross-multiplying the two equal fractions in each of equations (7.9)-(7.12), and hence are simply each the right hand side and left hand sides of an equation.

We have

$$(7.13) = (7.14)$$

$$(7.15) = (7.16)$$

$$(7.17) = (7.18)$$

$$(7.19) = (7.20)$$

Furthermore, each of these 8 expressions (7.13) through (7.20) can be written as the product of a row vector consisting of four individual frequency-dependent elements multiplied by the following common column vector **v** consisting of four phase-dependent elements.

$$\mathbf{v} \equiv \begin{bmatrix} \cos(\phi_1)\cos(\phi_2) \\ -\sin(\phi_1)\cos(\phi_2) \\ -\cos(\phi_1)\sin(\phi_2) \\ \sin(\phi_1)\sin(\phi_2) \end{bmatrix} \tag{7.21}$$

Expression (7.13) can be written as the product of the following two vectors.

$$\left[\cos(\omega_1 t_{11})\cos(\omega_2 t_{12}) \quad \sin(\omega_1 t_{11})\cos(\omega_2 t_{12}) \quad \cos(\omega_1 t_{11})\sin(\omega_2 t_{12}) \quad \sin(\omega_1 t_{11})\sin(\omega_2 t_{12})\right][\mathbf{v}] \tag{7.22}$$

Expression (7.14) can be written as the product of the following two vectors.

$$\left[\cos(\omega_1 t_{12})\cos(\omega_2 t_{11}) \quad \sin(\omega_1 t_{12})\cos(\omega_2 t_{11}) \quad \cos(\omega_1 t_{12})\sin(\omega_2 t_{11}) \quad \sin(\omega_1 t_{12})\sin(\omega_2 t_{11})\right][\mathbf{v}] \tag{7.23}$$

Expression (7.15) can be written as the product of the following two vectors.

$$\left[\cos(\omega_1 t_{21})\cos(\omega_2 t_{22}) \quad \sin(\omega_1 t_{21})\cos(\omega_2 t_{22}) \quad \cos(\omega_1 t_{21})\sin(\omega_2 t_{22}) \quad \sin(\omega_1 t_{21})\sin(\omega_2 t_{22})\right][\mathbf{v}] \tag{7.24}$$

Expression (7.16) can be written as the product of the following two vectors.

$$\left[\cos(\omega_1 t_{22})\cos(\omega_2 t_{21}) \quad \sin(\omega_1 t_{22})\cos(\omega_2 t_{21}) \quad \cos(\omega_1 t_{22})\sin(\omega_2 t_{21}) \quad \sin(\omega_1 t_{22})\sin(\omega_2 t_{21})\right][\mathbf{v}] \tag{7.25}$$

Expression (7.17) can be written as the product of the following scalars and two vectors.

$$(7.26)\quad \begin{matrix} H_1(\omega_1)H_2(\omega_2) \\ \times\left[\cos(\omega_1 t_{11})\cos(\omega_2 t_{21}) \quad \sin(\omega_1 t_{11})\cos(\omega_2 t_{21}) \quad \cos(\omega_1 t_{11})\sin(\omega_2 t_{21}) \quad \sin(\omega_1 t_{11})\sin(\omega_2 t_{21})\right][\mathbf{v}] \end{matrix}$$

Expression (7.18) can be written as the product of the following scalars and two vectors.

$$(7.27)\quad \begin{matrix} H_2(\omega_1)H_1(\omega_2) \\ \times\left[\cos(\omega_1 t_{21})\cos(\omega_2 t_{11}) \quad \sin(\omega_1 t_{21})\cos(\omega_2 t_{11}) \quad \cos(\omega_1 t_{21})\sin(\omega_2 t_{11}) \quad \sin(\omega_1 t_{21})\sin(\omega_2 t_{11})\right][\mathbf{v}] \end{matrix}$$

Expression (7.19) can be written as the product of the following scalars and two vectors.

$$(7.28)\quad \begin{matrix} H_1(\omega_1)H_2(\omega_2) \\ \times\left[\cos(\omega_1 t_{12})\cos(\omega_2 t_{22}) \quad \sin(\omega_1 t_{12})\cos(\omega_2 t_{22}) \quad \cos(\omega_1 t_{12})\sin(\omega_2 t_{22}) \quad \sin(\omega_1 t_{12})\sin(\omega_2 t_{22})\right][\mathbf{v}] \end{matrix}$$

Expression (7.20) can be written as the product of the following scalars and two vectors.

$$(7.29)\quad \begin{matrix} H_2(\omega_1)H_1(\omega_2) \\ \times\left[\cos(\omega_1 t_{22})\cos(\omega_2 t_{12}) \quad \sin(\omega_1 t_{22})\cos(\omega_2 t_{12}) \quad \cos(\omega_1 t_{22})\sin(\omega_2 t_{12}) \quad \sin(\omega_1 t_{22})\sin(\omega_2 t_{12})\right][\mathbf{v}] \end{matrix}$$

To make things less cumbersome, let us assign names to each of the row vectors in (7.22)-(7.29).

$$(7.30)\ \mathbf{qa} = \left[\cos(\omega_1 t_{11})\cos(\omega_2 t_{12}) \quad \sin(\omega_1 t_{11})\cos(\omega_2 t_{12}) \quad \cos(\omega_1 t_{11})\sin(\omega_2 t_{12}) \quad \sin(\omega_1 t_{11})\sin(\omega_2 t_{12})\right]$$

$$(7.31)\ \mathbf{za} = \left[\cos(\omega_1 t_{12})\cos(\omega_2 t_{11}) \quad \sin(\omega_1 t_{12})\cos(\omega_2 t_{11}) \quad \cos(\omega_1 t_{12})\sin(\omega_2 t_{11}) \quad \sin(\omega_1 t_{12})\sin(\omega_2 t_{11})\right]$$

$$(7.32)\ \mathbf{qb} = \left[\cos(\omega_1 t_{21})\cos(\omega_2 t_{22}) \quad \sin(\omega_1 t_{21})\cos(\omega_2 t_{22}) \quad \cos(\omega_1 t_{21})\sin(\omega_2 t_{22}) \quad \sin(\omega_1 t_{21})\sin(\omega_2 t_{22})\right]$$

$$(7.33)\ \mathbf{zb} = \left[\cos(\omega_1 t_{22})\cos(\omega_2 t_{21}) \quad \sin(\omega_1 t_{22})\cos(\omega_2 t_{21}) \quad \cos(\omega_1 t_{22})\sin(\omega_2 t_{21}) \quad \sin(\omega_1 t_{22})\sin(\omega_2 t_{21})\right]$$

$$(7.34)\ \mathbf{qc} = \left[\cos(\omega_1 t_{11})\cos(\omega_2 t_{21}) \quad \sin(\omega_1 t_{11})\cos(\omega_2 t_{21}) \quad \cos(\omega_1 t_{11})\sin(\omega_2 t_{21}) \quad \sin(\omega_1 t_{11})\sin(\omega_2 t_{21})\right]$$

$$(7.35)\ \mathbf{zc} = \left[\cos(\omega_1 t_{21})\cos(\omega_2 t_{11}) \quad \sin(\omega_1 t_{21})\cos(\omega_2 t_{11}) \quad \cos(\omega_1 t_{21})\sin(\omega_2 t_{11}) \quad \sin(\omega_1 t_{21})\sin(\omega_2 t_{11})\right]$$

$$(7.36)\ \mathbf{qd} = \left[\cos(\omega_1 t_{12})\cos(\omega_2 t_{22}) \quad \sin(\omega_1 t_{12})\cos(\omega_2 t_{22}) \quad \cos(\omega_1 t_{12})\sin(\omega_2 t_{22}) \quad \sin(\omega_1 t_{12})\sin(\omega_2 t_{22})\right]$$

$$(7.37)\ \mathbf{zd} = \left[\cos(\omega_1 t_{22})\cos(\omega_2 t_{12}) \quad \sin(\omega_1 t_{22})\cos(\omega_2 t_{12}) \quad \cos(\omega_1 t_{22})\sin(\omega_2 t_{12}) \quad \sin(\omega_1 t_{22})\sin(\omega_2 t_{12})\right]$$

Because, as before, each of (7.13) & (7.14), (7.15) & (7.16), (7.17) & (7.18) and (7.19) & (7.20) is an equal pair, we can write

(7.38) $$[\mathbf{qa}][\mathbf{v}] = [\mathbf{za}][\mathbf{v}]$$

(7.39) $$[\mathbf{qb}][\mathbf{v}] = [\mathbf{zb}][\mathbf{v}]$$

(7.40) $$H_1(\omega_1)H_2(\omega_2)[\mathbf{qc}][\mathbf{v}] = H_2(\omega_1)H_1(\omega_2)[\mathbf{zc}][\mathbf{v}]$$

(7.41) $$H_1(\omega_1)H_2(\omega_2)[\mathbf{qd}][\mathbf{v}] = H_2(\omega_1)H_1(\omega_2)[\mathbf{zd}][\mathbf{v}]$$

Let us further consolidate all the **q** row vectors into a matrix **Q** and all the **z** row vectors into a matrix **Z** as follows:

(7.42) $$\mathbf{Q} \equiv \begin{bmatrix} \mathbf{qa} \\ \mathbf{qb} \\ H_1(\omega_1)H_2(\omega_2)\mathbf{qc} \\ H_1(\omega_1)H_2(\omega_2)\mathbf{qd} \end{bmatrix}$$

(7.43) $$\mathbf{Z} \equiv \begin{bmatrix} \mathbf{za} \\ \mathbf{zb} \\ H_2(\omega_1)H_1(\omega_2)\mathbf{zc} \\ H_2(\omega_1)H_1(\omega_2)\mathbf{zd} \end{bmatrix}$$

We can now combine (7.38) through (7.41) as follows:

(7.44) $$\mathbf{Qv = Zv}$$

Applying the distributive law for matrix multiplication we have

(7.45) $$\mathbf{(Q - Z)v = 0}$$

Let us now name this difference matrix as **D**.

(7.46) $$\mathbf{D \equiv Q - Z}$$

We then have

(7.47) $$\mathbf{Dv = 0}$$

Now that we have the preliminary algebra out of the way, let us examine the implications of this important result.

First, we note that all the elements of **D** depend only on the frequencies ω_1 and ω_2, and do not have any dependency on phase. Conversely, all the elements of **v** depend only on the phases ϕ_1

and ϕ_2, and do not have any dependency on frequency. Thus, separation of frequency and phase has been achieved which will be very useful in the following discussion.

7.3 Solving for the Nullspace

7.3.1 Relation between Frequencies and Phases

The solution to a set of linear equations $\mathbf{Ax = 0}$ has a number of names. It is known variously as the homogeneous solution, the kernel of the matrix **A**, or the nullspace of the matrix **A**. In this work, we will use the term nullspace. Recall that if the matrix **A** is invertible, then we can multiply both sides by $\mathbf{A^{-1}}$ to get $\mathbf{A^{-1}Ax = A^{-1}0 = 0}$, meaning that the trivial solution $\mathbf{x = 0}$ is the only solution, and the nullspace consists of only that single point.

If, however, the matrix **A** is not invertible, then the nullspace consists of nontrivial solutions, as well. The key result we obtained in Section 7.2 is that **the phase vector v is the nullspace of the frequency matrix D.** The importance of this is that if we have knowledge of the frequencies alone, we can compute the unknown phases by a choice of linear algebra techniques to be discussed shortly.

7.3.2 Background

Before we discuss methods for computing the nullspace, we review a number of theorems of linear algebra to provide some background.

If a square matrix is invertible, then its determinant is nonzero and it has full rank, meaning that the rank is equal to the number of rows or columns of the matrix. Rank is the number of linearly independent rows and columns within a matrix. If a matrix has full rank, then no row (column) can be expressed as a linear combination of the remaining rows (columns).

If a matrix is noninvertible, then its determinant is zero and it has less than full rank, meaning that one or more of its rows (columns) can be expressed as a linear combination of the remaining rows (columns).

There is an important theorem in linear algebra which is termed by (Strang, 1993) as The Fundamental Theorem of Linear Algebra, and is also known as the Rank-Nullity Theorem. If a matrix has *n* rows and *n* columns, and its rank is *r*, then the dimension of the nullspace is $n - r$.

7.3.3 Solution for Frequency

In our situation, we know at the outset that there are nontrivial solutions, since the phases may be any arbitrary values we choose to use for generation of the signal. **The determinant of the frequency matrix D must, therefore, be 0**, therefore, the rank of the matrix must be 3 or less. It turns out, as we will see, that the rank in fact is exactly 3. From this, another key point emerges, since the dimension of the nullspace of the frequency matrix is $n-r$, or 1, there can be no other solution. All solutions must lie in a straight line, and are proportional to each other. Therefore, the **solution uniquely determines the phase vector** to within a proportionality factor. Normalization then limits the solution set to a single (and correct) value.

We note that while there is a formula for solving equations of the form $\mathbf{Ax = b}$ (Cramer's Rule), if one tries to apply it to equations of the form $\mathbf{Ax = 0}$, it will yield only the trivial solution $\mathbf{x = 0}$. Therefore, the standard technique for obtaining the nullspace is usually through row reduction.

We cite from (Akritas, Malaschonok and Vigklas, 2006) the following algorithm for computing the nullspace of a matrix.

• Algorithm to determine a basis for $N(\mathbf{A})$, the right nullspace of matrix **A**, i.e., the space spanned by solutions to $\mathbf{Ax = 0}$:

1) Find the reduced echelon matrix, RE, for the input matrix **A**.

2) Identify free variables and pivot variables of RE.

3) Set one free variable to 0, other free variables to 1, and solve for pivot variables.

4) Repeat step 3 for each free variable.

5) A basis for $N(\mathbf{A})$ is the set of special solution vectors from each step 3.

The following example appears in Wikipedia: Null Space:

Consider the matrix

$$\mathbf{A} = \begin{bmatrix} -2 & -4 & 4 \\ 2 & -8 & 0 \\ 8 & 4 & -12 \end{bmatrix}$$

To find its nullspace, one should find all vectors $\mathbf{v}$ such that $\mathbf{Av} = \mathbf{0}$. One proceeds by transforming $\mathbf{A}$ to reduced row echelon form.

$$\mathbf{E} = \begin{bmatrix} 1 & 0 & -4/3 \\ 0 & 1 & -1/3 \\ 0 & 0 & 0 \end{bmatrix}$$

One has that $\mathbf{Av} = \mathbf{0}$ if and only if $\mathbf{Ev} = \mathbf{0}$. Using the notation $\mathbf{v} = [x, y, z]^{\mathrm{T}}$, the latter equality becomes

$$\begin{bmatrix} 1 & 0 & -4/3 \\ 0 & 1 & -1/3 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}; \quad \begin{bmatrix} x - 4z/3 \\ y - z/3 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}; \quad \begin{bmatrix} x = 4z/3 \\ y = z/3 \\ 0 = 0 \end{bmatrix}; \quad \begin{bmatrix} x = 4s/3 \\ y = s/3 \\ z = s \end{bmatrix}$$

Thus, the null space of A is a one dimensional space,

$$\mathbf{v} = \begin{bmatrix} 4s/3 \\ s/3 \\ s \end{bmatrix}$$

One comment regarding these computations in Matlab® is that for many Matlab® functions, one needs to specify a tolerance in order to obtain correct results, given the presence of numerical roundoff error. We have found using a sampling rate of 1 MHz, and using test signals up to about 25 Hz that the value `0.001` seems to work fine for operations such as `rank` and `rref`.

Before proceeding further, let us clarify the requirement and issue of normalization which we have mentioned a few times. Since the nullspace is actually parameterized in terms of one or more free variables (the exact number depending on the rank of the matrix as $n - r$), one could obtain results which are beyond the range of the real trigonometric functions. The solution is to normalize, as we now demonstrate. While there are many norms defined in the world of mathematics, the conventional ℓ^2 norm is the one we need here.

Recall that (Eq. 7.21) **v** is defined as

$$\mathbf{v} = \begin{bmatrix} \cos(\phi_1)\cos(\phi_2) \\ -\sin(\phi_1)\cos(\phi_2) \\ -\cos(\phi_1)\sin(\phi_2) \\ \sin(\phi_1)\sin(\phi_2) \end{bmatrix}$$

Let us square each element and add.

(7.48) $$\mathbf{v}\bullet\mathbf{v} = \cos^2(\phi_1)\cos^2(\phi_2) + \sin^2(\phi_1)\cos^2(\phi_2) + \cos^2(\phi_1)\sin^2(\phi_2) + \sin^2(\phi_1)\sin^2(\phi_2)$$

But this can be factored as:

(7.49) $$\begin{aligned} \mathbf{v}\bullet\mathbf{v} &= [\cos^2(\phi_1) + \sin^2(\phi_1)][\cos^2(\phi_2) + \sin^2(\phi_2)] \\ &= [1]\bullet[1] \\ &= 1 \end{aligned}$$

by the trigonometric identity $\sin^2(x) + \cos^2(x) = 1$.

Thus, conventional normalization is the correct operation.

7.3.4 Determining Phase

We next discuss how to obtain the individual phases from the phase vector **v**. This is not difficult, and probably the easiest way to obtain ϕ_1 is to simply divide element 2 by element 1

(7.50) $$\frac{v(2)}{v(1)} = \frac{-\sin(\phi_1)\cos(\phi_2)}{\cos(\phi_1)\cos(\phi_2)} = -\tan(\phi_1)$$

(7.51) $$\because \phi_1 = \tan^{-1}\left[\frac{-v(2)}{v(1)}\right]$$

The one caveat is to pay careful attention to the quadrants to make sure one is not calculating the complement or supplement of the true phase angle due to the ambiguities (multiple values) inherent in the inverse trigonometric functions.

Note that one could just as well divide element 4 by element 3 to obtain

(7.52) $$\frac{v(4)}{v(3)} = \frac{\sin(\phi_1)\sin(\phi_2)}{-\cos(\phi_1)\sin(\phi_2)} = -\tan(\phi_1)$$

(7.53) $$\because \phi_1 = \tan^{-1}\left[\frac{-v(4)}{v(3)}\right]$$

Equating (7.51) and (7.53) leads to a constraint on the solution for the phase vector **v.**

$$\frac{v(2)}{v(1)} = \frac{v(4)}{v(3)} \tag{7.54}$$

which implies

$$v(1)v(4) = v(2)v(3) \tag{7.55}$$

Note that this relation can be seen by inspection of the definition of **v**.

This constraint on **v** in turn leads to a second constraint on the frequency matrix **D**. We will shortly make the exact form of this second constraint explicit. (Recall that the first constraint was $\det(\mathbf{D}) = 0$).

As we will emphasize repeatedly, the utility of this analytic approach is not only to calculate phase from frequency, but also to provide consistency checks on the process of frequency estimation, as well.

In a similar manner, we can solve for ϕ_2 by dividing element 4 by element 2 as follows:

$$\frac{v(4)}{v(2)} = \frac{\sin(\phi_1)\sin(\phi_2)}{-\sin(\phi_1)\cos(\phi_2)} = -\tan(\phi_2) \tag{7.56}$$

$$\because \phi_2 = \tan^{-1}\left[\frac{-v(4)}{v(2)}\right] \tag{7.57}$$

As before, we can alternately compute ϕ_2 by dividing element 3 by element 1 to get

$$\frac{v(3)}{v(1)} = \frac{-\cos(\phi_1)\sin(\phi_2)}{\cos(\phi_1)\cos(\phi_2)} = -\tan(\phi_2) \tag{7.58}$$

$$\because \phi_2 = \tan^{-1}\left[\frac{-v(3)}{v(1)}\right] \tag{7.59}$$

As before, we can equate (7.56) and (7.58) to obtain.

$$\frac{v(4)}{v(2)} = \frac{v(3)}{v(1)} \tag{7.60}$$

from which the same constraint we saw in (7.55) again emerges.

(7.61) $$v(1)v(4) = v(2)v(3)$$

We note that if one uses the ratios (7.51), (7.53), (7.57), and (7.59) to compute the phases, then it is strictly unnecessary to normalize **v,** as ratios among the elements will be invariant.

7.3.5 Computing Nullspace from SVD

We note another very useful and quick method for computing **v** from **D**. It turns out that if one computes the Singular Value Decomposition (SVD) of an $n \times n$ matrix **A** as in `[U,E,V]=svd(A)`, then the first *r* columns of the matrix **V** are an orthonormal basis for the row space of **A,** and the remaining *n-r* columns are a basis for the nullspace of **A**. Therefore, for our purposes, it is only necessary to look at the last column vector of **V** and immediately recognize the presence of **v.** (The similarity in the letters is a coincidence). This will only be the case if the frequency matrix has rank 3. If it has rank 4, then the frequencies are incorrect, as the determinant is non-zero. The SVD will still produce 4 column vectors, but the last will not be the nullspace. A way to recognize this incorrect vector is that in many cases, it will not obey the product constraints (7.55) and (7.61). A more accurate way of recognizing an instance in which there is no nullspace is to examine the singular values in matrix **E.** If the matrix is singular (noninvertible), it will have one or more (in our case, one) singular values which are equal to zero. If it is nonsingular (full rank), then all the singular values will be non-zero.

However, we note that perturbing the true frequency by a small amount will only change the SVD by a small amount, as well. One may have singular values which are close to zero, but not exactly. Similarly, roundoff errors may also produce small nonzero values. Therefore, as pointed out in (Nordberg and Farneback, 2001) there is some arbitrariness in setting the cutoff threshold in terms of establishing the rank of a matrix. This is also true in the computation of eigenvalues. One can have small values of which one might be unsure as to whether they should be considered zero or not, for the purpose of establishing the rank of a matrix. Still another similar situation could arise if one looks at the last coefficient of the characteristic polynomial of a matrix to determine rank. Ideally, if zero, one has a singular matrix, and if non-zero, then a nonsingular matrix. In all these tests roundoff error may confuse matters. This is another reason why we have used such high sampling rates, in order to keep any source of errors to an absolute minimum. And this is also a reason why it is useful to have multiple constraints on the frequency estimates. Often, one of them, such as the determinant, might be

somewhat ambiguous, but another one, such as the product constraint might be clearer, and vice versa. The oscillatory nature of the multiple trigonometric functions in the matrix **D** makes for unusual behavior.

7.4 Explicit Formula for Nullspace

Although we have two reliable methods, Row Reduction and SVD, for computing the phase vector **v** from the frequency matrix **D**, there is reason to search further. The reason being that our holy grail is to obtain closed-form solutions to the parameter-estimation problem. The motivation is, first of all, the ease and speed of computation, but much more importantly, the insight it gives into what is really happening. As reliable as the row reduction and SVD methods are, they can both be described as procedures, rather than formulas. They should probably be considered numerical methods, rather than algebraic methods. One can't easily predict, understand or even plot the effect of changing one of the frequency estimates on the resulting phase vector **v**, without actually computing each point. This makes our ultimate goal of accurate frequency estimation harder to attain. One of our primary goals for improving on the iterative algorithms we had been using was to understand the process of frequency estimation, and how the positions of the local maxima determine the frequencies of both components. The previous methods are only useful for brute-force computation of phase, but give no insight as to the nature of the solution.

7.4.1 Nullspace of Q and Z Matrices Separately

An initial attempt was to use the Matlab® Symbolic Toolbox command `null` to compute the nullspace of **D**. This fails and returns an empty result, since in general, the matrix appears to Matlab® as full-rank, and only certain combinations of frequency (which we seek) will reduce the rank. It is interesting that using `null` on **Q** and **Z** alone do return valid solutions. They were found to be

$$null(\mathbf{Q}) = \begin{bmatrix} \tan(\omega_2 t_{22})\tan(\omega_1 t_{11}) \\ -\tan(\omega_2 t_{22}) \\ -\tan(\omega_1 t_{11}) \\ 1 \end{bmatrix} \tag{7.62}$$

(7.63) $$null(\mathbf{Z}) = \begin{bmatrix} \tan(\omega_2 t_{11})\tan(\omega_1 t_{22}) \\ -\tan(\omega_2 t_{11}) \\ -\tan(\omega_1 t_{22}) \\ 1 \end{bmatrix}$$

While these are indeed closed-form solutions, exhaustive research in the literature revealed that in general, there is no special relationship between the nullspace of a sum or difference and the sum or difference of nullspaces. I.e.,

(7.64) $$null(\mathbf{A}-\mathbf{B}) \neq null(\mathbf{A}) - null(\mathbf{B})$$

Therefore, this doesn't shed any light on the nullspace of **D**. A few parting observations on this pair of equations.

1) They are identities. The nullspace will always exist regardless of frequency, i.e., even if frequency estimates are completely wrong. Therefore they do not give us any useful information on frequency estimation.

2) The rank of both **Q** and **Z** is always 3, again, regardless of frequency.

3) These nullspaces depend only on frequency and not on phase. It is puzzling that the nullspace of the difference matrix depends only on phase and not on frequency. (This author has no clear explanation, although not for lack of trying.)

4) Neither member of this pair of nullspaces depends in any way on the filter terms $H_i(\omega_j)$.

5) Neither member of the pair depends on the times t_{12} or t_{21}. They depend only on t_{11} and t_{22}. This is also very puzzling, as all 4 times seem to appear symmetrically in the matrices, as can be seen by going back to the original derivation.

6) The nullspaces are extremely compact, given the extreme complexity of the two matrices.

7) There is an interesting duality in the form of the pair.

8) The nullspaces of both **Q** and **Z** obey the same product condition as does the phase vector **v**. The product of elements 1 and 4 equals the product of elements 2 and 3. However, this is identically true, and so does not yield any criterion for sifting correct frequency estimates from incorrect estimates.

9) Despite much effort, this author could glean nothing useful from these relations, as of this writing.

7.4.2 Method of Dai-Jones

Intense research of the literature uncovered the following gem. In a paper by (Dai and Jones, 2002), there appears a systematic formula for constructing a nullspace of a matrix without use of any numerical procedures. The paper expands on a method described by (Aitken, 1964). The method is as elegant as Cramer's rule, as it gives an explicit formula for the homogeneous case. This author is surprised that what seems like such a fundamental method for solving homogeneous equations seems to be relatively unknown and buried in obscure literature.

The method involves constructing cofactors of an augmenting row of the matrix at hand. Recall that the cofactor of an element is the determinant of the matrix formed by the remaining rows and columns, after the row and column of the current element is eliminated. The sign of each cofactor alternates, beginning with a plus sign in the upper left corner, and changing as the element in question moves along successive rows and columns.

We therefore have the following explicit formula for **v**. For simplicity, we use Matlab® syntax, in which submatrices consisting of particular rows and columns from a larger matrix can be easily specified.

$$\mathbf{v} = \begin{bmatrix} \det\{\mathbf{D}([1\ 2\ 3],[2\ 3\ 4])\} \\ -\det\{\mathbf{D}([1\ 2\ 3],[1\ 3\ 4])\} \\ \det\{\mathbf{D}([1\ 2\ 3],[1\ 2\ 4])\} \\ -\det\{\mathbf{D}([1\ 2\ 3],[1\ 2\ 3])\} \end{bmatrix} \tag{7.65}$$

We explain in words the meaning of this expression. One computes cofactors along the bottom row of **D**. This means each element of **v** is actually a 3×3 determinant of a submatrix of **D** with the appropriate alternating sign. One starts at the lower-left corner of **D**. One strikes out the first

column and the bottom row. One computes the 3×3 determinant of the top 3 rows in the second through fourth columns. This becomes the first element of **v,** with a positive sign.

One next moves one column to the right, remaining in the bottom row. One now strikes out the bottom row and the second column. One then computes the determinant of the top three rows and columns 1,3,4. This becomes the second element of **v,** with a negative sign**.**

One next moves one more column to the right, remaining in the bottom row. One now strikes out the bottom row and the third column. One then computes the determinant of the top three rows and columns 1,2,4. This becomes the third element of **v,** with a positive sign**.**

Finally, one moves one more column to the last one on the right, remaining in the bottom row. One now strikes out the bottom row and the fourth column. One then computes the determinant of the top three rows and the first 3 columns, i.e., 1,2,3. This becomes the fourth element of **v,** with a negative sign**.**

The proof is found in the above-referenced paper. We verified both numerically and symbolically. (Symbolic Toolbox confirms that multiplication of the resulting symbolic nullspace by the rows and columns of symbolic matrix **D** yields **0,** identically. Although Matlab® could not on its own compute the nullspace, it was able to verify the above formula, despite requiring cancellation of hugely complicated expressions of trigonometric functions which were not at all obvious.)

Recall that, as before, one needs to take the arctangent of the appropriate ratio of the elements of **v** to compute the individual phases.

7.5 Graphical Bounds on Frequency

We discuss one further issue with regard to frequency estimation that gives some insight into the behavior of local maxima in mixtures of various types. Referring back to the first equation in set (7.1), we have a mixture of two sines in two channels which are differently weighted, according to the respective functions $H_1(f)$ and $H_2(f)$. For simplicity, we assume the weights are as follows:

$$H_1(f_1) = 2$$
$$H_1(f_2) = 1$$
$$H_2(f_1) = 1$$
$$H_2(f_2) = 2$$

For this example we let

$$f_1 = 5$$
$$f_2 = 9$$

We then have

$$x_1 = H_1(f_1)\sin(2\pi f_1 t) + H_1(f_2)\sin(2\pi f_2 t) = 2\sin(2\pi f_1 t) + \sin(2\pi f_2 t)$$

$$x_2 = H_2(f_1)\sin(2\pi f_1 t) + H_2(f_2)\sin(2\pi f_2 t) = \sin(2\pi f_1 t) + 2\sin(2\pi f_2 t)$$

The two signals are plotted in Figure 96.

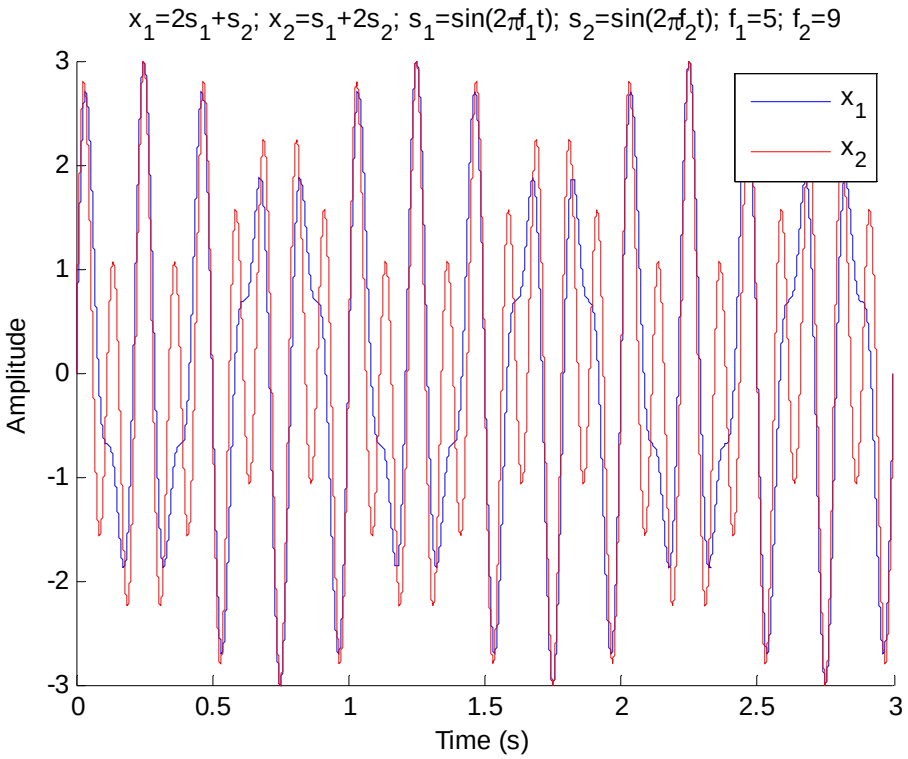

Figure 96. Two differently weighted mixtures of a pair of sinusoids of frequencies 5 and 9 Hz, respectively.

The peaks of the two mixtures are at different locations than the peaks of the original sines, as expected. Figure 97 illustrates the times at which peaks of each mixture occur relative to the underlying sines.

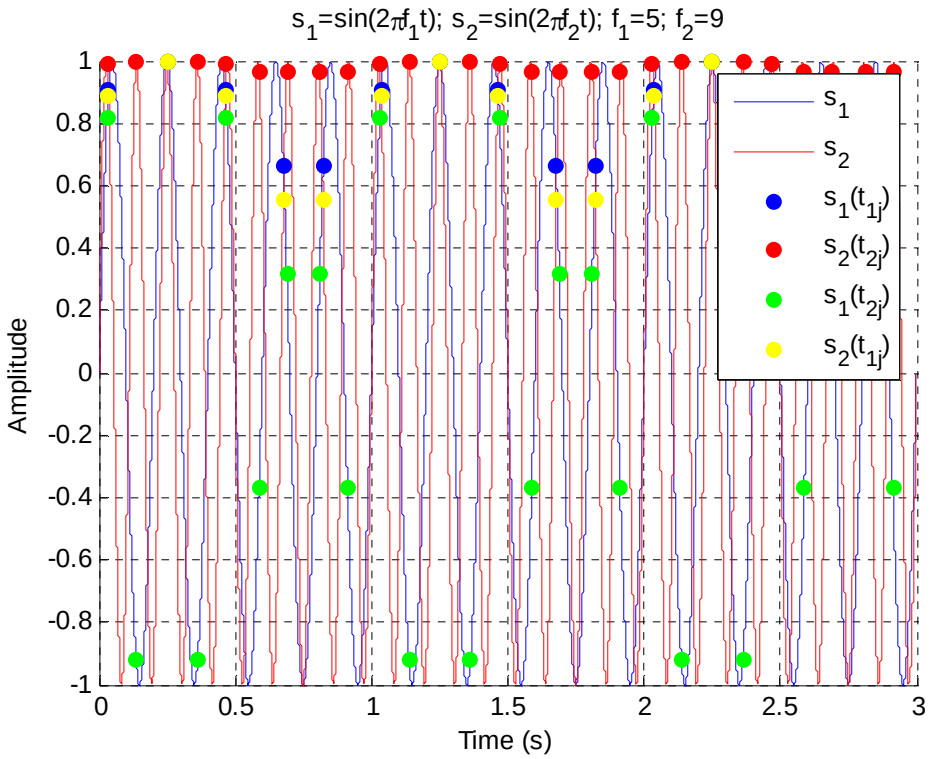

Figure 97. The original sines are shown with markers indicating times of peaks of the two mixtures. As expected, they do not coincide with the times of the peaks of the original sines. Blue dots mark times of peaks of mixture x1 on s1. Yellow marks times of peaks of x1 on s2. Red marks times of peaks of x2 on s2. Green marks times of peaks of x2 on s1. Red and green always line up vertically, as do blue and yellow.

From the derivative equations of set (7.3) at beginning of the chapter, we have for any peak j in channel 1 or 2, using the parameters of our example, that

(7.66) $$\frac{\cos(\omega_1 t_{1_j} + \phi_1)}{\cos(\omega_2 t_{1_j} + \phi_2)} = -\frac{a_2\omega_2 H_1(\omega_2)}{a_1\omega_1 H_1(\omega_1)} = .9000$$

(7.67) $$\frac{\cos(\omega_2 t_{2_j} + \phi_2)}{\cos(\omega_1 t_{2_j} + \phi_1)} = -\frac{a_1\omega_1 H_2(\omega_1)}{a_2\omega_2 H_2(\omega_2)} = .2778$$

Figure 98 shows a plot of the two cosines c_1 and c_2. (c_2 is inverted due to the negative sign in the equation.)

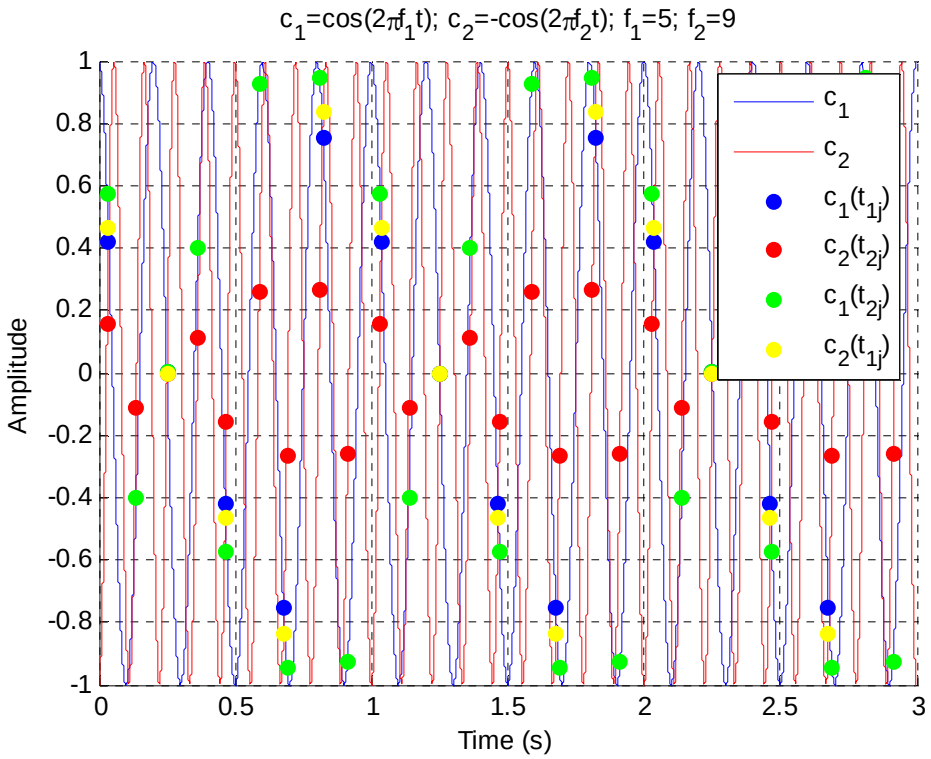

Figure 98. The two cosines c1 and c2 which arise from the derivative equations are plotted with markers indicating times of peaks in the two mixtures. Blue dots mark times of peaks of mixture x1 on c1. Yellow marks times of peaks of x1 on c2. Red marks times of peaks of x2 on c2. Green marks times of peaks of x2 on c1. Red and green always line up vertically, as do blue and yellow. From this figure, bounds on frequency estimates can be deduced, as in text.

Equation (7.66) indicates the ratio of the heights of the blue dots to the yellow dots. Equation (7.67) indicates the ratio of the heights of the red dots to the green dots. Since all the dots lie on cosine functions, and hence are limited to being ≤ 1, therefore the range of the red dot is restricted to ± 0.2778. The significance of this is seen when translating this height limitation to a time limitation. The time between successive red dots can be slightly more or slightly less than one period or 2π, but is limited to a fixed tolerance ε around the true period of the cosine c_2. Therefore, frequency estimates that one might make based on intervals between the peaks of the mixture x_2, while not exact, are limited to lie within a bounded radius surrounding the true value of the frequency of the original sinusoid s_2, since c_2 and s_2 share the same period.

A number of important observations follow:

1) The reason why we focused on bounding c_2, rather than c_1, is that it has a tighter bound. This is due to its higher frequency, and is thus another manifestation of the

fact we have noted on a number of occasions that for equal amplitude signals, the peaks of the higher frequency component dominate the estimation process.

2) It is possible for peaks of a mixture to yield estimates that are lower than the lowest or higher than the highest frequency component. Estimates based on peaks of mixtures do not always fall in between the two extremes. This is apparent from our example in which peaks may fall less than 2π from the previous peak, thus seeming to indicate a shorter period and higher frequency than the true value of 9 Hz.

3) This discussion highlights the strengths and weaknesses that underpin all of our numeric algorithms of Chapter 5 which were based on using peaks to initially estimate frequencies of mixture components. We again see that true accuracy is only obtained in a monocomponent signal, and not a mixture. The algorithms attempt to purify these initial estimates through iterative methods.

4) Since, as we have shown in this chapter that all other parameters follow from frequency, we have a range in which to do a possible grid search, and can use the two constraints of Section 7.3 to check the accuracy of each pair of frequency estimates.

7.6 Summary

We started with 8 coordinates from two pairs of local maxima in two bands. We needed to solve for 6 unknowns: the amplitudes, frequencies and phases of two sinusoids. We reduced the problem to finding the two frequencies, from which all the other variables can be found. In addition, we have found specific constraints on the two frequencies in the form of determinant equations, thus further limiting the possibilities.

These methods complement the computational work, and lend further support to the idea that the local maxima of multiple bands can determine the instantaneous parameters of the underlying signals, without the need for past history or future values.

The final step, to find an explicit formula for the frequencies remains elusive. In the absence, we suggested approximating with estimates based on peak intervals, and to use constraints to check accuracy.

It often happens, but is not always the case, that if one can organize a set of relations into matrix form, it is extendable to higher dimensions without major changes. If the frequencies can be found explicitly, the next logical step would be to work on this extension. These are both left for the future.

Chapter 8

Summary and Conclusion

8.1 Summary

We have examined the feasibility of using the assumption of comodulation as a basis for audio source separation. In Chapter 2, we saw that a number of researchers over the years have considered common amplitude and frequency trajectories to be useful cues for source separation. We posited that the reason why systems thus far have not been overly successful in separating mixtures of audio sources is due to an inability to properly allocate energy from multiple overlapping sources within a single frequency band. We noted that some authors disallow for the possibility, altogether, while others have imposed limits of about 25 Hz minimum separation between harmonics of competing sources. We also saw that at least one researcher wrote that merely developing reliable methods for visually separating sources in the time-frequency plane is a worthy goal, even if an actual resynthesis cannot be performed, so difficult is the source separation problem.

In Chapter 3 we examined the waveforms of various instruments and found that in certain cases there is evidence that comodulation may occur to a significant extent. The altoflute exhibited strong amplitude comodulation, the violin exhibited strong frequency comodulation, and the trumpet exhibited both.

We also examined the difference between interference and modulation, since both cause fluctuations in band envelopes. We suggested that study of the behavior of multiple harmonics can distinguish the two under the assumption of comodulation.

In Chapter 4 we developed a source-separation approach for constant-frequency coherent sources based on the technique of Non-Negative Matrix Factorization. We proved that in

general, the solution is non-unique, but under certain constraints involving the rows and columns of the matrix, a unique solution does exist. In our application, this translated to the requirement of a unique onset time and a single non-overlapped frequency channel per source, even if all remaining channels do contain overlapping contributions from multiple sources. This approach correctly allocates energy among competing sources in all the remaining channels. We offered both analytical and computational methods for finding this solution when it exists. We noted the biological plausibility of onset detection based on the accepted existence of cells within the cochlear nucleus that exhibit such a response.

In Chapter 5 we began to look at the general case of amplitude- and frequency-varying signals including speech waveforms, and saw that mixtures of different speakers exhibit interference within channels. We began to devote the rest of the thesis to the essential problem of how to recognize and separate closely spaced components from different sources on extremely short time scales. Because of the continually time-varying nature of real-world signals, there is only a very short time available to perform frequency analysis before the frequencies change. We sought ways to overcome the uncertainty principle.

We examined the use of information from multiple simultaneous frequency bands to obtain improved accuracy over what is achievable from a single band alone. We found that slight changes in the positions of local maxima from band to band can be exploited to uncover the frequencies of the original sources. We further explored properties of overlapping exponential filter banks and noted that they have properties which make them well-suited for dimensionality reduction, i.e., consolidating information from multiple channels and determining which is novel information, and which is redundant. In this way they assist in uncovering the true number of sources, which is not obvious from the often diverse channel data, in general. We noted possible biological plausibility here, as well, in the resemblance of auditory neural frequency-response curves to exponential shapes, and in the accepted concept of neural synchrony or phase locking (preferential firing at particular portions of the waveform), which fits well with our use of local maxima for enhanced frequency estimation. We compared three different variations on this theme with simple harmonic test sets. We also noted considerations for practical filter design in conjunction with these algorithms.

In Chapter 6 we continued with more realistic test sets involving AM and FM signals, and the addition of noise. We compared the performance of the Simultaneous-Equation method with the FFT, and noted that it is not encumbered by restrictions to predetermined digital frequencies, as is the FFT, and resolution appeared to surpass that of the FFT. We further conducted tests on actual speech recordings. We found that the sensitivity of our methods seemed to be high enough to resolve actual sidebands of modulated signals. These often appear to lie within a few Hz of each other, much less than the 25 Hz resolution limit noted by some other authors. We also noted that the results seemed to display time-locality, as well, and seem to fit at least to a certain degree the notion of instantaneous parameter estimation. Based on these results, we tried to better understand whether a conventional definition of instantaneous frequency is useful in practice, or whether its applicability needs to be rethought. We cited other authors who have themselves expressed doubts for reasons of their own. Finally we tried to qualitatively describe the effect of filter transient response on the performance of our algorithms.

In Chapter 7 we attempted to find closed-form solutions to the problem of exactly determining the amplitudes, frequencies, and phases of a mixture of two sinusoids given the positions of the local maxima (times and heights) of two differently weighted versions of the signal. We succeeded in showing that given the frequencies alone, one can analytically determine phases and amplitudes. We still seek the elusive last step of analytically determining the two frequencies. We state conditions that they must satisfy in terms of multi-term determinant equations, but do not yet have a complete solution.

8.2 Future Work

For the future, in addition to bolstering the analytical work underlying all our algorithms , we would like to further optimize filter design. Testing has been extremely slow due to the computational requirement of extremely high sampling rates to precisely locate the local maxima. While the use of interpolation has eased that burden to some extent, still more computational power is required than we have available. In addition, interpolation carries with it various inaccuracies of its own. We strongly believe that a hardware-based continuous time system would be the ideal platform upon which to do further testing of these types of

algorithms. We further note that it would be interesting to see whether additional increases in resolution can be achieved with the addition of more and more frequency bands, thus decreasing filter spacing beyond the 0.5 Hz that we are limited with our current platform, and by continuing to increase the sampling rate. We have suggested other enhancements in appropriate places in our work, for example the addition of a calibration step in Section 5.8.2. Further study of transient effects is necessary, as is a better understanding of speech sidebands.

A major task is the actual use of the algorithm to analyze whole utterances, and to resynthesize complete streams based on this analysis. One would need to run the algorithm at closely spaced points in time and follow the progression of spectral components. This would require very high computational power in the algorithm's current implementation. In the absence of a dedicated, parallel, analog platform or increased digital computing power, there is room for optimization of the code to make it more efficient. All convolution and interpolation operations need to be shortened to the exact minimum lengths necessary for accuracy. In the initial development of the various computational algorithms, we found it necessary to minimize all sources of error, thus being overly conservative in avoiding any shortcuts or truncation of data lengths, in order to understand the performance and limitations of the kernel of the algorithm, itself. When debugging, any unwarranted assumptions or extraneous code can only further confuse the situation, and make correct pinpointing of sources of error more difficult. However, once the core prototype has been found reliable, then certainly it is worthwhile and necessary to devote effort towards streamlining the actual implementation. A further benefit of trimming down the various convolution lengths is to see what is the minimum signal-length necessary to form a set of parameter estimates. This would give a better gauge of the degree of time-locality that the algorithm is capable of producing. In other words, how closely does it come to true instantaneous parameter estimation.

Among other directions that might be pursued, we suggest that it may be possible to further build on the Matrix Factorization approach in Chapter 4, by noting that once the row and column constraints are satisfied, the Non-Negativity requirement can possibly be relaxed. This would allow for the possibility of complex source signatures, thus properly accounting for phase effects, such as in the phase-cancellation example of Section 4.17.

A study of the best way to perform the grouping and resynthesis tasks is also a necessary component in developing an actual separation system. How can comodulation be incorporated in signals such as speech where more distant channel amplitudes may not be correlated with each other, but adjacent channels may in fact be (those that lie near the same formant)? Are sideband spacings a useful cue in detecting common modulation? How does one separate the AM and FM contributions to the sideband locations and trajectories? In cases of mixtures, how does one isolate the sidebands of each source when they appear to overlap, as we saw in the example of the speech mixture? Can we prevent sidebands from interfering with each other, so that their positions remain constant in the presence of other sources? Can we incorporate the framework of Section 3.5, where the use of multiple harmonics sheds light on the characteristics of a mixture?

We believe that the most rewarding part of all of our efforts was the work in Chapter 7, which attempts to provide a theoretical basis for the fundamental idea of using multiply weighted versions of a signal to uncover the underlying sources. As mentioned above, we found that for the 2-sinusoid case, a closed-form solution exists that can compute phases and amplitudes given the frequencies, using an elegant set of determinant equations. Work is needed for the last step, that of computing frequencies from the coordinates of the local maxima. If that can be done analytically, then an extension to higher dimensions should be sought. This may provide more information on some of the elusive questions we have raised in various places in this thesis, such as how tightly do filters need to be spaced, and how many filters are necessary to arrive at a unique solution. Possibly, incorporation of an exponential filter model, which we had found useful in the computational approaches of earlier chapters, may be of help here, as well, in simplifying the sets of equations, and in determining the optimal filter parameters. Ideally, we would like to replace the role of empirical observations in the choice of filter parameters with a systematic method or formula.

A closed-form solution will also allow us to separate and understand the effect of transients on algorithm performance, as we will be able to model them as a perturbation of the exact mathematical solution valid under ideal conditions. The idea is similar to studying the massless pulley or frictionless plane in beginning physics courses. Once one understands the basic principles involved, one can always add layers of complexity to account for more realistic

situations, but at least one has a frame of reference for what is to be ideally expected, and how far actual measurements deviate from this ideal.

Finally, a major advantage of the analytical approach is in computational efficiency. Our closed-form work returns an immediate answer, and doesn't require a long series of iterations which gradually approach the solution, but rather computes it in a single, well-prescribed step.

If the preceding tasks can be accomplished for a large number of channels and sources, then the final step is to let the filter spacing go to zero, and study what would be the effect of using an infinite number of filters (a continuum in frequency). Would this be a valid model for the auditory system, rather than using a discrete set of filters? Would such a continuum have the ability to represent and separate noiselike sources, whose spectra are themselves continuous in frequency? Can we use such a continuum to understand transients which we modeled as a spreading of spectral energy about the stimulus frequency in Section 6.8? There remain many interesting questions to further probe.

Appendix

Comparison of Independent Component Analysis and Comodulation

In recent years the method of ICA has become popular for source-separation tasks. We would therefore like to note a few similarities and differences between ICA and our comodulation approach.

The ICA formulation is based on a scenario in which there are assumed to be n sources and n detectors. We will look at a simple case where n=2. Each detector picks up sound from each source. However, they are weighted differently, in the sense that the first mike might pick up more of the first source than the second source, and vice versa for the second mike. The equations are as follows:

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t)$$
$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t)$$

where $s_1(t)$ and $s_2(t)$ are the source signals and $x_1(t)$ and $x_2(t)$ are the microphone outputs. The coefficients a_{11}, a_{12}, a_{21}, and a_{22} describe the weighting of the sources in each microphone. We can compact this into a matrix form, as we did for our comodulation equations in the earlier chapter as follows:

$$\mathbf{X} = \mathbf{AS}$$

A is called the mixing matrix. We are given only **X,** and need to find **A** and **S.** The number of sources are also unknown. We seek to find an unmixing matrix **W** which is the inverse of **A** so that we can recover the sources via

$$\mathbf{S} = \mathbf{WX}$$

We used a compact Matlab® algorithm which was provided by Prof. Gert Cauwenberghs, as follows:

```
clear
load numwav;% A file containing two sources to be mixed
mu=0.1;% learning rate
Id=eye(2);% Identity matrix
nrm=1/length(source1);
x=[source1+source2,source1-source2]';% Mix sources
W=0.1*(2*rand(2,2)-1);%unmixing matrix; initial
for iter=1:1000
  y=W*x;
  W=W+mu*(Id-nrm*sign(y)*y')*W;
end
```

After 1000 iterations, the algorithm is able to find the unmixing matrix **W**.

The theory of ICA is based on the fact that due to the Central Limit Theorem, sums of random variables tend to have Gaussian distributions. ICA algorithms, therefore, try to search for those combinations that are the least Gaussian in nature. How they do that is outside the scope of this work

The important point for our purposes is that the matrix formulation of ICA is very similar to our formulation of comodulation. The equations are identical. In both cases we need to factor a single matrix into two others. Such a factorization is not unique, as we have seen, unless additional constraints are placed on the system. For the comodulation formulation, we have used a nonnegativity constraint, coupled with requirements for each source to be a soloist at one point in time, and to have a component at one frequency which is not overlapped by any other source. However, this will not work in the ICA formulation, as the vectors in question are the entire waveforms, not the modulation envelopes. Envelopes do not go negative, so the nonnegativity constraint is natural when dealing with modulation envelopes. However, the actual waveforms do go negative, so one could not use such a constraint for ICA.

But what about the reverse? Could one use the algorithms of ICA to solve the comodulation set of equations? We attempted this for a simple case of two oppositely ramped envelopes, and found that it did not work. We illustrate the results. The first figure below shows the original ramped envelopes. The second shows a typical set of recovered envelopes. We note that in practice, the recovered envelopes are different in each trial, probably due to the random initial guess of the ICA algorithm (line 7) which changes from trial to trial.

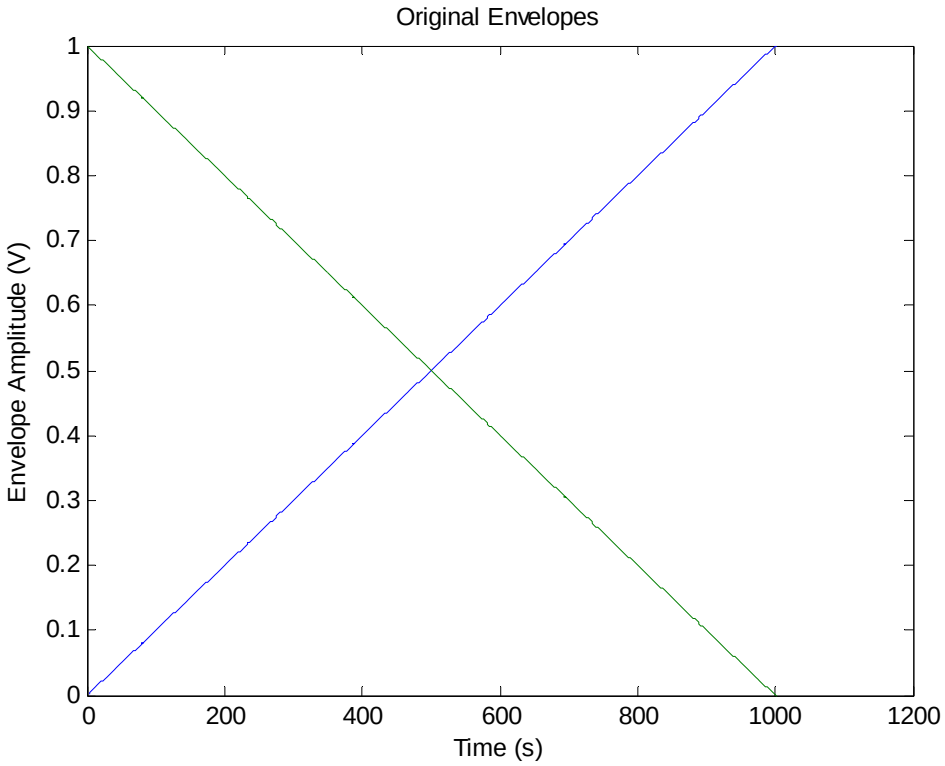

Figure 99. Envelopes of original pair of oppositely ramped signals that were added together and were to be unmixed.

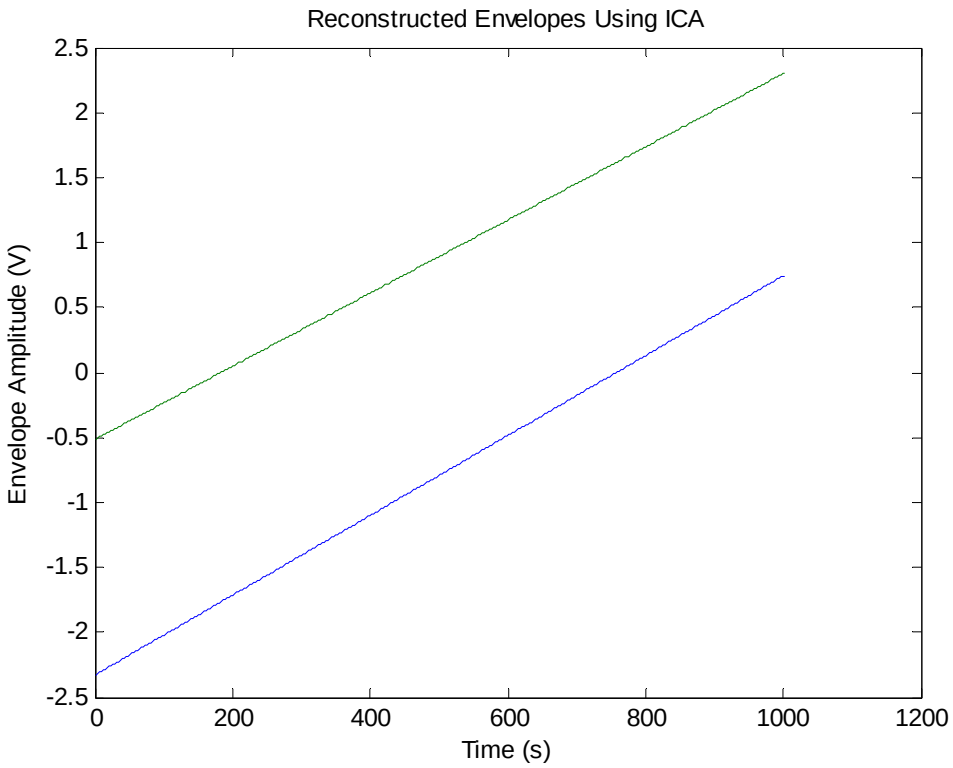

Figure 100. The pair of recovered envelopes using the ICA algorithm. Method fails to recover original pair for reasons discussed in text.

The reason for the failure is that ICA looks for a mixture matrix which yields the most non-Gaussian decomposition. However, for cases of simple envelope shapes such as a ramp, there is really no difference in the statistical distribution among the various ways of decomposing the sum of two ramps (which total a constant 1 volt) into various combinations. Only when dealing with waveforms which have extremely complex shapes would significant statistical differences appear. So the reverse is true as well, the ICA algorithm is not useful for the comodulation formulation. This behavior was correctly predicted by (Zweig, 2001) in personal discussions.

One final comment on this matter. In the comodulation formulation, we have a set of fixed source signature vectors (frequency is fixed) which are multiplied by a set of amplitude modulation vectors (functions of time). The product gives the band output envelopes which are roughly analogous to a spectrogram of the scene. In the ICA formulation, we have a set of fixed weights multiplied by a set of source signals (functions of time). The product is the set of microphone outputs. Note that if the frequency changes in the comodulation case, then the entire formulation is invalid, as we no longer have a product of two vectors in the usual sense. Similarly, in the comodulation case, if the weights change, then we can no longer can express

our original equation as a product. This points out a limitation of ICA for real world use. Having fixed weights implies that there is no movement between the mikes and the subjects for the entire duration of the recording. This would be difficult in a real time situation, as people tend to move around to some extent. For artificial mixtures of recordings, the method works well, where a single mixing matrix is initially applied to the entire duration of the source recordings. A formulation which can handle FM signals in the comodulation situation would be equivalent to solving the ICA equations in the case of moving targets.

Still another complex issue currently being studied by the ICA community is the problem of reverberation, since multiple delayed and weakened copies of each source from room reflections are added to the direct sources that are picked up by the microphones. This represents a convolution operation. (Zweig, 2001) also informally demonstrated that a delay by even a single sample of the recorded version of a source is enough to completely throw off current ICA algorithms. In order to correctly handle reverberation, an inverse filter needs to be found for each source-mike pair, not a just a single weight. (M. Cohen and Cauwenberghs, 1998) attempt such a solution in which the algorithm tries to calculate a matrix of filters, rather than the conventional matrix of weights, via a parallel, stochastic gradient-descent method. Further discussion is beyond the scope of this thesis.

References

A. C. Aitken (1964). *Determinants and Matrices*. Edinburgh, Oliver and Boyd.

A. G. Akritas, G. I. Malaschonok and P. S. Vigklas (2006). "The SVD-Fundamental Theorem of Linear Algebra." *Nonlinear Analysis: Modeling and Control* **11**(2): 123-136.

C. Alain (2005). Speech Separation: Further Results from Recordings of Event-related Brain Potentials in Humans. In *Speech Separation by Humans and Machines*. P. Divenyi, Ed., Kluwer Academic Publishers**:** 13-30.

A. J. Bell and T. J. Sejnowski (1995). "An information-maximization approach to blind separation and blind deconvolution." *Neural Computation* **7**: 1129-1159.

A. C. Bovik, J. P. Havlicek and M. D. Desai (1993). "Theorems for discrete filtered modulated signals." *IEEE International Conference on Acoustics, Speech, and Signal Processing* **3**: 153-156.

A. S. Bregman (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, MIT Press.

G. J. Brown (1992). Computational auditory scene analysis: a representational approach, Ph.D. Thesis, University of Sheffield.

G. J. Brown and M. Cooke (1994). "Computational auditory scene analysis." *Computer Speech & Language* **8**(44): 297-336.

J. F. Brugge (1996). Hearing and Balance, Department of Neurophysiology, University of Wisconsin.

J. F. Brugge, D. J. Anderson, J. E. Hind and J. E. Rose (1969). "Time structure of discharges in single auditory nerve fibers of the squirrel monkey in response to complex periodic sounds." *Journal of Neurophysiology* **32**: 386-401.

J. A. Cadzow (1988). "Signal enhancement—a composite property mapping algorithm." *IEEE Transactions on Acoustics, Speech and Signal Processing* **36**(1): 49-62.

J. Capon (1969). "High Resolution Frequency-Wavenumber Spectrum Analysis." *Proceedings of the IEEE* **57**(8): 1408-1418.

P. Cariani (2005). Recurrent Timing Nets for F0 Based Speaker Separation. In *Speech Separation by Humans and Machines*. P. Divenyi, Ed., Kluwer Academic Publishers**:** 31-54.

R. P. Carlyon (1991). "Discriminating between coherent and incoherent frequency modulation of complex tones." *Journal of the Acoustical Society of America* **89**(1): 329-340.

R. P. Carlyon (1994). "Further evidence against an across-frequency mechanism specific to the detection of frequency modulation (FM) incoherence between resolved frequency components." *Journal of the Acoustical Society of America* **95**(2): 949-952.

J. C. Catford (1988). *A Practical Introduction to Phonetics*. Oxford, Clarendon Press.

G. Cauwenberghs (1999). "Monaural Separation of Independent Acoustical Components." *Proceedings of the 1999 IEEE International Symposium on Circuits and Systems (ISCAS 1999)* **5**: 62-65.

L. Cohen (1995). *Time Frequency Analysis*. Englewood Cliffs, NJ, Prentice-Hall.

M. Cohen and G. Cauwenberghs (1998). "Blind separation of linear convolutive mixtures through parallel stochastic optimization." *Proceedings of the 1998 IEEE International Symposium on Circuits and Systems (ISCAS 1998)* **3**: 17-20.

M. P. Cooke (1991). Modeling auditory processing and organization, Ph.D. Thesis, Department of Computer Science, University of Sheffield.

J. S. Dai and J. R. Jones (2002). "Null-Space Construction Using Cofactors from a Screw-Algebra Context." *Proceedings: Mathematical, Physical and Engineering Sciences* **458**(2024): 1845-1866.

C. J. Darwin and V. Ciocca (1992). "Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component." *Journal of the Acoustical Society of America* **91**(1): 2281–3390.

L. Demany and C. Semal (1986). "On the detection of amplitude modulation and frequency modulation at low modulation frequencies." *Acustica* **61**: 243-255.

P. Divenyi (2005). *Speech Separation By Humans And Machines*, Kluwer Academic Publishers.

J. Durbin (1959). "The fitting of time-series models." *Review of the International Statistical Institute* **28**: 229–249.

B. W. Edwards and N. Viemeister (1994). "Modulation detection and discrimination with three-component signals." *Journal of the Acoustical Society of America* **95**(4): 2202-2212.

D. P. W. Ellis (1996). Prediction-driven computational auditory scene analysis. Cambridge, MA, Ph.D. Thesis, Department of Electrical Engineering, MIT.

G. Fant (1960). *Acoustic Theory of Speech Production*, Mouton De Gruyter.

S. A. Fulop and K. Fitz (2006). "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications." *Journal of the Acoustical Society of America* **119**(1): 360-371.

D. Geisler (1998). *From Sound to Synapse: Physiology of the Mammalian Ear*, Oxford University Press.

J. E. Greenberg and P. M. Zurek (1992). "Evaluation of an adaptive beamforming method for hearing aids." *Journal of the Acoustical Society of America* **91**(3): 1662-1676.

L. J. Griffiths and C. W. Jim (1982). "An alternative approach to linearly constrained adaptive beamforming." *IEEE Transactions on Antennas and Propagation* **30**: 27-34.

J. H. Grose and J. W. Hall III (1996). "Across-frequency processing of multiple modulation patterns." *Journal of the Acoustical Society of America* **99**(1): 534-541.

C. F. Halpin (1997). Lecture at Massachusetts Eye and Ear Infirmary, Department of Audiology.

B. A. Hanson and D. Y. Wong (1984). "The harmonic magnitude suppression technique for intelligibility enhancement in the presence of interfering speech." *IEEE International Conference on Acoustics, Speech, and Signal Processing* **9**: 65-68.

W. M. Hartmann (1998). *Signals, Sound and Sensation*, American Institute of Physics.

W. M. Hartmann and G. M. Hnath (1982). "Detection of Mixed Modulation." *Acustica* **50**: 297-312.

S. Haykin (1986). *Adaptive Filter Theory*, Prentice Hall.

A.-M. Higgins (2001). Timbre for Transition Year. **1**.

N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung and H. H. Liu (1998). "The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis." *Proceedings of the Royal Society of London: Mathematical, Physical and Engineering Sciences* **A454**(1971): 903-995.

A. Hyvarinen and E. Oja (2000). "Independent Component Analysis: Algorithms and Applications." *Neural Networks* **13**(4-5): 411-430.

M. Ito and M. Yano (2007). "Sinusoidal modeling for nonstationary voiced speech based on a local vector transform." *Journal of the Acoustical Society of America* **121**(3): 1717-1727.

B. D. Jacobson, G. Cauwenberghs and L. M. Litvak (2001). "A mathematical theory of comodulation." *Journal of the Acoustical Society of America* **109**(5): 2494.

D. H. Johnson (1980). "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones." *Journal of the Acoustical Society of America* **68**(4): 1115-1122

G. Jones and B. Boashash (1990). "Instantaneous frequency, instantaneous bandwidth and the analysis of multicomponent signals." *IEEE International Conference on Acoustics, Speech, and Signal Processing* **5**: 2467-2470.

S. M. Kay (1988). *Modern Spectral Analysis: Theory and Application*, Prentice Hall.

S. M. Kay and S. L. Marple Jr (1981). "Spectrum Analysis—A Modern Perspective." *Proceedings of the IEEE* **69**(11): 1380-1419.

M. Kubovy (1981). Concurrent-pitch segregation and the theory of indispensable attributes. In *Perceptual organization*. M. Kubovy and J. R. Pomerantz, Eds. Hillside, NJ, Lawrence Erlbaum Assoc.**:** 55-98.

R. T. Lacoss (1971). "Data Adaptive Spectral Analysis Methods." *Geophysics* **36**(4): 661-675.

D. D. Lee and H. S. Seung (1999). "Learning the parts of objects by non-negative matrix factorization." *Nature* **401**(6755): 788-791.

T.-W. Lee (2005). Blind Source Separation Using Graphical Models. In *Speech Separation by Humans and Machines*. P. Divenyi, Ed., Kluwer Academic Publishers**:** 55-64.

N. Levinson (1947). "The Wiener RMS (root mean square) error criterion in filter design and prediction." *Journal of Mathematical Physics* **25**(4): 261–278.

B. Libbey (2006). Frequency Modulation. National Instruments Developer Zone.

J. S. Lim and D. W. Griffin (1985). "A New Model-Based Speech Analysis/Synthesis System." *NSA Speech Research Symposium, San Diego, CA* **2**.

J. S. Lim, A. Oppenheim and L. Braida (1978). "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition." *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**(4): 354-358.

J. Y. Lin and W. M. Hartmann (1998). "The pitch of a mistuned harmonic: Evidence for a template model." *Journal of the Acoustical Society of America* **103**(5): 2608-2617.

B. F. Logan Jr (1977). "Information in the zero crossings of bandpass signals." *Bell System Technical Journal* **56**: 487-510.

P. J. Loughlin and B. Tacer (1997). "Comments on the interpretation of instantaneous frequency." *IEEE Signal Processing Letters* **4**(5): 123-125.

D. Maiwald (1967a). "Die Berechnung von Modulationsschwellen mit Hilfe eines Funktionsschemas." *Acustica* **18**: 193-207.

D. Maiwald (1967b). "Ein Funktionsschema des Gehors zur Beschreibung der Erkennbarkeit kleiner Frequenz-und Amplitudenanderungen." *Acustica* **18**: 81–92.

L. Mandel (1974). "Interpretation of Instantaneous Frequencies." *American Journal of Physics* **42**(10): 840-846.

S. McAdams (1984). Spectral fusion, spectral parsing and the formation of auditory images, Ph.D. Thesis, Stanford University.

H. McGurk and J. MacDonald (1976). "Hearing lips and seeing voices." *Nature* **264**(5588): 746-48.

B. C. J. Moore and A. Sek (1992). "Detection of combined frequency and amplitude modulation." *Journal of the Acoustical Society of America* **92**(6): 3119-3131.

B. C. J. Moore and A. Sek (1995). "Effects of carrier frequency, modulation rate, and modulation waveform on the detection of modulation and the discrimination of modulation type (AM vs. FM)." *Journal of the Acoustical Society of America* **97**(4): 2468-2478.

J. A. Naylor and S. F. Boll (1987). "Techniques for suppression of an interfering talker in co-channel speech." *IEEE International Conference on Acoustics, Speech, and Signal Processing* **12**: 205-208.

W. Nho and P. J. Loughlin (1999). "When is instantaneous frequency the average frequency at each time?" *IEEE Signal Processing Letters* **6**(4): 78-80.

K. Nishi and S. Ando (1998). "An Optimal Comb Filter for Time-Varying Harmonics Extraction." *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences* **81**(8): 1622-1627.

A. M. Noll (1967). "Cepstrum Pitch Determination." *Journal of the Acoustical Society of America* **41**(2): 293-309.

K. Nordberg and G. Farneback (2001). Rank complement of diagonalizable matrices using polynomial functions. Report LiTH-ISY. Linkoping, Sweden, Computer Vision Laboratory, Linkoping University.

P. M. Oliveira and V. Barroso (1999). "Instantaneous frequency of multicomponent signals." *IEEE Signal Processing Letters* **6**(4): 81-83.

T. W. Parsons (1975). Separation of simultaneous vocalic utterances of two talkers, Ph.D. Thesis, Polytechnic Institute of New York.

T. W. Parsons (1976). "Separation of speech from interfering speech by means of harmonic selection." *Journal of the Acoustical Society of America* **60**(4): 911-918.

P. M. Peterson (1989). Adaptive array processing for multiple microphone hearing aids. Cambridge, MA, Ph.D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

P. M. Peterson, S. M. Wei, W. M. Rabinowitz and P. M. Zurek (1990). "Robustness of an adaptive beamforming method for hearing aids." *Acta Otolaryngol Suppl* **469**: 85-90.

V. F. Pisarenko (1973). "The retrieval of harmonics from covariance functions." *Geophysical Journal of the Royal Astronomical Society* **33**(3): 347-366.

A. N. Popper and R. R. Fay (1992). *The Mammalian Auditory Pathway*, Springer.

B. G. R. d. Prony (1795). "Essai experimental et analytique: sur les lois de la dilatabilite de uideselastique et sur celles de la force expansive de la vapeur de l'alkool, direntes temperatures." *Journal de l'Ecole Polytechnique* **1**(22): 24-76.

T. F. Quatieri (2002). *Discrete-time speech signal processing: principles and practice*, Prentice Hall.

T. F. Quatieri (2004). Personal Communication.

T. F. Quatieri and R. J. Danisewicz (1990). "An Approach to Co-Channel Talker Interference Suppression Using a Sinusoidal Model for Speech." *IEEE Transactions on Acoustics, Speech, and Signal Processing* **38**(1): 56-69.

T. F. Quatieri, T. E. Hanna and G. C. O'Leary (1997). "AM-FM Separation Using Auditory-Motivated Filters." *IEEE Transactions on Speech and Audio Processing* **5**(5): 465-480.

S. Roweis (2005). Automatic Speech Processing by Inference in Generative Models. In *Speech Separation by Humans and Machines*. P. Divenyi, Ed., Kluwer Academic Publishers**:** 97-134.

R. Roy and T. Kailath (1989). "ESPRIT-estimation of signal parameters via rotational invariancetechniques." *IEEE Transactions on Acoustics, Speech, and Signal Processing* **37**(7): 984-995.

G. P. Scavone, J. S. Abel and D. P. Berners (2007). MUS320: Introduction to Digital Audio Signal Processing, Online Course Notes, Center for Computer Research in Music and Acoustics (CCRMA), Department of Music, Stanford University.

E. D. Scheirer (2000). Music-Listening Systems. Cambridge, MA, Ph.D. Thesis, Program in Media Arts and Sciences, School of Architecture and Planning, Massachusetts Institute of Technology.

M. Schiff (1997). Spectrum Analysis Using Digital FFT Techniques. Application Note AN106A, Elanix, Inc.

R. Schmidt (1986). "Multiple emitter location and signal parameter estimation." *IEEE Transactions on Antennas and Propagation* **34**(3): 276-280.

M. R. Schroeder (1968). "Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurement." *Journal of the Acoustical Society of America* **43**(4): 829-834.

S. Seneff (1984). "Pitch and spectral estimation of speech based on auditory synchrony model." *IEEE International Conference on Acoustics, Speech, and Signal Processing* **9**: 45-48.

V. C. Shields (1970). Separation of additive speech signals by digital comb filtering. Cambridge, MA, M.S. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

B. Shivapuja (1991). Acoustics, Physiological. Encyclopedia of Applied Physics. M. A. Ruggero and M. N. Semple, VCH Publishers. **1**.

M. Slaney (2005). The History and Future of Computational Auditory Scene Analysis. In *Speech Separation by Humans and Machines*. P. Divenyi, Ed., Kluwer Academic Publishers**:** 199-212.

M. Slaney, D. Naar and R. Lyon (1994). "Auditory model inversion for sound separation." *IEEE International Conference on Acoustics, Speech, and Signal Processing* **2**: 77-80.

P. Smaragdis (2005). Exploiting Redundancy to Construct Listening Systems. In *Speech Separation by Humans and Machines*. P. Divenyi, Ed., Kluwer Academic Publishers**:** 83-96.

R. Stern (2005). Signal Separation Motivated by Human Auditory Perception. In *Speech Separation by Humans and Machines*. P. Divenyi, Ed., Kluwer Academic Publishers**:** 135-154.

K. N. Stevens (1999). *Acoustic Phonetics*. Cambridge, MA, MIT Press.

P. Stoica and R. L. Moses (1997). *Introduction to spectral analysis*. Upper Saddle River, NJ, Prentice Hall.

G. Strang (1993). "The Fundamental Theorem of Linear Algebra." *American Mathematical Monthly* **100**(9): 848-855.

Q. Summerfield, A. Lea and D. Marshall (1990). "Modelling auditory scene analysis: strategies for source segregation using autocorrelograms." *Proceedings of the Institute of Acoustics* **12**(10): 507-514.

E. Terhardt (1974). "On the perception of periodic sound fluctuations (roughness)." *Acustica* **30**(4): 201–213.

W. P. Torres and T. F. Quatieri (1999). "Estimation of modulation based on FM-to-AM transduction: two-sinusoid case." *IEEE Transactions on Signal Processing* **47**(11): 3084-3097.

J. A. Tropp (2003). Literature Survey: Non-Negative Matrix Factorization, Institute for Computational Engineering and Sciences, The University of Texas at Austin.

D. Tufts and R. Kumaresan (1982). "Singular value decomposition and improved frequency estimation using linear prediction." *IEEE Transactions on Acoustics, Speech, and Signal Processing* **30**(4): 671-675.

J. Ville (1948). "Theorie et applications de la notion de signal analytique." *Cables et Transmission* **2**(1): 61–74.

A. L.-C. Wang (1994). Intantaneous and Frequency-Warped Signal Processing Techniques for Auditory Source Separation. Palo Alto, CA, Ph.D. Thesis, Department of Electrical Engineering, Stanford University.

D. L. Wang (2005). On Ideal Binary Mask as the Computational Goal of Auditory Scene Analysis. In *Speech Separation by Humans and Machines*. P. Divenyi, Ed., Kluwer Academic Publishers**:** 181-198.

D. L. Wang and G. J. Brown (1999). "Separation of speech from interfering sounds based on oscillatory correlation." *IEEE Transactions on Neural Networks* **10**(3): 684-697.

M. Weintraub (1985). A theory and computational model of auditory monaural sound separation, Stanford University.

B. Widrow (2001). "A microphone array for hearing aids." *Circuits and Systems Magazine, IEEE* **1**(2): 26-32.

B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong and R. C. Goodlin (1975). "Adaptive noise canceling: Principles and applications." *Proceedings of the IEEE* **63**(12): 1692–1716.

J. D. Wise, J. A. Caprio and T. W. Parks (1976). "Maximum likelihood pitch estimation." *IEEE Transactions on Acoustics, Speech, and Signal Processing* **24**(5): 418-423.

X. Yang, K. Wang and S. A. Shamma (1992). "Auditory representations of acoustic signals." *IEEE Transactions on Information Theory* **38**(2 Part 2): 824-839.

L. Young Jr and J. Goodman (1977). "The effects of peak clipping on speech intelligibility in the presence of a competing message." *IEEE International Conference on Acoustics, Speech, and Signal Processing* **2**: 216-218.

W. I. Zangwill (1969). *Nonlinear programming: a unified approach*, Prentice-Hall, Englewood Cliffs, NJ.

G. Zweig (2001). Private communication and demonstration of Signition platform.

E. Zwicker (1956). "Die elementaren Grundlagen zur Bestimmung der Informationskapazität des Gehörs." *Acustica* **6**: 365–381.