Model-driven Time-varying Signal Analysis and its Application to Speech Processing

by

Steven Sandoval

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved March 2016 by the
Graduate Supervisory Committee:

Antonia Papandreou-Suppappola, Chair
Julie Liss
Pavan Turaga
Narayan Kovvali

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

This work examines two main areas in model-based time-varying signal processing with emphasis in speech processing applications. The first area concentrates on improving speech intelligibility and on increasing the proposed methodologies application for clinical practice in speech-language pathology. The second area concentrates on signal expansions matched to physical-based models but without requiring independent basis functions; the significance of this work is demonstrated with speech vowels.

A fully automated Vowel Space Area (VSA) computation method is proposed that can be applied to any type of speech. It is shown that the VSA provides an efficient and reliable measure and is correlated to speech intelligibility. A clinical tool that incorporates the automated VSA was proposed for evaluation and treatment to be used by speech language pathologists. Two exploratory studies are performed using two databases by analyzing mean formant trajectories in healthy speech for a wide range of speakers, dialects, and coarticulation contexts. It is shown that phonemes crowded in formant space can often have distinct trajectories, possibly due to accurate perception.

A theory for analyzing time-varying signals models with amplitude modulation and frequency modulation is developed. Examples are provided that demonstrate other possible signal model decompositions with independent basis functions and corresponding physical interpretations. The Hilbert transform (HT) and the use of the analytic form of a signal are motivated, and a proof is provided to show that a signal can still preserve desirable mathematical properties without the use of the HT. A visualization of the Hilbert spectrum is proposed to aid in the interpretation. A signal demodulation is proposed and used to develop a modified Empirical Mode Decomposition (EMD) algorithm.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

LIST OF ABBREVIATIONS AND ACRONYMS

Symbol

AAoVSA	Automatic Assessment of Vowel Space Area
AM	Amplitude Modulation
AM–FM	Amplitude Modulation–Frequency Modulation
AR	Auto Regressive
AS	Analytic Signal
ASR	Automatic Speech Recognition
CELP	Code-Excited Linear Prediction
CMU	Carnegie Mellon University
CR	Cauchy-Riemann
DARPA	Defense Advanced Research Projects Agency
DCT	Discrete Cosine Transform
DDK	Diadochokinetic
DFT	Discrete Fourier Transform
EEMD	Ensemble Empirical Mode Decomposition
EM	Expectation Maximization
EMD	Empirical Mode Decomposition
EMS	Envelope Modulation Spectrum
ESA	Energy Separation Algorithm
FM	Frequency Modulation
FS	Fourier Series
FT	Fourier Transform
GMM	Gaussian Mixture Model
GSM	Global System for Mobile communications
HHT	Hilbert-Huang Transform
HSA	Hilbert Spectral Analysis
HT	Hilbert Transform

Symbol

HTK	Hidden Markov Model Toolkit
IA	Instantaneous Amplitude
IF	Instantaneous Frequency
IIR	Infinite Impulse Response
IMF	Intrinsic Mode Function
IPA	International Phonetic Alphabet
ISDN	Integrated Services for Digital Network
ITU	International Telecommunication Union
ITU-T	ITU Telecommunication Standardization Sector
kbps	Kilobits Per Second
LC	Inductor/Capacitor
LPC	Linear Predictive Coding
LSA	Latent Signal Analysis
LTAS	Long-Term Average Spectrum
LTI	Linear Time-Invariant
MELPe	Mixed-Excitation Linear Predictive Enhanced
MFC	Mel-Frequency Cepstrum
MFCC	Mel-Frequency Cepstral Coefficient
MFT	Mean Formant Trajectory
MIT	Massachusetts Institute of Technology
MOS	Mean Opinion Score
PESQ	Perceptual Evaluation of Speech Quality
PM	Phase Modulation
POTS	Plain Old Telephone Service
RMS	Root Mean Square
SA-EMD	Signal-Assisted Empirical Mode Decomposition

Symbol

SHC	Simple Harmonic Component
SLP	Speech Language Pathologist
SNR	Signal-to-Noise-Ratio
SPL	Sound Pressure Level
SR	Speaker Recognition
SRI	Stanford Research Institute
STFT	Short-Time Fourier Transform
STFTM	STFT Magnitude
TEO	Teager Energy Operator
TI	Texas Instruments
VoIP	Voice over Internet Protocol
VSA	Vowel Space Area
WVD	Wigner-Ville Distribution

Chapter 1

INTRODUCTION AND WORK MOTIVATION

1.1 Evaluation of Speech

Speech is arguably one of the most important signals experienced in everyday life. Not only is speech essential for communication, but it also contributes toward our cognitive and social-emotional development. Two of the most important properties of the speech signal are *quality* and *intelligibility*. Although speech quality and speech intelligibility are related, they do not provide the same information. For example, if we consider a speech signal transmitted over a poor communications channel, the presence of noise and signal distortions can affect the quality (in terms of the pleasantness of the sound). However, it is quite possible that the intelligibility (in terms of conveying the relevant information stream) may not be affected. Another example is a noise of sufficient loudness; this type of disturbance may affect both quality and intelligibility.

Dysarthria affects approximately 46 million people worldwide, three million of whom live in the US. Many individuals do not have access to treatment by trained Speech-Language Pathologists (SLPs), leaving them with a persisting inability to communicate. Telemedicine, along with the growing use of mobile devices to augment clinical practice, provides the impetus for the development of remote, mobile applications to augment the work of SLPs. Vowel Space Area (VSA) is an attractive metric for the study of speech production deficits and reductions in intelligibility, in addition to the traditional study of vowel distinctiveness. Because abnormal vowel formant reduction (centralization) is a common feature of speech production deficits,

there has been a longstanding interest in using VSA estimations for characterizing speech motor control, including speech development [9, 10], speech disorders [11–14], and speech interventions [15]. Despite the intuitive appeal of using VSA as an index of speech motor control and intelligibility, its success has been limited and modest [16]. For instance, VSA was minimally predictive of overall intelligibility for individuals with dysarthria, secondary to Parkinsons disease and multiple sclerosis (between 6% and 13%) [17, 18]. More optimistic relationships (over 40%) were reported when examining the same relationship for speakers with dysarthria, secondary to amyotrophic lateral sclerosis (ALS) [19, 20]. The most promising predictive relationship of VSA and intelligibility was demonstrated by Higgins and Hodge [21], in an assessment of a heterogenous sample of children with dysarthria. Attempts to modify the VSA estimate to more sensitively account for differences in the front-back and high-low dimensions have offered some benefit [22]. Such modifications may be preferable for mapping VSA to perceptual measures and speaker classification [11, 23–25]. However, it is likely that more extensive modifications are required to obtain VSA estimates that hold clinical utility for speech production deficits and the resulting decrements in speech intelligibility. Such information, particularly if fully automated and robust to speech sample, would provide an important objective assessment to augment and support clinical practice.

1.2 Mean Formant Trajectories

The use and study of formant frequencies for the description of vowels is commonplace in acoustical phonetics, with uses ranging from quality description, to identification/classification, and perception. However, numerous studies have shown that vowels are more effectively separated when the acoustic parameters are based on spectral information extracted at multiple time points, rather than at a single time

instance. This suggests that spectral dynamics play an integral part in phonetic specification.

We provide an analysis of the mean trajectories of the first two formant frequencies using two popular speech databases. Unlike previous studies of formant trajectories, we analyze speech samples that exhibits a wide range of speakers, dialects, and coarticulation contexts. We illustrate how the formant trajectories vary with gender and, to a lesser extent, with age. Additionally, we provide mean formant trajectories (MFTs) for phoneme groups that are not typically considered. Furthermore, we point out that phonemes which have close $F1$ and $F2$ values at the temporal midpoint, often exhibit formant trajectories progressing in different directions, promoting the importance or formant trajectory progression. Finally, we briefly consider three-dimensional MFTs.

1.3 Hilbert Spectral Analysis

Feature selection and dimensionality reduction are of particular interest during analysis, especially when dealing with big data. The ability to reduce a 10 thousand sample signal to a small number of features, for example 2 to 10, can result in savings on storage space, processing cost, etc. Typically, signal expansions are used for feature extraction. For example, a Fourier series may extract a few number of frequencies that make up 95% of the signal energy. However, these expansions also most always assume linear independence between basis functions in order to preserve desirable signal properties (e.g., energy). It is common practice to choose a signal expansion by making some assumption on the form of the signal. These assumption may not always match the physical nature of the data; in that case, we usually get an infinite number of components, which defeats the purpose of feature extraction and dimensionality reduction. Furthermore, when linear independence of the basis function in a signal expansion is relaxed, a unique expansion is given up, but the potential for huge

reduction in data dimension is possible.

Different tools have been proposed in the literature for analyzing speech and its properties. However, the short-time speech spectrum is the de facto analysis tool used in nearly all areas of speech analysis and applications [7, 26–28]. The spectrogram is a visualization of the energy structure of a signal in the coordinates of time and frequency obtained from the Short-Time Fourier Transform (STFT) [29]. The spectrogram can display a great deal of information about the properties of the speech utterance, including fundamental and formant frequencies [30].

Furthermore, time-frequency analysis is central to practically all areas of signal processing [31]. However, the concept of "instantaneous frequency" is often controversial [32–34]. Instantaneous frequency is that provided by Gabor [35], but other attempts to define instantaneous frequency, particularly those based on the Wigner-Ville have been proposed [31, 34]. Nonetheless, all such definitions suffer the same defect. That is, that the definition applies to one and only one frequency value of the signal at the desired time instant, and this is certainly not the case, for example, for signals that allow for more than one component [31].

On the other hand, Amplitude Modulation–Frequency Modulation (AM–FM) models that allow for a signal representation using multiple instantaneous frequencies have arisen [36–46, 46–51]. These AM–FM models rely on a rigid narrowband component, which limits their utility. Analysis using wideband components in the general form, has not been considered.

As our starting point, we abandon Gabor's complex extension and re-evaluate fundamental principles of time-frequency analysis. We provide a multicomponent model of a signal that enables rigorous definition of instantaneous frequency on a per-component basis. Within our framework, we have shifted all uncertainty of the latent signal to its imaginary part. In this approach, uncertainty is not a funda-

mental limitation of analysis, but rather a manifestation of the limited view of the observer. With the appropriate assumptions made on the signal model, the instantaneous amplitude and instantaneous frequency can be obtained exactly, hence an exact representation of a signal in the coordinates of time and frequency can be achieved. However, uncertainty now arises in obtaining the correct assumptions, i.e., how to correctly choose the imaginary part of the signal components.

1.4 Contributions in the Evaluation of Speech

As traditional VSA estimates are not sufficiently sensitive to map to production deficits, we propose an automated algorithm for evaluating VSA using healthy, connected speech rather than single syllables to estimate the entire vowel working space rather using only corner vowels. Our analyses reveal a strong correlation between the traditional VSA and automated estimates. Further, when the proposed and traditional methods diverge, the automated method is conjectured to provide a more accurate area since it accounts for all vowels.

We propose an application for recording speech samples and providing a variety of derived calculations, novel and traditional, to assess speech production. This includes an individual's pathology fingerprint and identification of which parameters of the intelligibility disorder, such as rhythm, are most disrupted. The automation of this assessment allows SLPs to treat patients remotely, thus permitting for the widespread, worldwide impact of highly skilled assessment, something currently lacking in underdeveloped parts of the world. The individualized selection of desired information for incorporation into a report template will be available. The reports are designed to mimic those generated manually by SLPs today. The published papers related to the contributed work, on the evaluation of speech, are the following [52–57]. The paper submitted for publication, related to the study of MFTs, is the

following [58].

1.5 Contributions in Hilbert Spectral Analysis

We present the Latent Signal Analysis (LSA) problem as a recasting of the classic "complex extension problem". Almost universally, the solution approach has been to use the Hilbert Transform (HT) to construct Gabor's Analytic Signal (AS). This approach relies on Harmonic Correspondence (HC), which may lead to incorrect Instantaneous Amplitude (IA) and Instantaneous Frequency (IF) parameters. We show that by relaxing HC, the resulting complex extension can still be an analytic function and we can arrive at alternate IA/IF parameterizations which may be more accurate at describing the latent signal. Although the existence of other IA/IF parameterizations is not new, Vakman in [59–63] argued that the AS is the only physically-justifiable complex extension. We argue that by modifying the differential equation for simple harmonic motion [64, 65], our parameterizations are also physically justified.

We present HSA as a generalized LSA problem. In the general problem, we seek a representation of a complex time-domain signal consisting of a superposition of latent components, i.e., a multicomponent model. Furthermore, rather than seeking a single IA/IF pair for the signal, we seek a set of IA/IF pairs each associated with the components. Although time-frequency analysis has been extensively studied [3, 66–69], the use of a generalized AM–FM model for this analysis, without the HC condition, has never been proposed. As we prove, using this model leads to non-unique signal decompositions. However, by imposing assumptions on the form of the AM–FM component, a unique parameterization in terms of IA and IF can be obtained. This analysis requires abandoning Gabor's complex extension and instead allowing the assumptions to imply the complex extension. This model enables us to analyze signals with very few restrictions resulting in alternate and possibly more

useful decompositions, especially for non-stationary signals.

Numerical algorithms are given for estimating the IAs and IFs of the components in the AM–FM model where the component is assumed to be an Intrinsic Mode Function (IMF). These algorithms first decompose the signal into IMFs using an improved version of Huang's original Empirical Mode Decomposition (EMD) algorithm [70] and second, demodulate the IMFs to obtain the instantaneous parameters. Unlike previous studies, we closely consider the assumptions made in the definition of the IMF which are carried forward to the demodulation step, thereby avoiding any ambiguity associated with obtaining the instantaneous parameters. We begin with a comprehensive review of EMD, and several variations, and propose an algorithm for the computation of the Hilbert spectrum assuming IMF components. As we demonstrate, while IMFs can be considered latent AM–FM components, there are other classes of AM–FM components that are not IMFs. Examples using the proposed algorithm are provided that highlight alternative decompositions compared to traditional Fourier analysis and demonstrate the advantages of using the HSA framework. The paper submitted for publication, related to the contributed work on Hilbert spectral analysis, is the following [71]. The published paper related to the contributed work, on the evaluation of speech, is the following [72].

1.6 Report Organization

This report is organized as follows. Chapters 2-5 focus on the evaluation of speech quality and speech intelligibility whereas Chapters 6-8 focus on theory related to Hilbert spectral analysis and Chapter 9 focuses on Hilbert spectral analysis of speech. More specifically, in Chapter 2 we discuss the traditional methods implored in the evaluation of speech quality and and speech intelligibility. In Chapter 3, we propose an automated method to assess the VSA of a speaker and discuss the potential

applications in the evaluation of speech intelligibility. In Chapter 4, we propose an augmentative tool for practice in speech-language pathology. In Chapter 5, we compute and investigate MFTs of health speech. In Chapter 6, we propose LSA as a recasting of the classic complex extension problem. In Chapter 7, we propose HSA as a generalized LSA problem and also as a generalization of Fourier Analysis. In Chapter 8, we present numerical algorithms based on a modified EMD for computing the Hilbert spectrum assuming IMFs. We compute and analyze the Hilbert spectrum of speech where we assuming IMFs. In Chapter 9, we provide some conclusive remarks; we also discuss our future work on the use of HSA theory to the speech analysis problem.

Chapter 2

SPEECH QUALITY AND SPEECH INTELLIGIBILITY

2.1 The Evaluation of Speech Quality

There are many assessment methods that have been proposed to evaluate speech in terms of speech quality and intelligibility. Quality assessment can be performed using subjective listening tests or objective quality measures. Quality is only one of many attributes of the speech signal and is highly subjective in nature and difficult to evaluate reliably. This is partly because individual listeners have different internal standards of what constitutes "good" or "poor" quality, resulting in large variability in rating scores among listeners. Quality measures assess *how* a speaker produces an utterance, including attributes such as "natural," "raspy," "hoarse," "scratchy," and so on. Quality measures can also be affected by noise and artifacts introduced by coding or transmission of the speech signal [4].

Subjective evaluation involves comparisons of original and coded/decoded speech signals by a group of listeners who are asked to rate the quality of speech along a predetermined scale [4]. *Objective evaluation* assesses speech using similarity measures between the original and decoded speech signals. For the objective measure to be valid it needs to correlate well with subjective listening tests [4].

2.1.1 Background on Subjective Quality Measures: Mean Opinion Scores

The most widely used direct method of subjective quality evaluation is the category judgment method [28]. This method allows listeners to rate the quality of a test signal using a five-point numerical scale ranging scores from 1 to 5. A score of 1

(bad) implies that the level of distortion is very annoying and objectionable where as a score of 5 (excellent) implies that the level of distortion is imperceptible; the other three scores are 2 (poor), 3 (fair), and 4 (good) [73]. The measured quality of the test signal is obtained by averaging the scores obtained from all listeners. The average score is called the Mean Opinion Score (MOS) [4].

The MOS test is administered in two phases: training and evaluation (testing). During the training phase, listeners hear a set of reference signals that exemplify the high (excellent), low (bad), and middle judgment categories. This phase is very important in subjective evaluation as it is required to equalize the subjective range of quality ratings of the listeners. During the evaluation phase, subjects listen to the test signal and rate the quality of the signal from 1 to 5. Although subjective quality measures may provide the most reliable method for assessing speech quality, they can be time-consuming and costly as they require access to trained listeners [4]. Thus, automated methods have been investigated with the aim of modeling human behavior and objectively predicting the subjective MOS [5].

2.1.2 Objective Quality Measures

In most objective quality measures, a similarity or distance measure is computed between the original speech signal and the processed (encoded/decoded) signal. The distance measure is then usually mapped to a scale from 1 to 5 for comparison to MOS scores. Ideally, the distance measure should correlate well with subjective tests [4].

Given a clean reference signal, objective measures of speech quality are typically implemented by first segmenting the speech signal into 10-30 ms frames and then computing a distortion measure between the original and processed signals. A single, global measure is then computed by averaging the distortion measures of the frames.

Distortion measures can be performed either in the time domain or in the frequency domain [4]. Next, we discuss three quality objective measures.

2.1.2.1 Segmental Signal-to-Noise Ratio Measures

The segmental Signal-to-Noise Ratio (SNR) is quite possibly the most popular example of an objective quality measure. It can be evaluated in either the time or frequency domains. For this measure to be meaningful, it is important to align the original and processed signals in time (frame-by-frame) and correct any phase errors present. The segmental SNR (SNRseg) is defined as

$$\text{SNRseg} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1}x^2[n]}{\sum_{n=Nm}^{Nm+N-1}(x^2[n]-\hat{x}^2[n])} \tag{2.1}$$

where M is the number of frames, N is the length of the frame, $x[n]$ is the original speech signal, and $\hat{x}[n]$ is the processed speech signal. Note that SNRseg is based on the geometric mean of the SNRs across all frames. During intervals of silence, SNRseg may be negative, and thus it can bias the overall measure. Therefore, silence frames are usually removed prior to the computation of the measure [4].

As it is well known, the human auditory system has a frequency sensitivity which is not uniform, i.e., certain frequencies are perceived louder than others even though they have they same loudness [27]. Figure 2.1 illustrates the perceived equal loudness curves [4]. Because of the non-uniform frequency sensitivity of the human auditory system, the signals are often first processed using perceptual weighting filters before computing SNRseg, resulting in the perceptually-weighted segmental SNR. Table 2.1 provides the perceptual filter parameters for different spectral bands [1, 2]. The perceptually-weighted segmental SNR has been found to yield a high correlation with subjective listening tests [4].

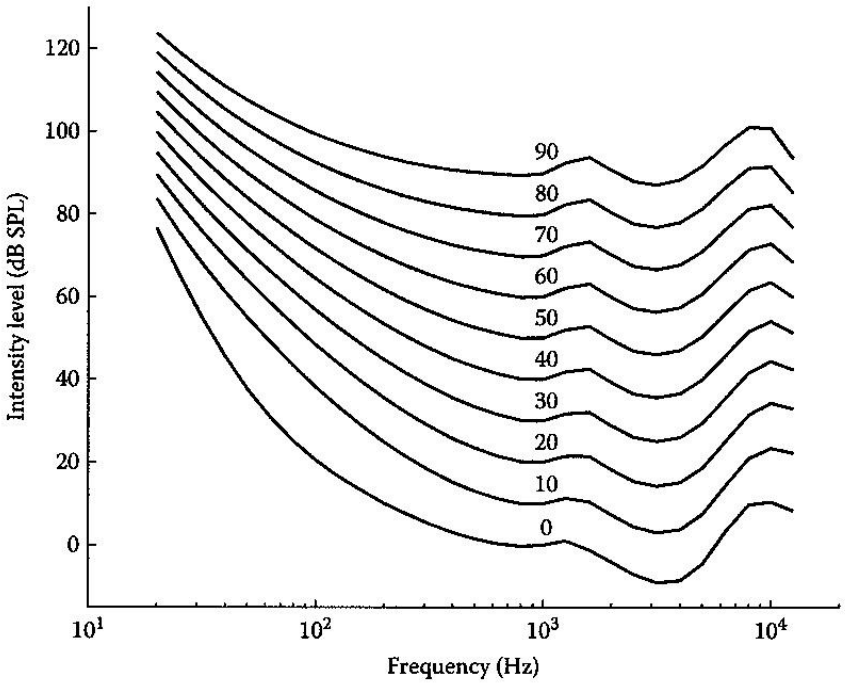


Figure 2.1: Equal loudness curves are a measure of Sound Pressure Level (SPL) in dB, over the frequency spectrum for which a listener perceives a constant loudness when presented with pure steady tones. This figure was taken from [2].

2.1.2.2 Perceptual Evaluation of Speech Quality ITU-T P.862

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications [74]. The ITU Telecommunication Standardization Sector (ITU-T) is responsible for studying technical, operating and tariff questions and issuing recommendations on them with a view to standardizing telecommunications on a worldwide basis.

In 2000, a competition was held by the ITU-T study group to select a new objective measure capable of performing reliably across a variety of codec and network conditions [5]. The Perceptual Evaluation of Speech Quality (PESQ) was selected and standardized in 2001 as the ITU-T recommendation P.862 [75]. PESQ was specifically developed to be applicable to end-to-end voice quality testing under real network conditions, like Voice over Internet Protocol (VoIP), Plain Old Telephone Service

Table 2.1: Perceptual weighting filters: center filter frequencies (Hz), corresponding articulation index weights, and the weights used for computing the weighted spectral distance measure. This table was taken from [1, 2].

Band Number	Center Frequency (Hz)	Weight	Band Number	Center Frequency (Hz)	Weight
1	50	0.003	14	1148	0.032
2	120	0.003	15	1288	0.034
3	190	0.003	16	1442	0.035
4	260	0.007	17	1610	0.037
5	330	0.010	18	1794	0.036
6	400	0.016	19	1993	0.036
7	470	0.016	20	2221	0.033
8	540	0.017	21	2446	0.030
9	617	0.017	22	2701	0.029
10	703	0.022	23	2978	0.027
11	798	0.027	24	3276	0.026
12	904	0.028	25	3597	0.026
13	1020	0.030			

(POTS), Integrated Services for Digital Network (ISDN), Global System for Mobile communications (GSM), etc. [76]. The internal signal representations that are used by the PESQ cognitive model to predict the perceived speech quality are calculated based on psychophysical equivalents of frequency (pitch measured in Barks) and intensity (loudness measured in Sones) [77]. The PESQ algorithm is designed to predict subjective mean opinion scores of an audio sample. PESQ returns a score from 4.5 to 0.5, with higher scores indicating better quality. The PESQ objective metric has been shown to have the highest correlation between objective metrics and subjective signal quality [2]. The block diagram of the algorithm is shown in Figure 2.2 [4]. It is important to note that this algorithm requires a clean reference signal; such algorithms are called full-reference (or double-ended) models and are commonly used for intrusive (or active) measurements [5].

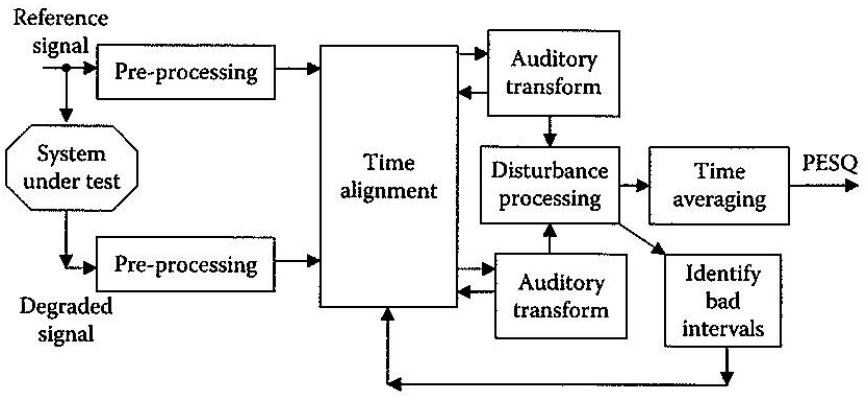


Figure 2.2: Block diagram of PESQ algorithm. This figure was taken from [4].

2.1.2.3 Single-Ended Method for Objective Speech Quality ITU-T P.563

In 2002, the ITU-T held another competition with the aim of standardizing a method that does not need a reference signal for estimating voice quality. In turn, the ITU-T P.563 was selected and standardized in 2004 [5]. Previous methods for speech quality assessment of systems, such as ITU-T Rec. P.862 [75, 76], required a reference signal or only calculated quality indexes based on a restricted set of parameters like level, noise in speech pauses and echoes [5]. ITU-T P.563 is the first recommended method for single-ended non-intrusive measurement applications. It takes into account the full range of distortions occurring in public switched telephone networks and thus is able to predict the speech quality on a perception-based MOS scale according to ITU-T Rec. P.800.1 [73]. The algorithm consists of three stages: preprocessing, distortion estimation, and perceptual mapping. Internally, several parameters, such as the shape of the vocal tract and the naturalness of the vocal quality, are estimated. A block diagram of the P.563 algorithm is shown in Figure 2.3.

2.2 The Evaluation of Speech Intelligibility

Intelligibility is another speech attribute that is different form speech quality. Intelligibility measures the perception of *what* the speaker said. As a result, it is

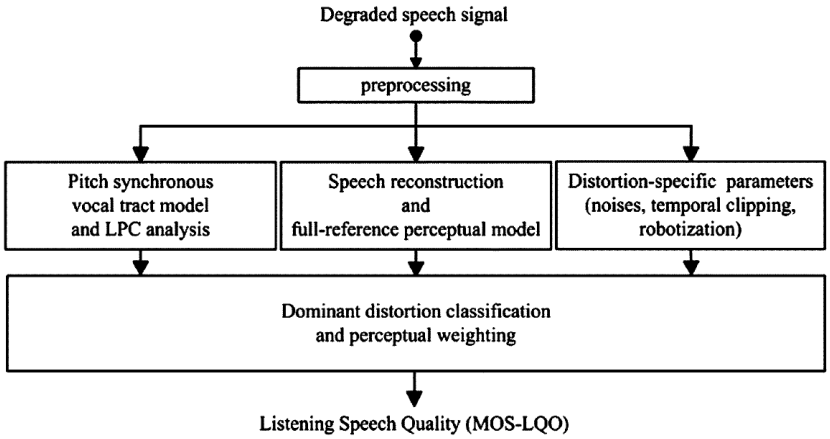


Figure 2.3: Block diagram of ITU-T P.563. This figure was taken from [5].

not subjective and is typically measured by presenting speech material to a group of listeners and asking them to identify the words spoken. Intelligibility is typically quantified by counting the number of words or phonemes correctly identified by a listener [4].

An important problem for individuals with speech disorders is the estimation of speech intelligibility [78–84]. However, this problem is not yet completely understood, even in the case of normal speech. That is, even for normal speech, the exact characteristics of a speech signal that affect intelligibility are unknown.

2.2.1 Standard Assessments of Speech Intelligibility

Intelligibility of patients with speech intelligibility disorders is currently assessed through subjective tests performed by trained speech-language pathologists. Subjective tests, however, tend to be inconsistent, costly and, oftentimes, not repeatable. In fact, research has shown poor inter-rater and intra-rater reliability of intelligibility assessments in clinical studies. Furthermore, clinicians working with patients can form a bias based on their interactions, resulting in intelligibility assessment of limited validity and reliability [85–88]. It is for this reason that the development of

inexpensive, yet subjective, unbiased, repeatable, and automated methods has gained interest. One example of a standard assessment is the Franchay Dysarthria Assessment which is relatively quick to perform and has been found to have acceptable inter-rater reliability [89]. One of the major drawbacks of nearly all assessments of speech intelligibility is the fact that results do not shed light on the nature of the intelligibility degradation. However, information pertaining to the nature of the degradation is important clinically, in order to be able to treat and monitor it.

2.2.2 Automated Assessments of Speech Intelligibility

There have been few research studies in the area of automated assessment of speech intelligibility in the literature. In [78], rhythm metrics are estimated through envelope modulation spectra to classify between different dysarthria types. In [90], acoustic cues are used to detect Parkinson's disease using only speech. In [79–81], an algorithm is developed for assessing intelligibility using a regression scheme that makes use of a number of acoustic cues. In [81–84], several schemes are presented to assess speech quality and intelligibility by comparing to a clean reference signal. As discussed next, the most common methods for automatic estimation of speech intelligibility either quantify intelligibility by counting the number of words or phonemes correctly identified, or they attempt to map a set of features to an intelligibility score.

2.3 Intelligibility Estimation via Automatic Speech Recognition Systems

By far the most common and straightforward approach for the estimation of speech intelligibility is the use of existing Automatic Speech Recognition (ASR) systems. Utilizing ASR systems, intelligibility is typically quantified by counting the number of words or phonemes correctly identified [4]. There exists several freely available high quality open source ASR systems online such as the Carnegie Mellon University

(CMU) Sphinx [91], Hidden Markov model Toolkit (HTK) [92], and Kaldi [93]. However, most modern ASR systems operate using similar feature sets, most notably, the Mel-frequency cepstral coefficients. A major limitation of the ASR based approach is the fact that, for areas of high interest, such as the intelligibility of dysarthric speakers, ASR systems do not typically perform well due to the highly varied and atypical speech waveforms generated by these speakers. Cepstrum-based features and their computation are presented next.

2.3.1 Cepstral Analysis for Speech Characterization

Often, a speech signal is assumed to be the output of a Linear Time-Invariant (LTI) system as it can be obtained from the convolution of an input (the vocal chords) and the system's impulse response (the vocal tract). Using the LTI characterization for speech signals, it is important to identify the LTI parameters that can be used as speech features. One such important parameter is the (power) cepstrum [94]. The computation of the cepstrum is a homomorphic transformation; that is, the convolution of two signals becomes equivalent to the sum of their cepstra. When the cepstrum of a speech signal is computed, the resulting deconvolution causes the lower order coefficients to represent the filter shape and the higher order coefficients to represent the filter excitation. It is for this reason that cepstral analysis is so popular for speech processing algorithms. Although the cepstrum is very useful in speech processing due to its homomorphic property, it has large dimensionality. This dimensionality is the same as the the length of the Fourier Transform (FT) used in the cepstrum computation and it usually becomes a problem in speech classification. As an alternative to the cepstrum, the Mel-Frequency Cepstrum (MFC) provides a low-dimensional representation of the key acoustic aspects of speech [95]. Since MFC efficiently represents speech features, it has been used in ASR systems for

nearly 30 years [95] and has been used for over 20 years in speaker recognition tasks where speaker-dependent statistical models (such as Gaussian Mixture Models) are employed for speaker identification [96].

The low-dimensional representation provided by the MFC can be used as features of speech necessary for machine recognition. The MFC has frequency bands that are spaced on the Mel scale rather than the linear scale of the cepstrum. The Mel scale, proposed by Stevens, Volkman and Newman in 1937 [97], is a perceptual scale of pitches judged by listeners to be similar to one another. The MFC is composed of several Mel-Frequency Cepstral Coefficients (MFCCs), one for each of the Mel scale frequency bands. Encapsulated in the MFCCs is information related to the vocal tract configuration (formant frequencies) and glottal pulse (pitch) which together determine the "acoustics of speech" [28].

2.3.2 Cepstrum Computation

In order to compute the cepstrum, the sampled speech signal, $s[m]$, is first windowed to obtain

$$x_r[m] \;=\; s[rR+m]w[m] \tag{2.2}$$

where $w[m], m = 0, \ldots, L-1$ is a window of length L, R is the window or frame advance in samples, and r denotes the frame index. In vector form, the $(L \times 1)$ speech frame is represented as

$$\mathbf{x} \;=\; [x_r[0]\ x_r[1]\ \ldots\ x_r[L-1]]^T \tag{2.3}$$

where T denotes vector transpose; note that we drop the subscript r to simplity notation. The spectrum is obtained by taking the Discrete Fourier Transform (DFT)

of $x[m]$ as

$$X[k] \;=\; \sum_{m=0}^{L-1} x[m] \exp(-2\pi jkm/L). \tag{2.4}$$

In vector form, we use the operator $\mathcal{F}$ to denore the DFT samples as

$$\mathbf{X} = \mathcal{F}\{\mathbf{x}\} = [X[0]\ X[1]\ \ldots\ X[L-1]]^T. \tag{2.5}$$

The cepstrum of $\mathbf{x}$ is defined as [98]

$$\mathcal{C} \;\equiv\; \mathcal{F}^{-1}\left\{\log\left|\mathbf{X}\right|\right\} \tag{2.6}$$

where the inverse DFT operator $\mathcal{F}^{-1}$ is applied to the log-magnitude spectrum of $\mathbf{x}$. Here $|X|$ is the absolute value of the elements of X.

2.3.3 Mel-Frequency Cepstrum Computation

The Mel-weighted power spectrum in (2.8) can be expressed in matrix form as

$$\mathbf{Y} = \mathbf{\Phi}\left|\mathbf{X}\right|^2 \tag{2.7}$$

where $\mathbf{Y}$ is $J \times 1$, the weighting matrix $\mathbf{\Phi}$ is $J \times L$ and has columns ϕ_j, $j = 1, \ldots, J$, and $\left|\mathbf{X}\right|^2$ is $L \times 1$.

We can compute the MFCC vectors $\mathcal{M}$ by applying the Discrete Cosine Transform (DCT) operator to the log operation of the Mel-weighted power spectrum [95]

$$\mathcal{M} \;=\; \mathrm{DCT}\left\{\log \mathbf{\Phi}\left|\mathbf{X}\right|^2\right\}. \tag{2.8}$$

The Mel-weighting matrix $\mathbf{\Phi}$ is based on the human perception of pitch [97], and its columns ϕ_j are in the form of a bank of filters, each with a triangular frequency response [95]. Assuming a sampling frequency of 8 kHz and $J = J_1 + J_2$ the Mel-scale weighting filters are generally derived from J_1 triangular weighting filters

that are linearly-spaced between 0 and 1 kHz, and J_2 triangular weighting functions logarithmically-spaced between 1 and 4 kHz [95].

Additionally in [98, 99], two "half-triangle" weighting filters centered at 0 kHz and 4 kHz are tallied in J_1 and J_2, respectively, since these directly affect the number of MFCCs. The use of the two "half-triangle" weighting filters improves the quality of reconstructed speech obtained using MFCC inversion. In usual implementations, the lenght of J os the DCT in (2.8) is less than the frame length L in (2.3) Thus this weighting may also be thought of as a perceptually-motivated dimensionality reduction [98]. An example of a Mel-filter bank is shown in Figure 2.4.

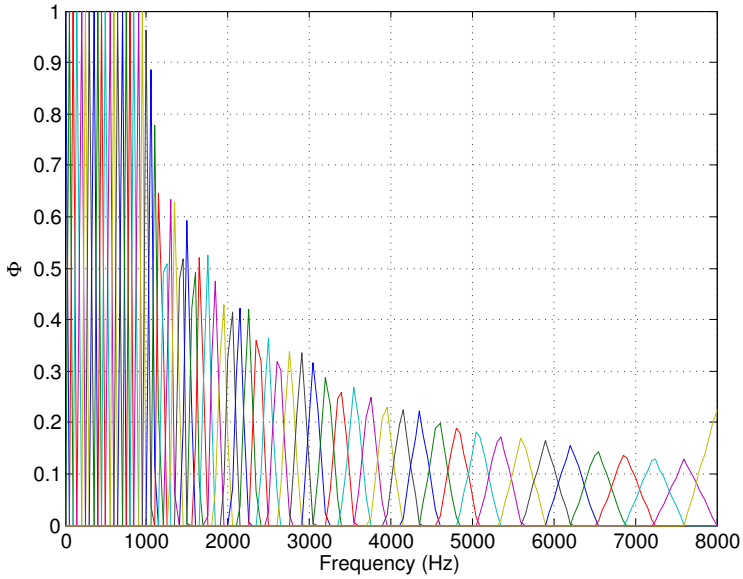


Figure 2.4: A high resolution Mel-filter bank including two "half-triangle" weighting functions centered at 0 Hz and the Nyquist frequency that are necessary for good quality MFCC inversion. A "half-triangle" filter is clearly illustrated with maximum value at the Nyquist frequency.

2.4 Intelligibility Estimation via Feature Mapping

Another approach that can be used for the estimation of speech intelligibility is the mapping of acoustic features to perceptual correlates using machine learning algorithms. There are various advantages to this approach. Specifically it avoids both

biased human interactions and ASR systems. Also, if a certain subset of features can be linked to a perceptual description of speech (such as vowel articulation, prosody, hoarseness, and hypernasality), then mappings could also be created to estimate these perceptual qualities. This additional information could then be used in a clinical setting to influence treatment or monitor the progression of a patient. However, there are two problems associated with this method: 1) selecting appropriate features, and 2) mapping the features to an intelligibility estimate. Next, we discuss several features commonly used in the literature to obtain intelligibility estimates.

2.4.1 Envelope Modulation Spectrum Feature

The Envelope Modulation Spectrum (EMS) [78] is a representation of the slow amplitude modulations in a signal and the distribution of energy in the amplitude fluctuations across designated frequencies, collapsed over time. It has been shown to be a useful indicator of atypical rhythm patterns in pathological speech. The speech segment, $x(t)$, is first filtered into 7 octave bands with center frequencies of 125, 250, 500, 1,000, 2,000, 4,000, and 8,000 Hz. Let $h_i(t)$ denote the filter associated with the ith octave. The filtered signal $x_i(t)$ is obtained from the ith filter $h_i(t)$, $i = 1, \ldots, 7$, using

$$x_i(t) = h_i(t) * x(t) \tag{2.9}$$

Where $*$ denotes convolution. The envelope in the ith octave, denoted by $e_i(t)$, is extracted by:

$$e_i(t) = h_{\text{LPF}}(t) * \mathcal{H}\{x_i(t)\} \tag{2.10}$$

where $\mathcal{H}\{\cdot\}$ is the Hilbert transform and $h_{\text{LPF}}(t)$ is the impulse response of a 20 Hz lowpass filter. Once the amplitude envelope of the signal is obtained, the low-frequency variation in the amplitude levels of the signal can be examined. Fourier analysis is used to quantify the temporal regularities of the signal. With this, six

EMS metrics are computed from the resulting envelope spectrum for each of the seven octave bands, $x_i(t)$, and the full signal, $x(t)$:

1. Peak frequency
2. Peak amplitude
3. Energy in the spectrum from 3-6 Hz
4. Energy in the spectrum from 0-4 Hz
5. Energy in the spectrum from 4-10 Hz
6. Energy ratio between the 0-4 Hz band and the 4-10 Hz band.

2.4.2 Long-Term Average Spectrum Feature

The Long-Term Average Spectrum (LTAS) [100] captures atypical average spectral information in the signal. Nasality, breathiness, and atypical loudness variation, are common causes of intelligibility deficits in pathological speech, and present themselves as atypical distributions of energy across the spectrum; LTAS attempts to measure these cues in each octave. For each of the 7 octave bands, $x_i(t)$, and the full signal, $x(t)$, the following features are extracted:

1. Average normalized Root Meamn Square (RMS) energy
2. RMS energy standard deviation
3. RMS energy range
4. Pairwise variability of RMS energy between ensuing 20 ms frames.

2.4.3 Internal ITU-T P.563 Features

ITU-T P.563 was previously described in Section 2.1.2.3. While previously used for evaluating speech quality, it is possible to remap the internal features of the algorithm to estimate speech intelligibility, rather than speech quality. In total, there are five major classes of internal features deemed appropriate for the study of speech intelligibility:

1. f_{basic} - basic speech descriptors, such as pitch and loudness information

2. f_{VT} - vocal tract analysis, including statistics derived from estimates of vocal tract area based on the cascaded tube model

3. f_{stat} - speech statistics, including the skewness and kurtosis of the cepstral and linear prediction coefficients, which model the source and filter responsible for production of the signal

4. f_{SSNR} - static SNR, measurements of SNR, estimates of background noise, and estimates of spectral clarity based on a harmonic-to-noise ratio

5. f_{SegSNR} - segmental SNR, or dynamic noise, where the SNR is calculated on a frame-by-frame basis.

The subjective rating (MOS score), which is a non-linear combination of the above features can also be used. Other internal P.563 features such as "Interruptions and Mutes," which are distortions, such as temporal speech clipping or speech interruption, that occur as a byproduct of signal transmission in telecommunications may not be relevant or deemed of interest if the goal is to evaluate a speaker's intelligibility and not the effects of signal transmission. A detailed description of all these internal features, including mathematical derivations, can be found in [5].

2.4.4 Linear Predictive Coding Features

Internal Linear Predictive Coding (LPC) features are utilized in the ITU-T P.563 algorithm. LPC is one of the most commonly used methods for encoding good quality speech at low bit rates [28]. The basics of LPC are discussed here; however, it should be noted that there are several advanced variations of the basic LPC algorithm, including the Code-Excited Linear Prediction (CELP) [101] and the Mixed-Excitation Linear Predictive enhanced (MELPe) algorithm [102–105].

LPC is a model based speech coding process that assumes that the elements of human speech production, the vocal cord and the vocal tract, can be modeled by an excitation source and a tube [27, 28]. The parameters of the model relating to the vocal cords consist of both intensity and pitch, while the parameters of the model related to the vocal tract are characterized by its resonances, or formant frequencies, and are modeled by an all-pole filter. The basic problem of an LPC system is to estimate the parameters of an Auto Regressive (AR) process. The formants are determined using a difference equation which represents each sample as a linear combination of the previous samples (AR processes). The equations for computing the least squares solution for the LPC coefficients are solved by using the Levinson-Durbin algorithm [28]. Once the parameters for the model have been calculated, they can be encoded, transmitted, and decoded. There are three steps to the decoder operation. In Step 1, an excitation signal is generated. In Step 2, the excitation signal is filtered using an all-pole Infinite Impulse Response (IIR) filter defined by the parameters of an all-pole model constructed using the formants. In Step 3, the filter outputs are overlapped (as defined in the encoder) and added to synthesize the speech signal. Since each frame corresponds to a speech frame, the length of the excitation signal must be the same as the frame [106].

2.4.4.1 The CELP Algorithm

The CELP class of algorithms has been proven to work reliably as well as provide good scalability. Some examples of CELP-based standard codecs consist of G.728 [107], which operates at 16 kilobits per second (kbps) and DoD CELP (Federal Standard 1016) [108], which operates at 4.8 kbps. The open-source SPEEX codec, also based on CELP, operates at a variety of bit rates ranging from 2150 bps to 44 kbps. There are several aspects of human speech that cannot be modeled by traditional linear prediction coefficients (LPCs). This results in a difference between the original and reconstructed speech. CELP-based codecs first compute the LPCs and then calculate the residue using traditional LPC. This residue is compared to a codebook and the codeword which best represents the residue is transmitted. Additionally in the CELP implementation, an adaptive codebook is utilized to further encode the residue in order to more accurately synthesize speech.

2.4.4.2 The MELPe Algorithm

The MELPe algorithm was derived using several enhancements to the original mixed-excitation linear predictive algorithm [102]. MELPe is also known as MIL-STD-3005 [103] and NATO STANAG-4591 [104] and supports bit rates of 1200 bps and 2400 bps. There also exists a proprietary 600 bps MELPe vocoder algorithm [105]. Traditional LPC algorithms use either periodic pulse trains or white noise as excitation for a synthesis filter. The MELPe family of vocoders use a mixed-excitation model of the human voice and extensive lookup tables to extract and regenerate speech. The MELPe codec also utilizes aperiodic pulse excitation, pulse dispersion to soften the synthetic sound of reconstructed speech, and adaptive spectral filtering to model the poles of the vocal tract.

Chapter 3

AUTOMATIC ASSESSMENT OF VOWEL SPACE AREA

3.1 Motivation of the Proposed Algorithm for Vowel Space Area Automatic Assessment Algorithm

The formants of a vowel are the frequencies at which a vowel resonates, and each vowel has its own signature formants. The Vowel Space Area (VSA) described in [109] as the two-dimensional area bounded by the lines connecting first ($F1$) and second ($F2$) formant frequency coordinates of corner vowels. Estimating the VSA has been shown to be important for studying vowel identity, speaker characteristics, speech development, speaking style and sociolinguistic factors that influence vowel production [7, 26, 110–116]. The traditional steps for computing the VSA are shown in Figure 3.1(a). A typical computation involves making static measurements of the $F1$ and $F2$ values for each of the four corner vowels (or 3 point vowels, /a, i, u/ for triangle) at 50% vowel duration, for several productions of each vowel [117]. The mean $F1$ and $F2$ values for each of the four corner vowels are then used to compute the area of the quadrilateral formed by the corner vowels. Since the frequencies of the first and second formants roughly relate to the size and shape of the cavities created by the jaw opening ($F1$) and tongue position ($F2$), the VSA is an acoustic proxy for the kinematic displacements of the articulators [118]. In general, studies have shown speech that is clearer and more intelligible is associated with larger values of VSA rather than smaller values of VSA [119]. This is interpreted as corresponding to greater articulatory excursions and more distinct acoustic-articulatory vowel targets. Thus, the VSA and other derived vowel metrics related to distinctiveness have been

quite successful in the study of speaking style, dialects and languages [113–115].

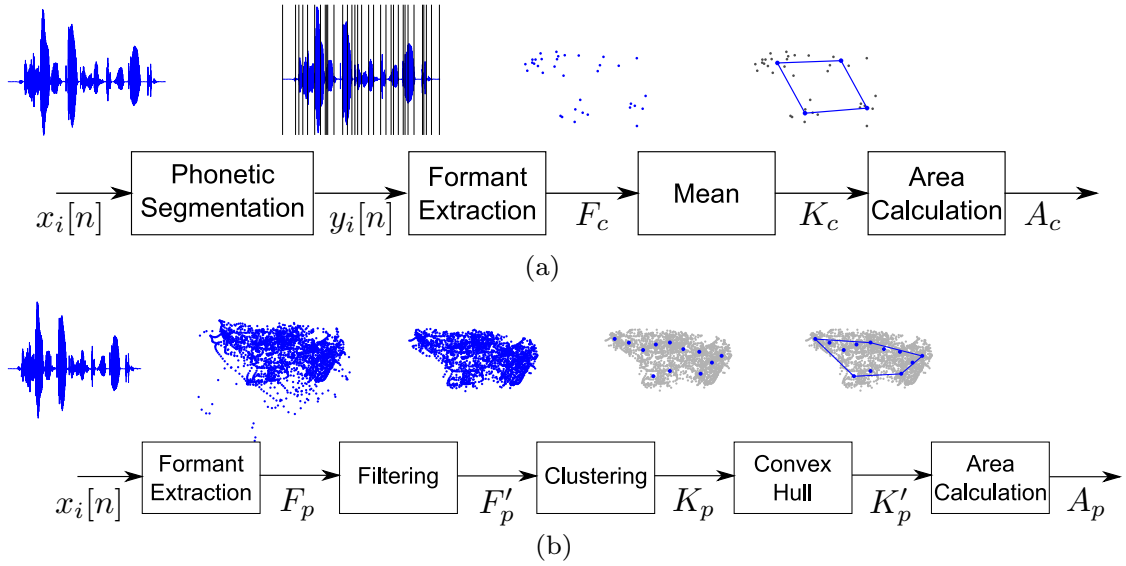


Figure 3.1: Block diagram for (a) the typical steps used in computing the VSA: speech samples are phonetically segmented, formants for the corner vowels are estimated, the mean value of each corner vowel is obtained, and the area bounded by the mean of the corner vowels is computed; (b) the proposed automatic assessment VSA method.

There are several significant limitations associated with existing VSA computation approaches in the context of speech production disorders. The first limitation is that VSA calculations are based only on point (triangle) or corner (quadrilateral) vowels, rather than all vowels. The second limitation is that these vowels are produced in isolation (typically hVd). Note that the original VSA computation method was borrowed from the study of vowel production in healthy speech to examine vowel distinctiveness [26]. Following this direction seems reasonable from the standpoint of trying to identify the most disparate regions of the vowel space (and, by extension, the maximal articulatory excursions) in a way that is free of extraneous coarticulatory influences.

However, when applying the original VSA computation methods to disordered speech, they failed to robustly capture speech production deficits and intelligibility in

a clinically meaningful way. This suggests that more sensitive methods are necessary, when the VSA is globally reduced, as in speech production disorders. One possible approach is to sample the entire articulatory working space and characterize its shape in order to fully account for the extent of articulatory displacements and their acoustic consequences. A second possible approach is to extract vowel formant information from productions in connected speech rather than single word productions to magnify the impact of the underlying production disorder. Finally, perhaps the most important limitation from an applied standpoint, is that the traditional VSA estimation process is cumbersome, requiring phonetic segmentation of input speech.

In an effort to overcome these limitations and move closer to a clinical tool, we propose a novel VSA computation approach that has the following advantages over the original methods: a) it is fully automated; b) it can be applied to any speech segment and to an type of speech that contains a range of vowels; c) it makes use of all vowel produced rather than using only three of four value to estimate the triangle or quadrilateral VSA shape. The proposed Automatic Assessment of Vowel Space Area (AAoVSA) algorithm relies on a series of automated tools for extracting all formants from voiced sections of speech, thereby removing the need for hand segmentation. This is followed by a clustering and area calculation algorithm based on the convex hull of the cluster centers to estimate the final VSA. The proposed algorithm is applied to healthy speech and then compared to the original quadrilateral VSA method by hand-segmenting the same speech samples [120]. Results show that the automated estimate exhibits a strong correlation with the hand-segmented estimate and often yields a more accurate estimate of the VSA. This proposed work was published in [52], and is described in detail in the remaining of this chapter.

3.2 Automatic Assessment Method

Figure 3.1(b) shows a block diagram of the proposed AAoVSA method for the automated estimation of the VSA. The algorithm can operate on any incoming speech signal that contains a range of vowels. The ith utterance of speech, $x_i(t)$, is analyzed on a frame-by-frame basis and, for each voiced frame, the first and second formants are estimated. Following, outliers are removed and the remaining points are clustered. The convex hull of the cluster centers is determined and the area of the resulting convex hull is calculated. In the following sections, these required steps are discussed in detail.

3.2.1 Formant Extraction

A Praat [121] script is used to automatically extract all pairs of $(F1, F2)$ that correspond to voiced frames. The Praat script assesses voicing on a frame-by-frame basis by estimating periodicity using an autocorrelation-based method. In this work we only considered the first two formants; however, using the recommended Praat values, five formants were extracted per frame below a ceiling value. This value was 5 kHz for male speakers and 5.5 kHz for female speakers. We also used the following settings: 1 ms frame advance; 50 ms analysis window; pre-emphasis starting from 50 Hz. Internally, Pratt computes estimates of the formants by resampling to twice the ceiling of the formant search range, then applying a pre-emphasis filter, windowing the speech in the time domain using a Gaussian window, and then estimating the Linear Predictive Coding (LPC) coefficients using the algorithm by Burg [122, 123]. Processing all input speech results in an $N \times 2$ matrix, $\mathbf{F}_p$, that stores all $(F1, F2)$ pairs for a particular speaker, where N is the number of formant observations for a particular speaker.

3.2.2 Filtering

Automated formant estimation algorithms can result in outliers. In order to identify the extrema, the probability distribution of each speaker's formants, is modeled using a Gaussian Mixture Model (GMM) and low-likelihood points are identified and removed. The use of GMMs is common in speech processing applications [96]. The weight, mean, and (full) covariance matrix for each of the 4 component densities in the Gaussian mixture are learned using the Expectation Maximization (EM) algorithm[124]. For each formant in $\mathbf{F}_p$, the log-likelihood is calculated and components with a likelihood less than $0.3\overline{L(\mathbf{F}_p)}$ are identified as outliers and are removed from downstream processing, where $\overline{L(\mathbf{F}_p)}$ denotes the mean likelihood of all observations in $\mathbf{F}_p$. The matrix formed using the filtered parametere set is denoted by $\mathbf{F}'_p$. Our simulations showed that the outlier filtering rejected approximately 15% of the total number of formant observations for a particular speaker.

3.2.3 Clustering

Following outlier rejection, the remaining formants in $\mathbf{F}'_p$ are clustered using the k-means algorithm [125]. Twelve cluster centers (with each cluster center corresponding to each of the 12 English vowels) are initialized using the mean $(F1, F2)$ values as reported by Hillenbrand [7] at 50% vowel duration. The cluster centers are initialized for adult males and females, using the respective reported values, and the returned values are used to form the matrix $\mathbf{K}_p$.

3.2.4 Convex Hull / Area Calculation

Using the implementation of the Quick-hull [126] algorithm in MATLAB [127], the convex hull of the set of points in the matrix $\mathbf{K}_p$ is found. The clockwise ordered

endpoints (beginning and ending with the same point) of the resulting convex polygon are stored in a matrix $\mathbf{K}'_p$. The area of the convex polygon with m corners is then given, with slight abuse of the determinant notation, by

$$A_p = \frac{1}{2}\,|\mathbf{K}'_p| \;= \frac{1}{2}\begin{vmatrix} F1_1 & F2_1 \\ F1_2 & F2_2 \\ \vdots & \vdots \\ F1_m & F2_m \\ F1_1 & F2_1 \end{vmatrix} = \frac{1}{2}\sum_{i=1}^{m}(F1_i F2_{i+1} - F2_i F1_{i+1}) \tag{3.1}$$

3.2.5 Stimuli

Speech samples were drawn from the TIMIT [120] corpus commissioned by (Defense Advanced Research Projects Agency) DARPA. The TIMIT corpus consists of 6,300 sentences, 10 sentences spoken by 630 speakers from 1 of 8 major dialect regions [128] of the United States. The TIMIT corpus includes hand verified, time-aligned orthographic, phonetic and word transcriptions as well as 16-bit, 16 kHz speech waveform files for each utterance. The corpus design was a joint effort among the Massachusetts Institute of Technology (MIT), Stanford Research Institute International (SRI) and Texas Instruments, Inc. (TI). The speech samples consists of phonetically-diverse sentences intended to expose dialectal variants of the speech.

3.3 Results and Discussion

We simulated our proposed AAoVSA algorithm and compared it to the original quadrilateral VSA using hand-segmented speech, including several derivations of the Pearson correlation coefficient [129].

3.3.1 *Performance Analysis*

In order to assess the performance of the proposed method, several comparisons were made between the proposed method and a control method. The control method uses the original VSA computation paradigm by utilizing the meta-data provided with the TIMIT corpus. More specifically, for each occurrence of the corner vowels, estimates of the means of the formant frequencies were calculated and the area of the resulting quadrilateral was computed. An estimate of the VSA for each of the 630 speakers utilizing all ten sentences per speaker ($x_i[n], i = 1, \ldots, 10$) for both the proposed and control methods were computed. When separated by gender, male and female speakers yielded correlation coefficients of $\rho = 0.790980$ and $\rho = 0.74681$, respectively. The proposed method yielded a correlation coefficient of $\rho = 0.77553$ when computed over all 630 speakers. A scatter plot of the data is shown in Figure 3.2, and the results are summarized in Table 3.1.

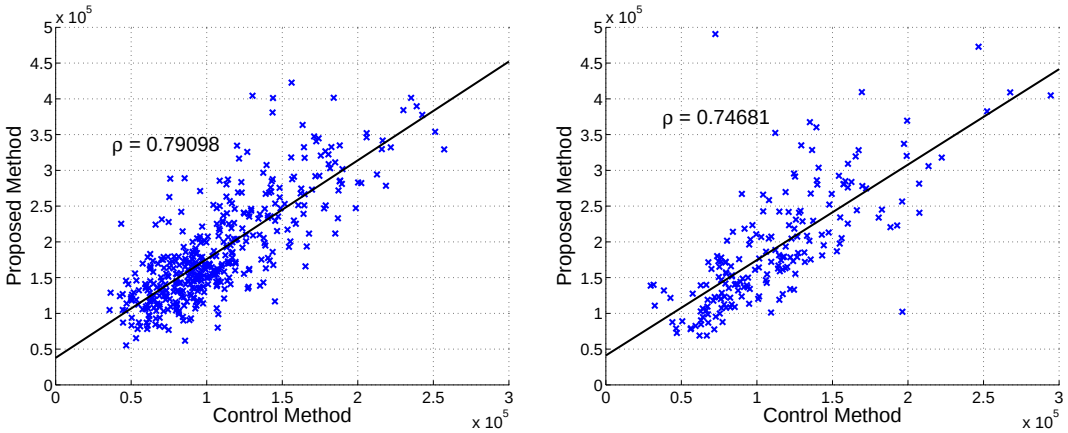


Figure 3.2: A scatter plot showing the estimated VSA obtained using the proposed and control methods for each of the 630 speakers in the TIMIT corpus for (a) male speakers (b) female speakers. Male and female speakers yielded correlation coefficients of $\rho = 0.790980$ and $\rho = 0.74681$, respectively. The proposed method yielded a correlation coefficient of $\rho = 0.77553$ over all speakers.

Similar analyses comparing estimates of the VSA corresponding to an entire di-

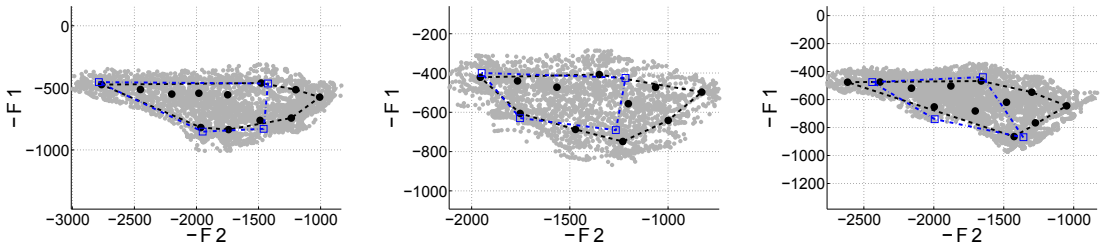
alect region were performed. When estimating the VSA for the 8 dialect regions by gender, estimates yielded correlation coefficients of $\rho = 0.50937$ and $\rho = 0.52836$, for male and female speakers, respectively. The proposed method yielded a correlation coefficient of $\rho = 0.60118$ when estimating the VSA for a dialect region using both male and female speakers. The corresponding results are also summarized in Table 3.1.

Table 3.1: Correlation between the proposed and control methods.

Case	**By Speaker**	**By Dialect Region**
Male	0.79098	0.50937
Female	0.74681	0.52836
All	0.77553	0.60118

Overall, the proposed AAoVSA method has demonstrated high correlation when compared to the original quadrilateral VSA method. However, the proposed method has the potential to yield a more accurate VSA estimation result than the original method. This is because the original method limits the computation of the VSA to only three of four corner vowels from the twelve English vowels. In reality, there are many occurrences of $(F1, F2)$ pairs that occur outside of this space and that contribute to the overall shape of the vowel space. This is readily seen in Figure 3.3, by comparing the VSA as bounded using the proposed and control methods. The proposed metric results in consistently larger VSA estimates, but it also more accurately accounts for the actual shape of the VSA. This may provide a more complete assessment of the contribution of VSA to intelligibility and subsequent decrements.

It is important to note that a key requirement of the proposed algorithm is that the vowel space is adequately sampled. This means that the analyzed content must



(a) (b) (c)

Figure 3.3: The VSA for three speakers as bounded using the proposed (dashed line) and control (dash-dot line) methods overlaid on the filtered points F'_p (small grey dots). The mean corner vowels K_c (large squares) and the cluster centers K_p (large dots) are also shown. The proposed method better accounts for the actual shape of the VSA. The axes have been chosen so that the plots have the same orientation as the standard International Phonetic Alphabet vowel trapezium.

be phonetically balanced or consistent across individuals for comparison. By design, the TIMIT corpus indeed satisfied this requirement. For clinical applications of this work, clinicians will have the option of specifying the spoken text, ensuring that the incoming speech stream is balanced.

Chapter 4

SPEECH ASSIST: AN AUGMENTATIVE TOOL FOR PRACTICE IN SPEECH-LANGUAGE PATHOLOGY

4.1 Motivation Proposed for the Mobile Application Suite: Speech Assist

Abnormal speech production is a common symptom, and often a leading indicator, in a number of neurological disorders, requiring the expertise of a Speech Language Pathologist (SLP) for evaluation and treatment. Telemedicine, utilizing videoconferencing and other software, has been advanced as a partial solution to the longstanding shortage of SLPs, particularly in rural and underserved areas. The advent of affordable, portable, and capable telecommunication devices (e.g., tablets, smart phones) allows for the development of more powerful approaches [55]. Here, we propose a mobile application suite that allows patients to remotely record speech and video samples and automatically provides a report of the integrity of speech production based on a variety of acoustic calculations. This work was published in [55] and [54].

4.2 Speech Assist Methods

4.2.1 Automated Vowel Space Area

In Chapter 3, we proposed a novel method for automatic estimation of Vowel Space Area (VSA). In general, studies have shown that VSA is larger in speech that is clearer and more intelligible than speech associated with smaller VSAs [119]. This can be interpreted as corresponding to greater articulatory excursions and more distinct acoustic-articulatory vowel targets [118]. Thus, the VSA and other derived vowel metrics related to vowel distinctiveness have been quite successful in the study

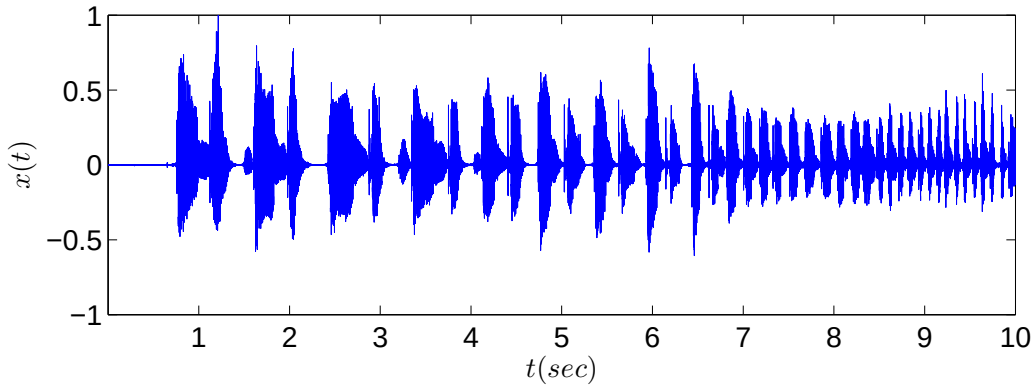
Table 4.1: Statistical summary of the automated DDK method for the example waveform in Figure 4.1(a).

Description	**Value**
Mean Rate	3.63 (Hz)
Standard Deviation	1.46 (Hz)
Minimum Rate	1.33 (Hz)
Maximum Rate	6.00 (Hz)

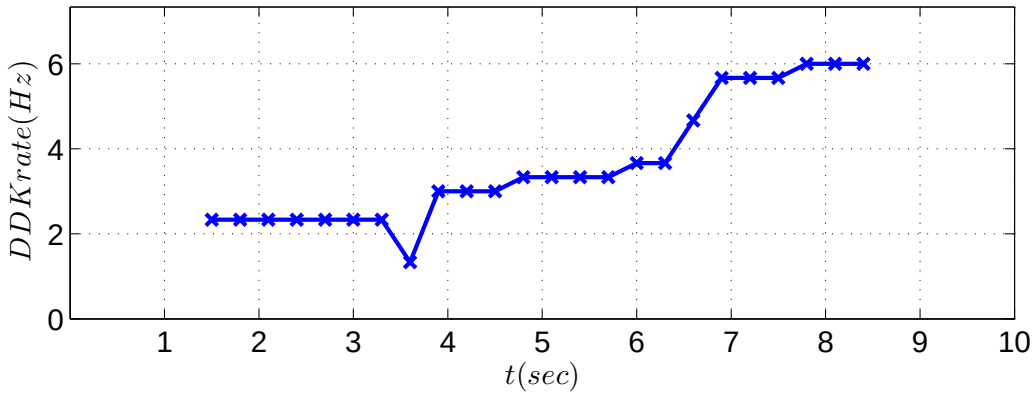
of speaking style, dialects and languages [113–115]. Because abnormal vowel formant reduction (centralization) is a common feature of speech production deficits, there has been a longstanding interest in using VSA estimation for characterizing speech motor control, including speech development [9, 10], speech disorders [11–14], and speech interventions [15].

4.2.2 Automated Diadochokinetic Rate

Diadochokinetic (DDK) rate refers to an assessment used by SLPs, that measures how quickly an individual can accurately produce a series of rapid and alternating movements of the oral articulators. We develop an automated method that provides the average rate of movement, including statistics that describe the consistency with which these movements occur. Figure 4.1(a) shows a speech waveform to be used for evaluating an individual's DDK rate. Figure 4.1(b) shows the average rate of movement over a short time interval. In Table 4.1, we compute statistics that describe the consistency with which these movements occur.



(a)



(b)

Figure 4.1: (a) Waveform of a typical speech utterance used in DDK rate evaluation; (b) The DDK rate of the speech utterance in (a) using the automated DDK method developed.

4.2.3 Automated Perceptual Ratings

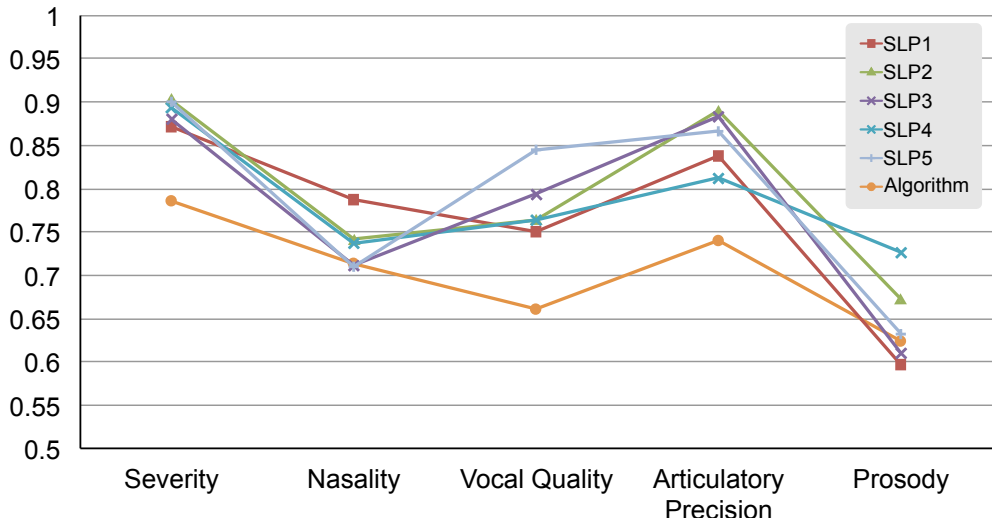
For our proposed mobile application suite, several other items require further development. First, automated SLP perception models to represent ratings of perceptual dimensions (e.g., nasality, prosody, articulatory precision, etc.) made by expert SLPs. Secondly, the use of these models towards developing mappings to acoustic features (e.g., shimmer, jitter, envelope spectra), utilizing machine and ensemble learning [53, 56, 130]. In our current work, we have identified acoustic features that correlate strongly with the ratings of pathological speech by SLPs; modeling SLP's

responses based on these features, we have shown success in automatically evaluating pathological speech. Thus, the output of the calculations maps to perceptually and clinically relevant aspects of speech, rather than to standardized values that may not be meaningful to the SLP. While still in infancy, the algorithms used for this process are undergoing iterative development.They provide a promising integrated signal processing-perception tool for mobile speech applications as they have the potential to maximize benefit.

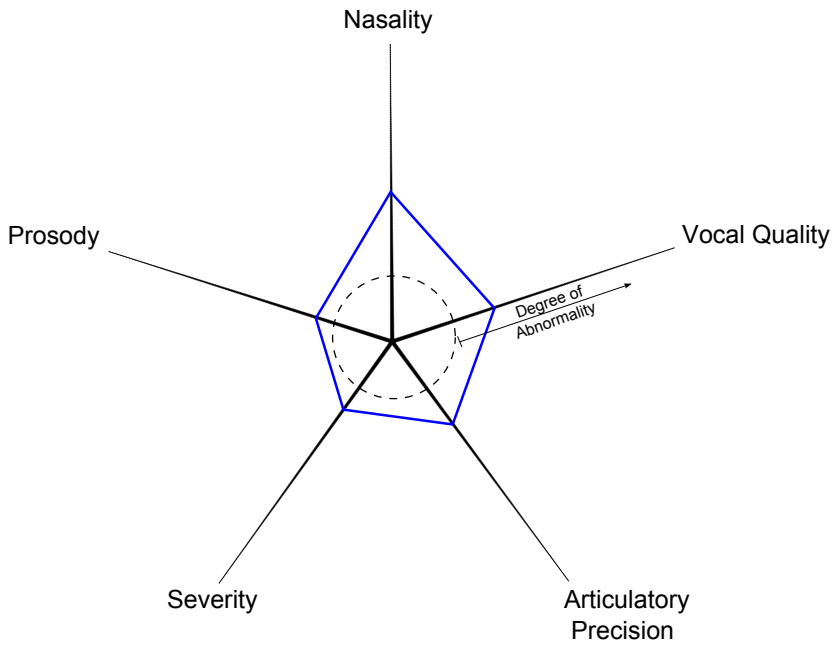
4.3 Conceptual Interface

Figure 4.3 demonstrates the potential conceptual interface. Next, we outline an example interaction between an SLP and a patient using the proposed mobile application suite.

1. Initially, an SLP assigns a set of evaluation modules to a particular patient.
    - A module corresponds to a set of designated set of acoustic features (e.g., the vowel space area, the DDK rates, etc.).
    - Standard reading passages used by SLPs are utilized (e.g., grandfather passage, rainbow passage), along with novel phrases.
2. Once received, the patient completes the requested recordings (e.g., speech and video or speech only).
    - The media is sent to the secure data center, and, once ready, it is downloaded by the SLP; an evaluation report that characterizes the speech signal is automatically generated.
3. The SLP can review recordings, modify the report, record a video to the patient, and assign new modules.



(a)



(b)

Figure 4.2: Example of (a) perceptual rating of SLPs and algorithm; (b) visual display of the deviation of ratings from normal.

Figure 4.3: Example interface design; here, we characterize the interaction between an SLP a the patient using the proposed mobile application suite.

Chapter 5

MEAN FORMANT TRAJECTORIES

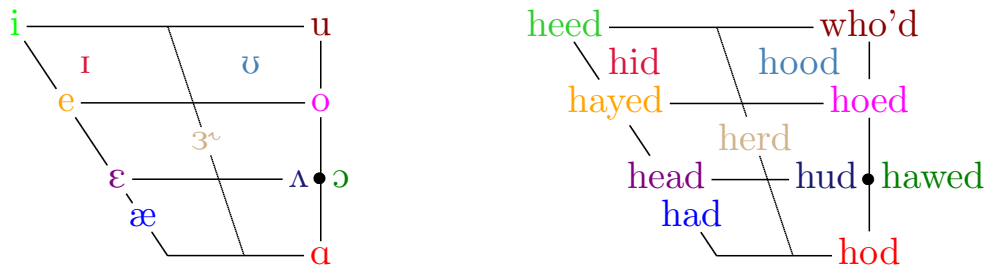
5.1 Motivation of Proposed Method for Quantifying Mean Formant Trajectories

The use of formant frequencies has played a central role in the development and testing of theories on vowel recognition since popularized by the seminal study of vowels by Peterson and Barney [26]. Over the last 60 years, many different studies have established the role of the first two formant frequencies, $F1$ and $F2$, as the main determinants of vowel quality [26–28, 109, 131]. These studies range from research on vowel recognition [132–140], speech perception [141, 142], articulatory-to-acoustic modeling [143, 144], and acoustic phonetic cues [26, 145]. All of the aforementioned studies have shown high correlation between the first two formant frequencies and phonetic height and backness. Since relative values of the first and second formants roughly relate to the size and shape of the cavities created by the jaw opening ($F1$) and tongue position ($F2$), the formant frequencies form an acoustic proxy for the kinematic displacements of the articulators [118]. The preceding insights have led to a convenient phonetic/acoustic/perceptual portrayal of vowels, called a vowel diagram, which is formed by arranging the vowel tokens in the $F1 - F2$ space [131, 146, 147] with axes chosen appropriately. An example of a vowel diagram and corresponding words in /hVd/ context [128] is shown in Figure 5.1.

As useful as $F1-F2$ measurements and the illustrative vowel diagram have proven to be, there is also a large body of evidence indicating that dynamic properties, such as duration [148–151] and spectral change [7, 150–157], play an important role in vowel perception. For example, some vowels may have long or short vowel onglides or

offglides, resulting in a considerable displacement of the formant frequencies across duration from the values at the temporal midpoint [131, 158–165]. Although the effectiveness of the first two formant frequencies in vowel identification is indisputable, it has also been recognized that information derived from beyond the temporal midpoint can provide many kinds of cues to vowel quality [131]. For example, acoustic classification studies [7, 164, 166–169] have shown that (a) vowels are more effectively separated when the acoustic parameters are based on spectral information extracted at multiple time points, rather than at a single time instance; (b) spectral change patterns aid in the statistical separation of vowels in both fixed and variable phonetic environments [169]; and (c) static vowel targets are not necessary for vowel identification, nor are they sufficient to explain the very high levels of vowel intelligibility reported in studies such as Peterson and Barney [26] and Hillendbrand et al. [7]. It was also shown that formant trajectory is beneficial for the within-class separation of the tense/lax monophthong pairs [131]. There have been many studies on vowel inherent spectral changes, that is formant changes associated with monophthongs [26, 153, 170–172]. In fact, all but a few nominally monophthongs show a significant amount of spectral movement through the courses of the vowel, even when those vowels are spoken in isolation [169]. However, the discussion of formant changes is far more prevalent in studies of diphthongs [173] than monophthongs, where vowel duration is typically used as an additional feature to classify vowels, rather than considering the formant trajectories [131].

The established practice of static vowel representation in phonetic/acoustic/perceptual space, rather than trajectories through that space, remains in use despite several authors pointing out that this oversimplification has fundamental limitations which are not always acknowledged in interpretation [169]. Although it has been suggested in the literature that spectral change, such as the trajectory of vowel for-



(a) (b)

Figure 5.1: The International Phonetic Alphabet (IPA) [6] vowel trapezium showing (a) American English vowels; and (b) the corresponding /hVd/ context words; used by Hillenbrand et al. [7].

mants, may be useful in the identification and classification of vowels, very little work has been done to quantify the progression of formant trajectories. Some initial attempts to quantify formant trajectories only utilize a coarsely sampled two point trajectory [153, 174, 175]. Other studies that included more detailed trajectories were limited to only a few speakers [168, 176–178], or a single dialect region [172, 179], or a specific range of ages [171], or a single word context (e.g., isolated vowels or single consonant-vowel or consonant-vowel-consonant context) [172, 180, 181].

In this work, we propose an initial analysis for quantifying formant trajectories using a wide range of speakers, dialects, and coarticulation contexts, while also assessing the formants throughout the full duration of phoneme production. For this analysis, we use two popular speech databases to offer Mean Formant Trajectories (MFTs) that are representative of standard American English using a the proposed automated analysis framework.

In the first study, we use the Hillenbrand database, allowing for the comparison of our proposed analysis to a widely cited assessment of vowel characteristics. In our study, we examine formant trajectories on the comprehensive TIMIT database, which offers several dialects and coarticulation contexts and allows the examination

of not only vowels but also other phoneme types. Our studies demonstrate that phoneme tokens which lie close to each other in the $F2-F1$ space, preventing easy discrimination based on the $F2-F1$ at the temporal midpoint, often exhibit formant trajectories progressing in different directions, allowing easy visual discrimination when a formant trajectory is utilized. Use of the third formant, $F3$, in MFT is also succinctly examined.

5.2 Analysis Study Using the Hillenbrand Database

The first study examines the Mean Formant Trajectories (MFTs) present in the database provided by Hillenbrand et al. [7]. The analysis method id based on averaging formant trajectories for each vowel token using their pre-computed values; tokens are considered for four classes of speakers based on gender and age.

5.2.1 Analysis Method to Quantify Mean Formant Trajectories

5.2.1.1 Database Description and Formant Frequencies

The Hillenbrand et al. [7] database consists of recordings of /hVd/ utterances spoken by 45 men, 48 women, and 46 children (27 boys and 19 girls) sampled at 16 kHz. The database includes values of the formant frequencies. These values were calculated using Linear Predictive Coding (LPC) [28] analysis with 16 ms Hamming window and an 8 ms frame advance. A three-point parabolic interpolator was used to achieve finer resolution than a 61.5 Hz frequency quantization. Note that formant frequency values were verified and hand edited to correct any tracking errors that occurred. The formant frequencies are provided for 10-80% vowel duration at 10% increments.

There are various limitations to this database. Specifically, as it only includes 139 subjects, it is relatively small in size. Also, it has limited dialect variation (87% of the subjects were raised in Michigan's lower peninsula), the words were spoken only

in /hVd/ context, and it only uses one instance of each word per speaker.

5.2.1.2 Computation of Mean Formant Trajectories

As the formant frequencies were pre-computed for the Hillenbrand data, we only need to perform trajectory averaging to obtain the MFTs. Using MATLAB [182], utterances corresponding to a common vowel token are collected and the mean formant values across the utterances, at each temporal point relative to the vowel duration, are computed. This results in a mean trajectory in the $F1 - F2$ space for each of the tokens in the database.

5.2.2 Results and Discussion

We demonstrate our results by plotting the MFT for each token in the database in the $F1 - F2$. The resulting plots differ from the standard vowel diagrams in that they represent each token by a curve, instead of a point, in the $F1 - F2$ space. We illustrate how a single formant trajectory is formed in Figure 5.2. Once a single trajectory if formed it is sampled at 10 points relative to its duration. The process is repeated and the mean values for each token relative to duration are computed. The mean trajectory for each token in the database can be plotted in the $F2 - F1$ space, and by choosing the axes properly results in a plot similar to the standard IPA vowel trapezium.

Figure 5.3 shows the MFTs for each of the tokens in the Hillenbrand database (i.e., 12 American English vowels) for each of the speaker groups. The Hillenbrand database can be used to highlight the difference in MFT based on age group, in addition to gender. The female and male children have very similar vowel trajectories; however, there is notably more variation and higher formant values among the female children when compared to the male children. Previously, Pettinato et al. [183] found that the

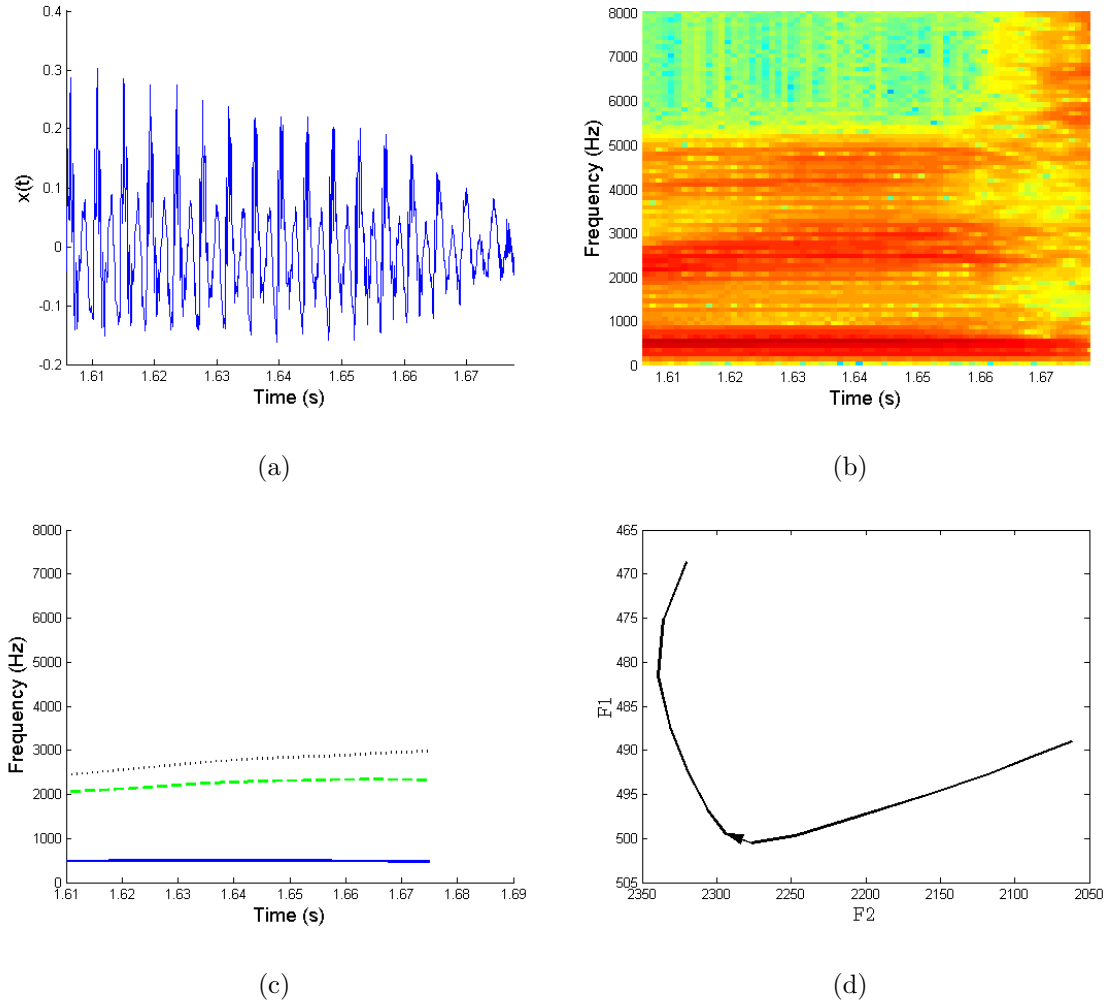


Figure 5.2: (a) A speech segment; (b) Short-time Fourier Log Magnitude Spectrum of the segment in (a); (c) The first three formants, $F1$ (blue solid line), $F2$ (green dashed line), and $F3$ (black dashed line), for the speech segment in (a) using the proposed method. (d) The formant trajectory of the segment in (a) formed by plotting $F1$ versus $F2$ with the axes chosen appropriately.

two-dimensional vowel space area, derived from the first and second formant frequency coordinates of vowels, was significantly larger for children compared to adults. In contrast, our results in this study show that the MFTs for adult females exhibit only slight compression and slightly lower formant values than the male children. Note however, that the adult male MFTs are noticeably lower than the corresponding values of the other three groups. As expected, the trajectory arrangement of the

vowels is, in general, consistent across age and gender, exhibiting only shifts in value and changes in scale. Importantly, the MFTs are nearly identical in direction of progression across the four groups.

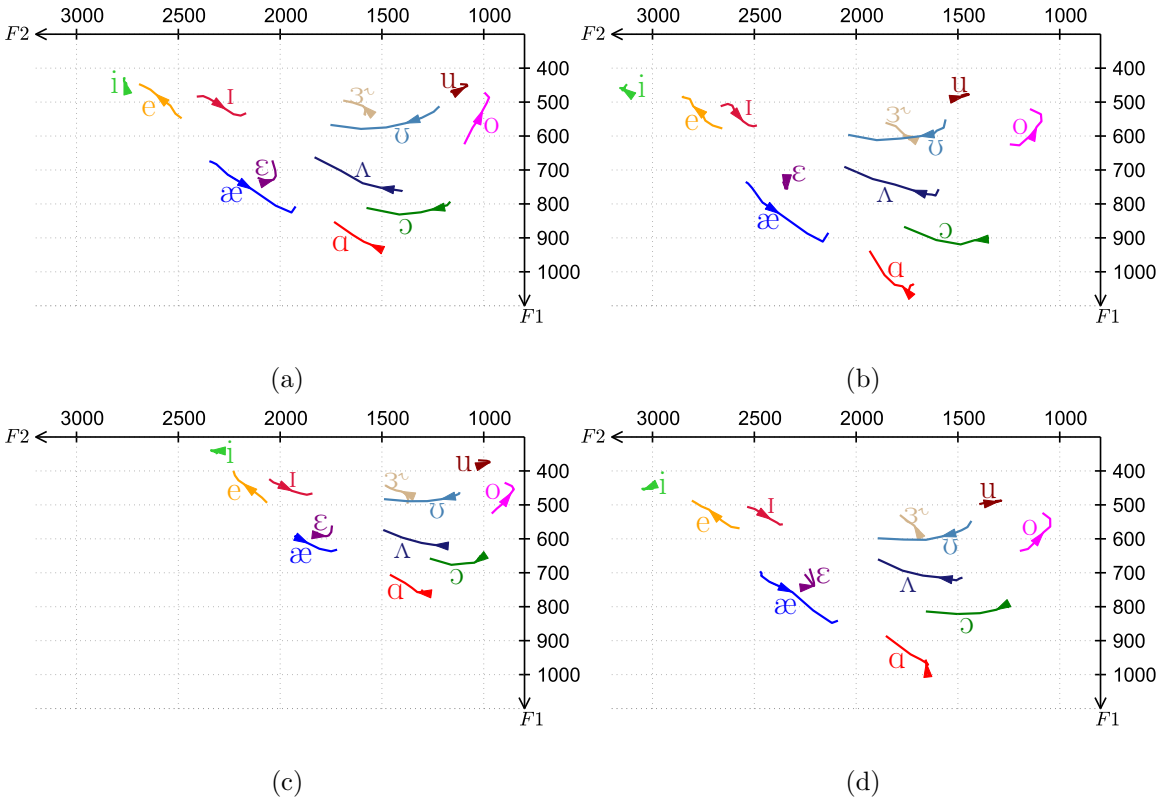


Figure 5.3: The MFTs for (a) adult females; (b) female children; (c) adult males; (d) male children; taken from the Hillenbrand database. The same axis limits are used in each of the plots to facilitate comparison and have been chosen so that the plots have the same orientation as the standard IPA vowel trapezium. Direction is indicated by an arrow (➙) which is placed at the mean $F1 - F2$ value at 50% vowel duration. Note that this arrow may not be centrally located along the length of the trajectory, thus it can be used to infer if there is more variation early in the trajectory or later in the trajectory.

Hillenbrand et al. [7] pointed out that formant frequencies $F1$ and $F2$, taken at a single time point, do not provide adequate predictors for vowel identification. The example in [7] using /æ/ and /ɛ/ were easily identified by listeners even though they were poorly separated in the static $F1 - F2$ space. We note that when the vowel trajectory is considered, we find that these tokens are nearly perpendicular to each

other. Similarly, /ʊ/ and /ɜ˞/ appear very close to one another at the temporal midpoints; however, they also exhibit trajectories that progress at approximately 45° from one another. This offers an explanation for the listeners' ability to accurately identify these tokens that is eluded by utilizing only midpoint measurements.

When considering the results from this study, it is important to note several limitations. First, the Hillenbrand database, albeit widely used, is relatively small and the speakers are quite homogeneous, in that they are all from the same dialectical region of the United States. Further, the vowels utilized in the study are all spoken in the /hVd/ context, providing a single articulatory and coarticulatory context. While this database provides an important foundational ground for the study of acoustical phonetics, it provides limited ecological validity for extrapolating findings. The results of this study provide substantial proof of concept of this method and a point of comparison for the use of a much larger, representative database, as described next.

5.3 Analysis Study Using TIMIT Database

The second study examines the MTFs present in the TIMIT database [120] for adult female and adult male speakers. The phonemes considered include vowels, as in the first study, along with diphthongs, semivowels, glides, stops, fricatives, and affricates [6, 128]. Results are provided in the form of MFT plots in the $F1 - F2$ space and in the form of tables with descriptive statistics.

5.3.1 Analysis Method to Quantity Mean Formant Trajectories

Next consider quantifying MFTs when the values of the formant frequencies have not been pre-computed. In order to determine the MFTs for each phoneme token, three steps are necessary. First, the formant frequencies must be extracted from the

acoustic signal. Second, the value of the formant frequencies must be determined at the relative temporal increments across the duration of each utterance. Finally, the MFT must be computed across utterances at each of the temporal points. This is performed for a series of sounds, as described next in detail. Moreover, although they are usually only considered in relation to vowels, formants can be similarly applied to other phonemes provided a formant is defined as a concentration of acoustic energy around a particular frequency. As such we extend the MFT analysis study to include phonemes as well as vowels.

5.3.1.1 Database Description and Formant Frequencies

In an attempt to overcome the limitations of the Hillenbrand database, we consider speech samples drawn from the TIMIT [120] database commissioned by DARPA. The TIMIT database consists of 6,300 sentences, with 10 sentences spoken by 630 speakers from 8 major dialect regions [128] of the United States. Although the database consists of only adults, it contains a wide variety of speakers. The TIMIT database includes hand verified and time-aligned orthographic and phonetic word transcriptions, as well as 16-bit, 16 kHz speech waveform files for each utterance. The database design was a joint effort among the Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI) International, and Texas Instruments (TI), Inc. The speech files in the TIMIT database consist of phonetically-diverse sentences intended to expose dialectal variants of speech. In particular, the files include sentences instead of isolated /hVd/ productions as in the Hillenbrand database. Note that for ease of comparison, our study maintained the grouping of the phoneme classes (vowel, semivowel or glide, stop, fricative or affricate, nasal), as specified in the TIMIT documentation. However, chose to separate the diphthongs and vowel variants (rhotic, centralized, fronted, and voiceless) from the rest of the vowels to allow for closer

comparison to the vowels in the Hillenbrand database.

5.3.1.2 Extraction of Formant Frequencies

We extract the formant frequencies using the method we proposed in Chapter 3 for automatic assessment of vowel space area [52]. A Praat [121] script is used to automatically extract formant frequencies on a frame-by-frame basis. The Praat script assesses voicing on a frame-by-frame basis by estimating periodicity using an autocorrelation-based method. In this study, we only consider the first three formants; however, using the recommended Praat values, five formants are extracted per frame below a ceiling value (5,000 male, 5,500 female) in Hz. The other settings in the algorithm are chosen as follows: 5 ms frame advance; 50 ms analysis window; pre-emphasis starting from 50 Hz. Internally, Pratt computes estimates of the formants by resampling to twice the ceiling of the formant search range, then applying a pre-emphasis filter, windowing the speech in the time domain using a Gaussian window, and estimating the LPC coefficients using the algorithm by Burg [122, 123].

5.3.1.3 Computation of Formant Trajectory

Due to the variation in phoneme duration, both across individual utterances and across speakers, we utilize time points corresponding to each utterance's relative phoneme duration to temporally capture the formant trajectory (e.g., formant values at 20 percent of the phoneme duration). Using MATLAB [182] and the meta-data provided with the TIMIT database, the start and end times of each vowel utterance are determined and used to calculate the times corresponding to 0-100% vowel duration at increments of 10%. The time corresponding to relative phoneme durations are likely to fall between the frames in which the formant frequencies are sampled (every 5 ms). As a result, we interpolate the values of the formant frequencies between

analysis frames using a cubic spline in order to obtain more precise temporal values. Processing all input speech results in an $N \times 20$ matrix, $\mathbf{F}$, that stores all $F1$ and $F2$ pairs for a particular phoneme token at each of the 10 temporal points, where N is the number of phoneme observations.

5.3.1.4 Formant Trajectory Averaging

Utterances corresponding to a particular phoneme token are collected and the mean formant values across the utterances, at each temporal point relative to the phoneme duration, are computed. This results in a mean trajectory in the $F2 - F1$ space for each of the tokens in the database.

5.3.2 Results and Discussion

5.3.2.1 Vowel MFTs

Table 5.1 summarizes the number of occurrences for each vowel token in the TIMIT database. Figure 5.4(a) and (b) show the MTFs for each of the vowels in the TIMIT database. Table 5.2 and Table 5.3 show a summary of the vowel MFTs values in the TIMIT database at 20%, 50%, and 80% duration for adult female and adult male speakers, respectively.

Although the arrangement of the vowel MFTs is similar in both the Hillenbrand and the TIMIT databases, there are some key differences. Particularly, the trajectories in the TIMIT database exhibit more of a curved trajectory and are more tightly arranged with smaller average $F1$ and $F2$ values. It is not apparent whether these differences result from speaker dialect or coarticulation effects or are due to the different method used in computing the formants. Unlike the vowels in the Hillenbrand database, which had some formant values very close to one another, the vowels in the TIMIT database appear to have a distinct region of occurrence. As expected, the

male vowel MFTs are noticeably compact and have lower value when compared to the female vowel MFTs in both databases.

5.3.2.2 Diphthong and Vowel Variant MFTs

Table 5.4 summarizes the number of occurrences for diphthongs and vowel variants in the TIMIT database. Figure 5.4(c) and (d) show the MFTs for each of the diphthongs and vowel variants in the TIMIT database overlaid on the vowel MFTs from the same database (from Figure 5.4(a) and (b)) for adult female and adult male speakers, respectively. Tables 5.5 and 5.6 show a summary of the average diphthong and vowel variant formant values in the TIMIT database at 20%, 50%, and 80% duration for adult female and adult male speakers, respectively. We note that, due to a lack of a standard IPA symbol for fronting, /u̟/ has been used to denote a fronted allophone of /u/.

In general, the female and male MFTs are in agreement; however, the male MFTs exhibit a very noticeable compacting and lowering of formant values. Additionally, there are some noticeable differences between the shape and direction of MFTs of male and females. For example: /ɨ/ resembles an upward angled cup for females (◡) and a downward angled cup for males (◠); /ə̹/ is mostly one directional for males but it is distinctly two directional for females; /ɜ˞/ and /ə˞/ start and end closer to the center of the vowel space for males but not for females.

Similarly to the vowel MFTs in the Hillenbrand study, the vowel MFTs in the TIMIT study that are close in the $F1-F2$ space have different directions.For example, /aɪ/ and /aʊ/ have very close formant values, especially at the temporal midpoint. However, their trajectories have opposite dierections and the formant values of /aɪ/ have an overall greater deviation from the temporal midpoint.

5.3.2.3 Semivowel and Glide MFTs

Table 5.7 summarizes the number of occurrences for semivowels and glides in the TIMIT database. Figure 5.4(e) and (f) show the MFTs for each of diphthongs and vowel variants in the TIMIT database overlaid on the MFTs (from Figure 5.4(a) and (b)) for adult female and adult male speakers, respectively. Tables 5.8 and 5.9 show a summary of the average semivowel and glide formant values in the TIMIT database at 20%, 50%, and 80% duration for adult female and adult male speakers, respectively.

As in the pervious phonemes, the female and male MFTs are very similar. Also, the male trajectories exhibit a very noticeable compacting and lowering of the trajectory values. Similar to the Hillenbrand vowels, tokens that are close in the $F2 - F1$ space travel in different directions. For example, /h/ and /ɦ/ are relatively close in the $F2 - F1$ space, but traverse in opposite directions. The same can be said about /l̥/ and /w/.

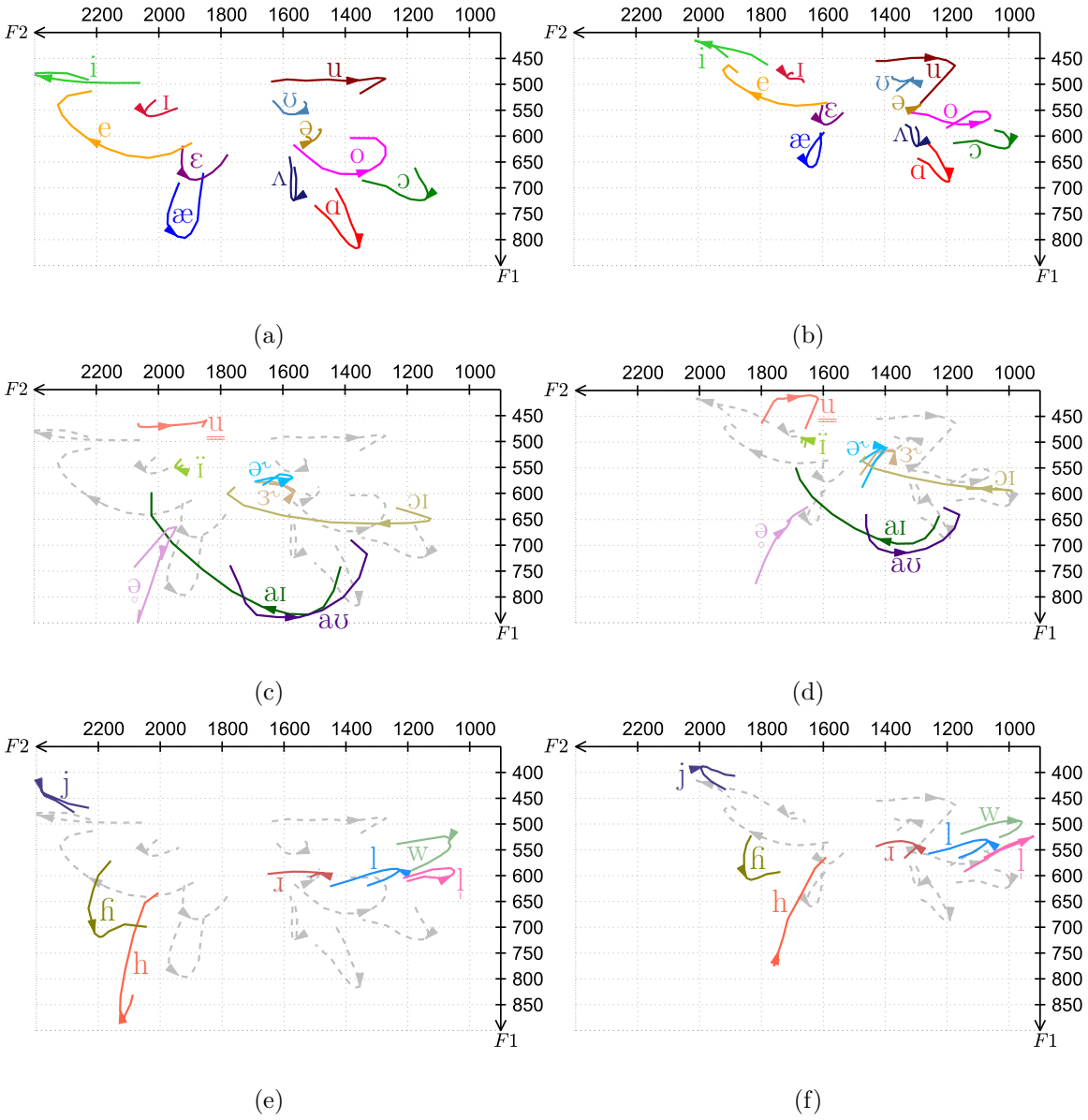


Figure 5.4: The average formant trajectories in the TIMIT database for adult (a) female vowels; (b) male vowels; (c) female diphthongs and vowel variants; (d) male diphthongs and vowel variants; (e) female semivowels and glides; (f) male semivowels and glides. Direction is indicated by an arrow (→) which is placed at the mean $F2 - F1$ value at 50% vowel duration. Note that this arrow may not be centrally located along the length of the trajectory, thus it can be used to infer if there is more variation early in the trajectory or later in the trajectory. Note that the vowel MRTs for females (from (a)) is shown overlaid in (c) and(e) for comparison using a grey dashed line (---). Similarly, the vowel MFTs for males (from (b)) is shown overliad in (d) and (f).

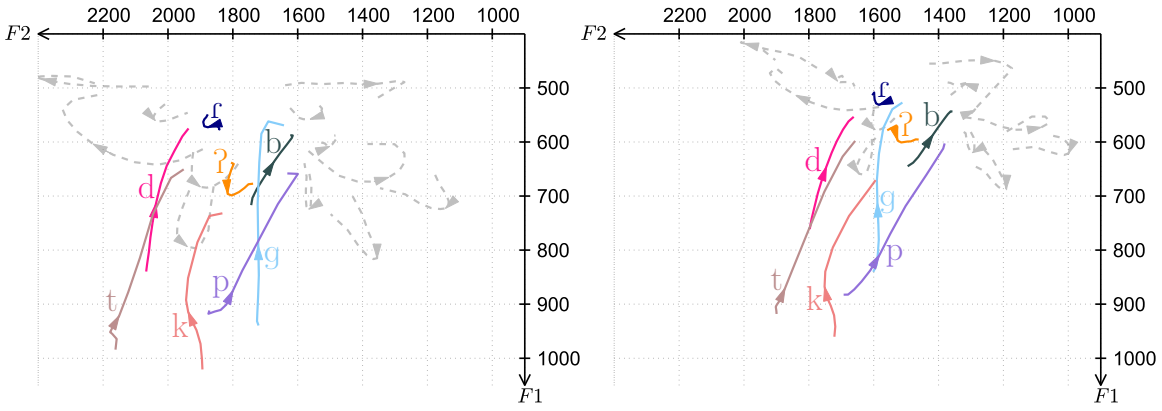


Figure 5.5: The stop MFTs for adult (a) female; (b) male; speakers in the TIMIT database. The stops have been overlaid on the vowels from Figure 5.4 and are displayed using a grey dashed line (---). Direction is indicated by an arrow (→) which is placed at the mean $F2 - F1$ value at 50% vowel duration. Note that this arrow may not be centrally located along the length of the trajectory, thus it can be used to infer if there is more variation early in the trajectory or later in the trajectory.

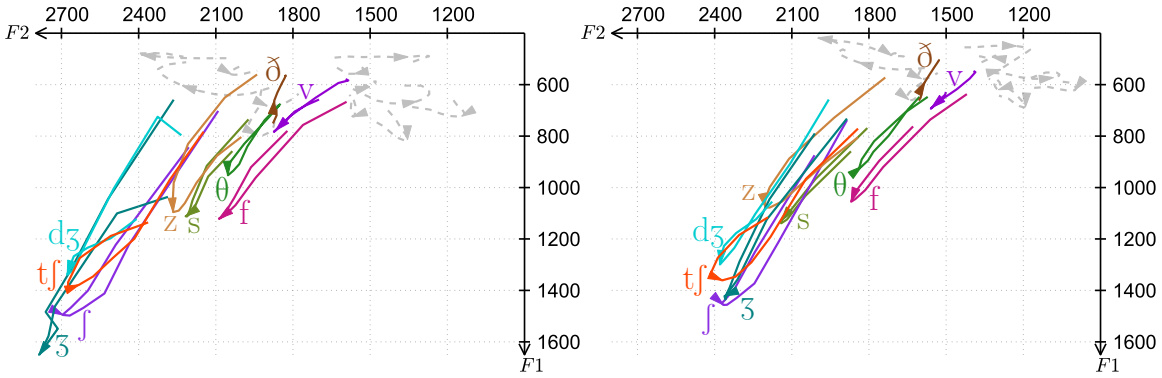
5.3.2.4 Stop MFTs

Table 5.13 summarizes the number of occurrences of stops in the TIMIT database. Figure 5.5 shows the MFTs for each of the stops in the TIMIT database overlaid on the vowel MFTs from the same database (from Figure 5.4(a) and (b)). Tables 5.14 and 5.15 show a summary of the average stop formant values in the TIMIT database at 20%, 50%, and 80% duration for adult female and adult male speakers, respectively.

As in the previous phonemes, the female and male MFTs are very similar with the male MFTs exhibiting a very noticeable compacting and lowering of the trajectory values. The MFTs of stop phonemes seem to appear in two categories: 1) /d/, /g/, /p/, /t/, and /k/ all begin with rather large $F1$ and $F2$ values which decrease significantly during the duration of the phoneme; and 2) /ɾ/, /ʔ/, and /b/ are located in the frequency range of the vowel trajectories and exhibit a relatively small amount of movement during the duration of the phoneme compared to other stop consonants.

5.3.2.5 Fricative and Affricate MFTs

Table 5.10 summarizes the number of occurrences for fricatives and affricates in the TIMIT database. Figure 5.6 shows the MFTs for each of fricatives and affricates in the TIMIT database overlaid on the vowel MFTs from the same database (from Figure 5.4(a) and (b)). Tables 5.11 and 5.12 show a summary of the average fricative and affricate formant values in the TIMIT database at 20%, 50%, and 80% duration for adult female and adult male speakers, respectively.



(a) (b)

Figure 5.6: The fricative and affricate MFTs for adult (a) female; (b) male; speakers in the TIMIT database. The diphthong and vowel variants have been overlaid on the vowels from Figure 5.4(a) and (b) respectively and are displayed using a grey dashed line (---). Direction is indicated by an arrow (→) which is placed at the mean $F2-F1$ value at 50% vowel duration. Note that this arrow may not be centrally located along the length of the trajectory, thus it can be used to infer if there is more variation early in the trajectory or later in the trajectory.

Most fricatives and affricates exhibit a positive swing in both $F1$ and $F2$, which is expected because these phonemes are traditionally characterized by relatively high frequency noise. Unlike the previous phoneme types, which exhibit a very noticeable compacting and lowering of the formant values for the male trajectories, this trend is less robust for fricatives and affricates. We conjecture that this is because the predominate determinant of fricative and affricate acoustics is the manner of articulation and place of constriction; this is in contrast to other phonemes (namely vowels), for

which variation of the overall vocal tract results in these differing characteristics. In other words, affricates and fricatives are possibly less influenced by the differences in male and female anatomies.

The fricative and affricate MFTs seem to appear in three categories: 1) /v/ and /ð/ have all $F1$ values less than 800 Hz and all $F2$ values less then 1,900 Hz; 2) /z/, /s/, /θ/, and /f/ have all $F1$ values less than 1,200 Hz and all $F2$ values less then 2,300 Hz; 3) /ʃ/, /ʒ/, /dʒ/, and /tʃ/ have all $F1$ values less than 1,700 Hz and all $F2$ values less then 2,800 Hz. Most of the fricative and affricates begin with relatively low $F1$ and $F2$ values which rapidly increase to a maximum near the temporal mid point, before then rapidly falling and returning to lower $F1$ and $F2$ values. This is in stark contrast to the stop formant trajectories where most of the phonemes begin with large $F1$ and $F2$ values that decrease during the duration of the phoneme. Also, unlike previous phoneme types considered, there is considerable overlap in the formant trajectories of phonemes within the class of fricatives and affricates. Interestingly, most of the overlapping trajectories have similar progressions and do not diverge in different directions.

5.3.2.6 Nasal MFTs

Table 5.16 summarizes the number of occurrences of nasals in the TIMIT database. Figure 5.7 shows the MFTs for each of the nasals in the TIMIT database overlaid on the vowel MFTs from the same database (from Figure 5.4(a) and (b)). Tables 5.17 and 5.18 show a summary of the average nasal formant values in the TIMIT database at 20%, 50%, and 80% duration for adult female and adult male speakers, respectively.

Unlike the rest of the phoneme types considered thus far, the configuration of nasal trajectories is substantially different when comparing female and male speakers. Only /ɾ̃/ seems to appear with some consistency in the two speaker groups. This may be

the case because /ɾ̃/ is not a formal nasal but rather a nasalized flap. We conjecture that this is secondary to the retention of the stop-like qualities of the flap, as this is consistent with the patterns seen when examining the non-nazalized version of this stop consonant. Furthermore, we conjecture that the substantial variation of the rest of the nasal trajectories results from the fact that the predominate determinant of nasal quality, the nasal cavity, cannot be reconfigured like the rest of the vocal tract, and, as a result, could exacerbate the speaker dependence of these sounds.

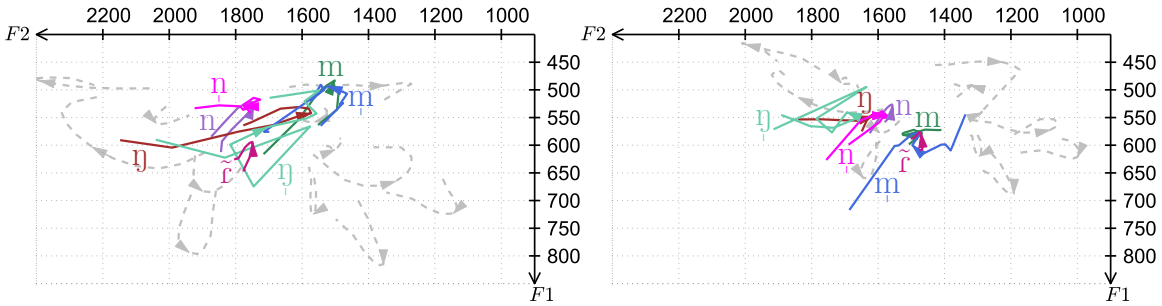


Figure 5.7: The nasal MFTs for adult (a) female; (b) male; speakers in the TIMIT database. The diphthong and vowel variants have been overlaid on the vowels from Figure 5.4(a) and (b) respectively and are displayed using a grey dashed line (---). Direction is indicated by an arrow (→) which is placed at the mean $F2 - F1$ value at 50% vowel duration. Note that this arrow may not be centrally located along the length of the trajectory, thus it can be used to infer if there is more variation early in the trajectory or later in the trajectory.

5.3.2.7 Three-dimensional MFTs utilizing $F3$

In this study, $F3$ values are but not reported, primarily due to the limitations of displaying three-Dimensional (3-D) data using two-Dimensional (2-D) media; nevertheless, $F3$ values have been found to be useful for distinguishing certain phoneme types, e.g., rhotic vowels and velar consonants. As a result, we provide animations of the MFTs in 3-D space using the first three formants, $F2$, $F1$, and $F3$, in the electronic version of this report in [184]. These illustrations utilize all of the phonemes

considered in our second study using the TIMIT database. They are plotted in a similar fashion to the previous figures, but in 3-D space and without token labels.

As the illustration shows, most of the phonemes lie very close to a 2-D hyperplane of the 3-D space. This graphical representation shows a general lack of independence between the first three formants and suggests that inclusion of $F3$ is superfluous for many, but not all, phonemes. It is interesting to note that /ɜ˞/, /ə˞/, and /ɹ/ appear with drastically lower $F3$ values than other phonemes with similar $F2 - F1$ values. The value of $F3$ is also noticeably lower at the start of /g/ and /k/, and the the end of /ɑ/. Likewise, /l/, /l̪/, and /j/ have larger $F3$ values than other phonemes with similar $F2 - F1$ values. A relative increase in $F3$ value is also true for /z/ and /s/; interestingly, the extent of this difference is far intensified in the MFTs of the adult male speakers compared to the adult female speakers.

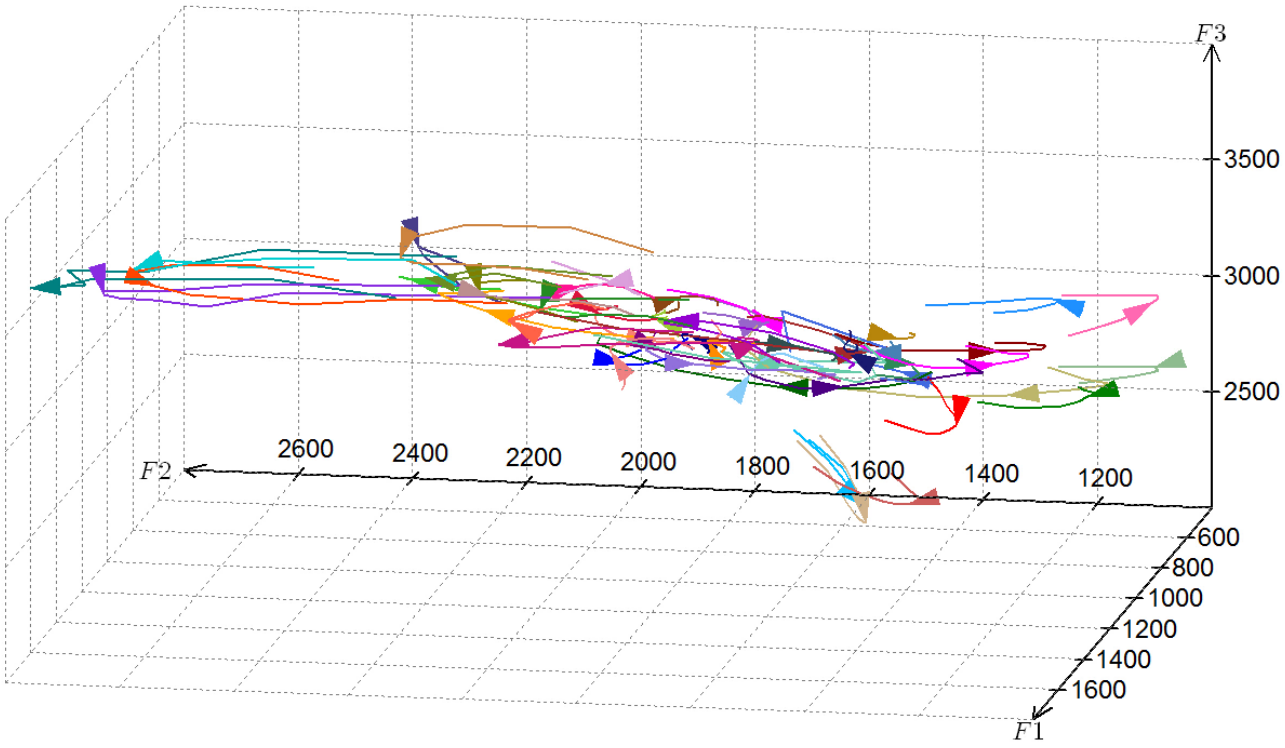


Figure 5.8: 3-D vowel MFTs for adult female speakers in the TIMIT database. Direction is indicated by an arrow (→) which is placed at the mean $(F2, F1, F3)$ value at 50% vowel duration. Note that this may not be centrally located along the length of the trajectory, thus this can be used to infer if there is more variation early in the trajectory or later in the trajectory. Observe that many of the phonemes lie approximately on a 2-D hyperplane of the 3-D space.

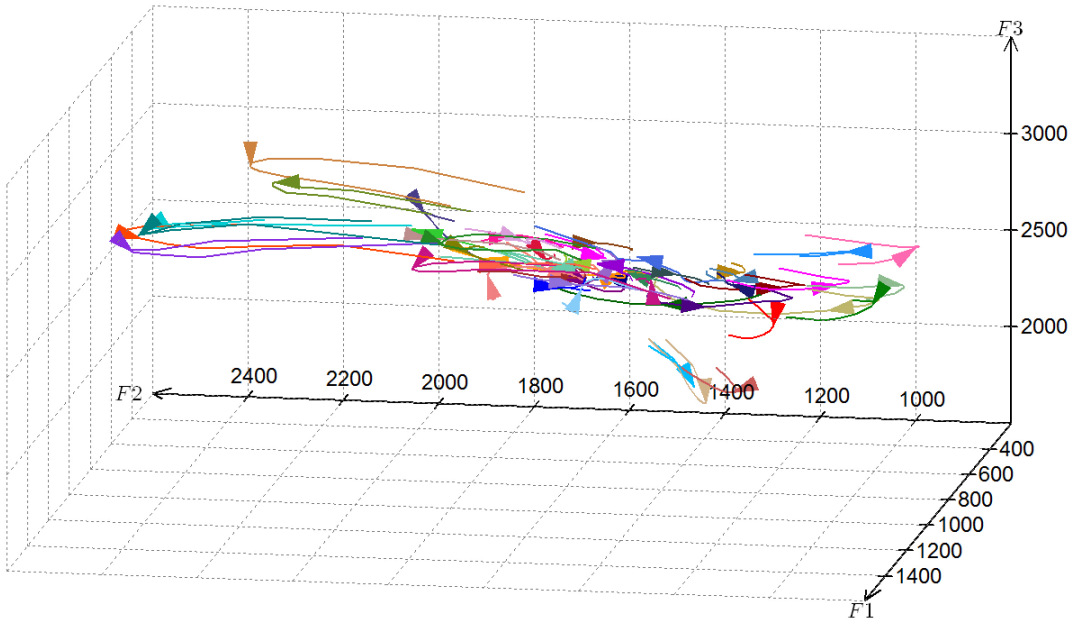


Figure 5.9: 3-D vowel MFTs for adult male speakers in the TIMIT database. Direction is indicated by an arrow (→) which is placed at the mean $(F2, F1, F3)$ value at 50% vowel duration. Note that this may not be centrally located along the length of the trajectory, thus this can be used to infer if there is more variation early in the trajectory or later in the trajectory. Observe that many of the phonemes lie approximately on a 2-D hyperplane of the 3-D space.

Table 5.1: Number of vowel token occurrences utilized in TIMIT database.

Token	/æ/	/ɑ/	/ɔ/	/ɛ/	/e/	/ʊ/	/ɪ/	/i/	/o/	/ə/	/ʌ/	/u/
Female	1,651	1,335	1,152	1,593	935	256	2,258	3,057	860	1,369	1,031	199
Male	3,753	2,859	2,942	3,700	2,152	500	4,498	6,604	2,051	3,584	2,152	524
Total	5,404	4,194	4,094	5,293	3,087	756	6,756	9,661	2,911	4,953	3,183	723

Table 5.2: Mean, μ, and standard deviation, σ, for the vowel MFTs of the female speakers in TIMIT at 20%, 50%, and 80% vowel duration.

	$F1$						$F2$					
	20%		50%		80%		20%		50%		80%	
Token	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
/æ/	750	87	794	86	763	93	1,971	274	1,940	250	1,875	255
/ɑ/	767	93	815	83	792	78	1,382	225	1,355	155	1,413	181
/ɔ/	697	107	723	105	713	90	1,143	201	1,144	162	1,225	192
/ɛ/	653	80	683	75	670	78	1,927	287	1,900	246	1,828	270
/e/	642	76	604	77	540	78	2,032	280	2,231	258	2,325	273
/ʊ/	545	64	558	71	546	68	1,528	312	1,562	289	1,609	297
/ɪ/	544	79	560	79	557	81	2,036	310	2,039	274	1,988	289
/i/	496	74	484	70	478	78	2,234	310	2,388	264	2,356	309
/o/	662	69	665	81	621	97	1,473	274	1,319	231	1,271	252
/ə/	606	91	608	93	592	94	1,512	256	1,501	247	1,479	262
/ʌ/	701	86	724	86	689	89	1,559	237	1,566	196	1,575	218
/u/	493	53	492	71	489	77	1,526	325	1,352	281	1,271	261

Table 5.3: Mean, μ, and standard deviation, σ, for the vowel MFTs of the male speakers in TIMIT at 20%, 50%, and 80% vowel duration.

	$F1$						$F2$					
	20%		50%		80%		20%		50%		80%	
Token	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
/æ/	629	59	659	58	639	70	1,642	179	1,645	157	1,611	171
/ɑ/	653	70	687	62	672	59	1,215	183	1,192	125	1,235	135
/ɔ/	602	75	621	75	617	71	1,002	202	999	160	1,064	164
/ɛ/	557	59	577	57	568	61	1,607	206	1,595	177	1,556	202
/e/	541	57	514	56	473	60	1,687	212	1,840	173	1,920	188
/ʊ/	487	78	492	61	491	75	1,309	282	1,322	243	1,363	256
/ɪ/	480	71	488	63	489	72	1,706	234	1,711	206	1,678	228
/i/	434	82	418	75	420	92	1,885	230	1,999	197	1,984	221
/o/	571	59	573	70	553	92	1,213	207	1,083	174	1,071	213
/ə/	543	81	542	79	538	89	1,297	203	1,288	191	1,284	227
/ʌ/	603	66	620	64	597	68	1,289	196	1,296	148	1,313	168
/u/	452	104	450	93	464	114	1,359	263	1,227	226	1,174	246

Table 5.4: Number of diphthong and vowel variant token occurrences utilized in TIMIT database.

Token	/aɪ/	/aʊ/	/ɝ/	/ɚ/	/u̟/	/ɨ/	/ə̹/	/ɔɪ/
Female	998	298	952	1,339	750	3,603	88	292
Male	2,243	647	1,894	3,451	1,738	7,979	405	655
Total	3,241	945	2,846	4,790	2,488	11,582	493	947

Table 5.5: Mean, μ, and standard deviation, σ, for the diphthong and vowel variant MFTs for the female speakers in TIMIT at 20%, 50%, and 80% vowel duration.

	$F1$						$F2$					
	20%		50%		80%		20%		50%		80%	
Token	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
/aɪ/	819	83	819	96	696	106	1,469	167	1,667	191	1,955	263
/aʊ/	812	95	839	87	763	102	1,720	264	1,548	214	1,354	205
/ɝ/	591	73	596	72	583	83	1,581	250	1,562	206	1,599	225
/ɚ/	571	80	571	73	564	84	1,605	260	1,570	232	1,590	262
/u̟/	472	57	469	58	462	61	2,054	267	1,957	265	1,856	282
/ɨ/	548	83	551	84	539	86	1,929	288	1,945	276	1,935	293
/ə̹/	789	423	727	474	665	415	2,030	421	1,991	462	1,942	448
/ɔɪ/	649	78	659	64	623	68	1,125	191	1,299	216	1,726	280

Table 5.6: Mean, μ, and standard deviation, σ, for the diphthong and vowel variant MFTs for the male speakers in TIMIT at 20%, 50%, and 80% vowel duration.

	$F1$						$F2$					
	20%		50%		80%		20%		50%		80%	
Token	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
/aɪ/	687	63	687	71	606	84	1,271	142	1,425	141	1,636	190
/aʊ/	690	69	714	60	667	69	1,450	192	1,326	157	1,176	149
/ɝ/	518	72	518	67	516	89	1,386	202	1,369	160	1,406	173
/ɚ/	514	94	511	86	523	124	1,419	217	1,395	194	1,414	217
/u̱/	416	111	411	99	416	138	1,748	211	1,672	212	1,617	239
/ɨ̈/	494	83	492	79	491	101	1,658	226	1,670	212	1,668	232
/ə̹/	696	350	666	349	654	347	1,762	357	1,719	354	1,716	376
/ɔɪ/	593	78	589	63	552	60	1,002	201	1,098	166	1,433	223

Table 5.7: Number of semivowel and glide token occurrences utilized in TIMIT database.

Token	/l/	/ɹ/	/w/	/j/	/h/	/ɦ/	/l̩/
Female	2,481	2,773	1,334	709	368	490	401
Male	5,671	6,288	3,043	1,640	945	1033	893
Total	8,152	9,061	4,377	2,349	1,313	1523	1,294

Table 5.8: Mean, μ, and standard deviation, σ, for the semivowel and glide token MFTs for the female speakers in TIMIT at 20%, 50%, and 80% vowel duration.

	$F1$						$F2$					
	20%		50%		80%		20%		50%		80%	
Token	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
/l/	598	108	587	105	597	103	1,245	255	1,235	256	1,324	312
/ɹ/	594	118	594	111	592	99	1,473	244	1,488	242	1,571	246
/w/	524	98	536	90	562	88	1,081	280	1,068	291	1,096	242
/j/	445	166	438	98	453	78	2,374	278	2,381	267	2,326	298
/h/	872	185	860	179	713	224	2,114	396	2,130	384	2,086	456
/ɦ/	628	200	712	199	708	173	2,217	435	2,212	403	2,166	416
/l̩/	604	74	599	73	585	79	1,113	186	1,059	157	1,063	179

Table 5.9: Mean, μ, and standard deviation, σ, for the semivowel and glide token MTFs for the male speakers in TIMIT at 20%, 50%, and 80% vowel duration.

	$F1$						$F2$					
	20%		50%		80%		20%		50%		80%	
Token	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
/l/	545	94	530	86	539	88	1,091	269	1,074	240	1,148	272
/ɹ/	549	130	537	110	534	105	1,306	201	1,311	193	1,372	198
/w/	495	107	494	89	509	78	1,020	365	962	336	975	278
/j/	404	175	388	118	398	96	1,987	200	1,988	196	1943	202
/h/	770	178	746	183	637	190	1,755	319	1,744	328	1,672	384
/ɦ/	560	143	605	148	597	121	1,856	347	1,847	328	1,783	334
/ɬ/	535	76	525	73	532	85	975	245	927	241	950	274

Table 5.10: Number of fricative and affricate token occurrences utilized in TIMIT database.

Token	/s/	/ʃ/	/z/	/ʒ/	/f/	/θ/	/v/	/ð/	/dʒ/	/tʃ/
Female	3,062	915	1,560	57	943	324	849	1,182	495	332
Male	7,051	2,118	3,483	168	2,184	694	1,855	2,691	1,085	748
Total	10,113	3,033	5,043	225	3,127	1,018	2,704	3,873	1,580	1,080

Table 5.11: Mean, μ, and standard deviation, σ, for the fricative and affricate variant MTFs for the female speakers in TIMIT at 20%, 50%, and 80% vowel duration.

	$F1$						$F2$					
	20%		50%		80%		20%		50%		80%	
Token	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
/s/	1,050	229	1,112	166	1,080	221	2,191	273	2,217	234	2,184	272
/ʃ/	1,401	421	1,496	382	1,412	401	2,597	338	2,694	289	2,533	360
/z/	831	426	1,096	311	982	342	2,208	416	2,266	354	2,175	359
/ʒ/	1,320	755	1,652	491	1,463	599	2,664	410	2,789	359	2,726	358
/f/	1,041	237	1,122	195	962	221	2,029	243	2,088	190	1,945	259
/θ/	830	274	952	199	840	231	1,990	266	2,053	216	1,977	249
/v/	594	201	778	395	706	299	1,623	349	1,867	454	1,794	382
/ð/	700	230	660	215	574	94	1,865	258	1,872	245	1,832	223
/dʒ/	1,234	394	1,339	506	1,005	623	2,598	333	2,678	308	2,497	339
/tʃ/	1,274	310	1,378	406	1,199	481	2,629	245	2,632	299	2,417	368

Table 5.12: Mean, μ, and standard deviation, σ, for the fricative and affricate MTFs for the male speakers in TIMIT at 20%, 50%, and 80% vowel duration.

	$F1$						$F2$					
	20%		50%		80%		20%		50%		80%	
Token	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
/s/	1,078	275	1,129	217	1,124	259	2,096	363	2,141	352	2,119	376
/ʃ/	1,391	353	1,456	317	1,373	342	2,321	274	2,366	249	2,245	322
/z/	888	441	1,078	351	992	377	2,111	520	2,202	458	2,070	445
/ʒ/	1,287	651	1,422	482	1,357	525	2,303	382	2,362	298	2,311	330
/f/	1,000	203	1,055	169	920	212	1,840	227	1,869	188	1,745	255
/θ/	820	253	918	183	798	218	1,780	275	1,829	215	1,723	230
/v/	576	201	690	286	658	258	1,408	326	1,557	400	1,525	367
/ð/	637	231	589	195	515	98	1,597	270	1,582	238	1,534	192
/dʒ/	1,177	426	1,299	508	1,101	614	2,317	309	2,380	289	2,241	325
/tʃ/	1,273	294	1,361	346	1,192	440	2,387	226	2,371	245	2,183	326

Table 5.13: Number of stop token occurrences utilized in TIMIT database.

Token	/b/	/d/	/g/	/p/	/t/	/k/	/ɾ/	/ʔ/
Female	943,	1,510	909	1,124	1,822	2,015	1,019	1,862
Male	2,074	3,253	1,856	2,417	4,070	4,468	2,629	2,969
Total	3,017	4,763	2,765	3,541	5,892	6,483	3,648	4,831

Table 5.14: Mean, μ, and standard deviation, σ, for the stop MFTs for the female speakers in TIMIT at 20%, 50%, and 80% vowel duration.

	$F1$						$F2$					
	20%		50%		80%		20%		50%		80%	
Token	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
/b/	687	236	636	193	593	155	1,726	351	1,673	356	1,618	393
/d/	803	229	714	246	617	212	2,059	266	2,037	249	1,981	247
/g/	915	217	792	218	585	189	1,724	376	1,721	390	1,712	429
/p/	914	228	874	227	715	200	1,855	271	1,799	302	1,661	394
/t/	955	201	921	281	733	283	2,172	255	2,150	271	2,046	279
/k/	973	165	914	192	786	253	1,900	395	1,938	415	1,909	436
/ɾ/	572	109	573	152	558	105	1,844	319	1,877	278	1,888	296
/ʔ/	651	167	696	183	685	154	1,805	425	1,816	459	1,764	504

Table 5.15: Mean, μ, and standard deviation, σ, for the stop MFTs for the male speakers in TIMIT at 20%, 50%, and 80% vowel duration.

	$F1$						$F2$					
	20%		50%		80%		20%		50%		80%	
Token	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
/b/	628	219	586	187	548	161	1,463	355	1,417	356	1,371	373
/d/	720	234	646	246	577	233	1,783	236	1,746	233	1,695	249
/g/	808	234	713	252	574	251	1,587	348	1,589	351	1,570	357
/p/	873	215	809	212	660	189	1,660	275	1,582	302	1,437	364
/t/	905	215	870	284	692	279	1,902	237	1,873	267	1,752	272
/k/	921	166	868	210	737	268	1,722	362	1,752	358	1,676	369
/ɾ/	521	130	531	161	513	133	1,554	254	1,584	242	1,598	246
/ʔ/	573	150	597	150	599	137	1,528	364	1,526	391	1,485	426

Table 5.16: Number of nasal token occurrences utilized in TIMIT database.

Token	/m/	/n/	/ŋ/	/m̥/	/n̥/	/ŋ̊/	/ɾ̃/
Female	1,701	3,099	535	45	246	16	281
Male	3,725	6,466	1,207	126	728	27	1,050
Total	5,426	9,565	1,742	171	974	43	1,331

Table 5.17: Mean, μ, and standard deviation, σ, for the nasal MFTs for the female speakers in TIMIT at 20%, 50%, and 80% vowel duration.

	$F1$						$F2$					
	20%		50%		80%		20%		50%		80%	
Token	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
/m/	503	168	483	188	517	193	1,490	318	1,502	335	1,567	364
/n/	572	126	527	138	533	164	1,812	347	1,746	383	1,794	381
/ŋ/	575	129	542	153	534	186	1,790	585	1,573	540	1,665	582
/m̥/	526	188	493	178	500	256	1,486	306	1,523	324	1,534	419
/n̥/	530	129	531	137	514	163	1,786	347	1,774	379	1,746	426
/ŋ̊/	566	195	567	211	521	178	1,576	460	1,706	543	1,593	534
/ɾ̃/	634	82	594	96	615	83	1,767	249	1,749	233	1,783	244

Table 5.18: Mean, μ, and standard deviation, σ, for the nasal MTFs for the male speakers in TIMIT at 20%, 50%, and 80% vowel duration.

	$F1$						$F2$					
	20%		50%		80%		20%		50%		80%	
Token	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
/m/	572	235	582	273	572	234	1,459	394	1,525	426	1,493	406
/n/	540	147	528	187	534	184	1,556	292	1,557	343	1,575	344
/ŋ/	555	132	548	162	546	172	1,699	450	1,615	454	1,620	450
/m̥/	599	210	600	279	600	269	1,398	348	1,497	428	1,537	433
/n̥/	543	182	546	204	543	216	1,603	314	1,573	355	1,578	377
/ŋ̊/	565	155	540	157	552	175	1,805	397	1,694	413	1,786	477
/ɾ̃/	602	108	578	130	585	104	1,466	228	1,471	218	1,488	202

Chapter 6

LATENT SIGNAL ANALYSIS AND THE ANALYTIC SIGNAL

6.1 Motivation for Proposed Latent Signal Analysis

The interpretation of frequency, as the number of cycles during on unit of time, is well understood. However, the concept of instantaneous frequency is often controversial [32–34, 185–189]. Historically, interest in the definition of instantaneous frequency coincides with the advent of frequency modulation for radio transmission [3, 190]. The most widely accepted methods of instantaneous frequency is that provided by Gabor and Ville [35, 191], but other attempts to define instantaneous frequency were proposed [31, 34]. According to Cohen [3], "[o]ne should keep and open mind regarding the proper definition of the complex signal, that is, the appropriate way to define phase, amplitude, and instantaneous frequency. Probably the last word on the subject has not yet been said."

In the time-frequency literature, the instantaneous frequency is most often interpreted as the average frequency at each unit of time; it is thus computed as the derivative of the phase function of a complex signal [34, 186, 187]. For man-made or naturally observed measurements, the phase function cannot be computed as the signals are real. In most cases, the Analytic Signal (AS) is used in place of the real signal as an approach to obtain some form of instantaneous frequency information. The AS signal is obtained from the Hilbert Transform (HT) of the real signal, and it can be shown to contain all the positive frequencies of the real signal [35, 191]. Note, however, that this approach relies on harmonic correspondence, and in some signal cases, it can be shown to lead to Instantaneous Amplitude (IA) and Instantaneous

Frequency (IF) parameters that do not have meaningful physical interpretations of the signals.

We propose a new approach to analyzing real signals by considering complex extensions of the signals whose IA and IF more accurately represent the signals physical characteristics. We call the approach Latent Signal Analysis (LSA) as it is used to find the latent (or hidden) imaginary part of the observed real signal for the complex signal model with the physically-matched IA and IF functions. Although the existence of other IA/IF parameterizations is not new, Vakman [59–61] argued that the AS is the only physically-justifiable complex extension. However, as we will demonstrate with examples and by using differential equations other than that describing simple harmonic motion, our LSA approach is also physically justified.

6.2 Background on Instantaneous Frequency

It has long been known that simple harmonic analysis of non-stationary signals may lead to incorrect interpretations of an underlying signal model [34]. In his seminal paper, Gabor writes [35]

> "The greatest part of the theory of communication has been built up on the basis of Fourier's reciprocal integral relations .... Though mathematically this theorem is beyond reproach ... even experts could not at times conceal an uneasy feeling when it came to the physical interpretation of results obtained by the Fourier method. After having for the first time obtained the spectrum of a frequency-modulated sine wave, Carson wrote: 'The foregoing solutions, though unquestionably mathematically correct, are somewhat difficult to reconcile with our physical intuitions ....' "

As Gabor noted, Carson in 1922 was the first to rigorously study an Frequency

Modulation (FM) signal and realize that the frequency components obtained from such a non-stationary signal do not describe the physical system in a meaningful way [192]. A similar argument regarding an Amplitude Modulation (AM) signal was made by Priestly, who wrote, "...a non-stationary process in general cannot be represented in a meaningful way by the simple Fourier expansion" [66]. A time domain signal example is given by [66]

$$x(t) = Ae^{-t^2/\sigma^2}\cos(\omega_0 t + \phi_0) \tag{6.1}$$

and its Fourier Transform (FT) consists of two Gaussian functions centered at frequencies $\pm\omega_0$. Thus, the FT contains an infinite number of Simple Harmonic Components (SHCs). This interpretation of the underlying signal model may be incorrect, because this the FT assumes an expansion into components with constant amplitude and constant frequency. For example, an alternative signal model can interpret the signal in (6.1) as having two components at constant frequencies $\pm\omega_0$, with each component having a *time-varying amplitude*, $(A/2)e^{-t^2/\sigma^2}$. Mathematically, these two representations are equally valid and correspond to different families of basic functions used for representation [34].

For non-stationary signals, SHCs may not provide an accurate representation, and the idea of time-varying components[1] , i.e., components with time-varying IA and IF, has arisen in order to account for the non-stationarity [34, 66]. However, the concept of IF is not without its own controversy; this is primarily due to the following four reasons:

1. There is an apparent paradox in associating the words "instantaneous" and "frequency" because frequency *usually* defines the number of cycles undergone during one unit of time [34].

[1]Some authors, such as Ville [191], refer to time-varying components as "instantaneous spectra", or more generally, as functions of time that give the structure of a signal at a given instant.

2. Without assumptions, instantaneous parameterizations of a signal are not unique [3, 32, 34, 193].

3. The commonly accepted definition of IF as the derivative of the phase function of the AS only holds for a limited class of signals [3].

4. Although different quantities, harmonic frequency and IF are often confused likely due to the term "frequency" attached to both [33, 194]. Harmonic frequency is a special case of IF, and the two are *only* equivalent under the assumption of SHCs.

Several authors over the previous decades have shown that problems and paradoxes exist that are related to the definition of IF [3, 31–33, 62, 66, 185, 187, 188, 193, 194]. However on the whole, these problems seem to have been forgotten or ignored.

6.3 Latent Signal Analysis

Many physical phenomena are characterized by the complex signal

$$z(t) = x(t) + jy(t) = \rho(t)e^{j\Theta(t)} \tag{6.2}$$

where $\rho(t)$ is the signal's IA, $\Theta(t)$ is the signal's phase function, and $\Omega(t) = \dfrac{d}{dt}\Theta(t)$ is the signal's IF. Here, we assume that only the real part, $x(t)$, is observed and the imaginary part, $y(t)$, is hidden, i.e., the act of observation corresponds to

$$x(t) = \Re\{z(t)\}, \tag{6.3}$$

where $\Re\{\cdot\}$ denotes the real operator. We thus refer to $z(t)$ as the *latent signal*.

It is often desirable to analyze the latent signal since the IA and IF of $z(t)$ in (6.2) completely characterize the signal physical properties. Thus, the LSA approach becomes that of determining $z(t)$ from the observation $x(t)$, i.e., finding the physically

matched $y(t)$ from the observation $x(t)$. We denote this using the operator $\mathcal{L}\{\cdot\}$ that obtains the estimate of $y(t)$,

$$\hat{y}(t) = \mathcal{L}\{x(t)\}. \tag{6.4}$$

These relations and the LSA approach are illustrated in Figure 6.1(a); the figure demonstrates many latent signal mappings to a single observation under the real operator. In the LSA approach, given $x(t)$, $z(t)$ need to be determined.

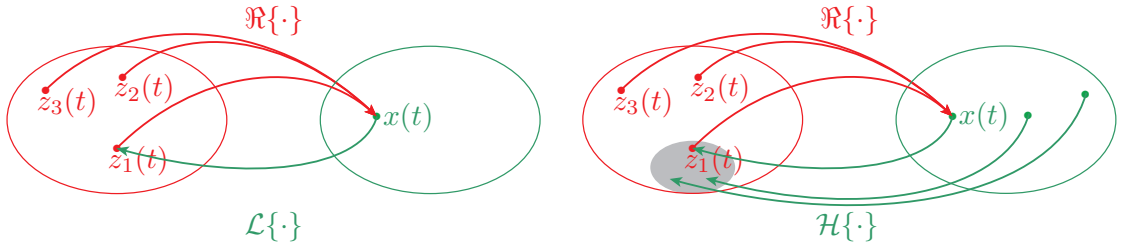


Figure 6.1: Set diagrams for the LSA problem. (a) There are an infinite number of latent signals, $z(t)$ that map to the observed signal $x(t)$, according to $x(t) = \Re\{z(t)\}$. A rule, $\mathcal{L}\{\cdot\}$ is sought out to determine an appropriate latent signal $z(t)$. (b) Most often, the analytic signal is selected using the Hilbert transform operator $\mathcal{H}\{\cdot\}$, thus this limits us to only a subset of latent signals as illustrated by the shaded part.

In the context of time-frequency signal analysis, the instantaneous parameters $\rho(t)$ and $\Omega(t)$ are the variables of interest. The geometric interpretation of the latent signal $z(t)$ in (6.2) is illustrated with the Argand diagram in Figure 6.2. Once a rule for $\hat{y}(t)$ is determined, the instantaneous estimates are given by

$$\hat{\rho}(t) = \pm\,|\hat{z}(t)| = \pm\sqrt{x^2(t) + \hat{y}^2(t)} \tag{6.5}$$

and

$$\hat{\Omega}(t) = \frac{d}{dt}\left[\arctan\left(\frac{\hat{y}(t)}{x(t)}\right)\right]. \tag{6.6}$$

The very definition of IA and IF depends on $\hat{y}(t)$ and hence $\mathcal{L}\{\cdot\}$.

The extension from the real signal to a complex signal is a well-known problem. In 1937, Carson and Fry formally defined the IF based on the phase derivative of a complex FM signal [190, 195]. This assumption, while perfectly valid in communication

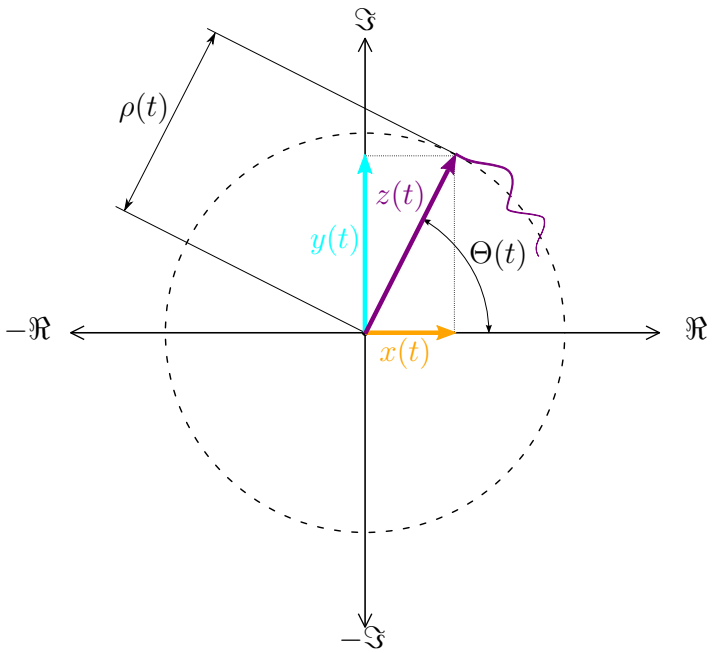


Figure 6.2: Argand diagram of the latent signal $z(t)$ in (6.2) at some time instant. By interpreting the latent signal $z(t)$ (—) as a vector, then the length of the vector is the latent signal's IA $\rho(t)$ and the vector's angular position is the latent signal's phase function $\Theta(t)$. The real part of the latent signal $x(t)$ (—) and the imaginary part of the latent signal $y(t)$ (—) are interpreted as orthogonal projections of vector $z(t)$. We have included an example path (—) taken by $z(t)$.

theory, implies a narrowband component when used in signal analysis. In Gabor's seminal 1946 paper [35], a practical approach for obtaining the complex signal extension of a real signal was introduced. Gabor's method assumed positive IF SHCs and was shown to be equivalent to the Hilbert transform [35]. Although Gabor's provides a unique method for complex extension, as Cohen pointed out, without strict adherence to the assumption, this results in many counter-intuitive consequences [3]. Ville defined the IF of a real signal by using Gabor's complex extension and then Carson's definition of IF [34]. By defining the IF as the derivative of the phase of Gabor's AS, Ville was able to show that, the average harmonic frequency is equal to the time average of the IF. He then formulated the Wigner-Ville Distribution (WVD) and showed that the first moment of the WVD with respect to frequency yields the

IF. Using Gabor's AS to extend Carson's definition of IF for a real signal results in a number of useful relations. As a result, Gabor's method is almost universally viewed as the correct way to define the complex signal, and subsequently, the correct way to define IA, IF, and phase for real signals, despite the inherent assumption of harmonic correspondence [3]. In this approach, the corresponding estimate of the imaginary part in (6.4) is given by

$$\hat{y}(t) = \mathcal{H}\{x(t)\} \tag{6.7}$$

where $\mathcal{H}\{\cdot\}$ is the HT operator and

$$\hat{z}(t) = x(t) + j\mathcal{H}\{x(t)\} \tag{6.8}$$

is termed the AS. This is illustrated in Figure 6.1(b), where the HT is used to estimate the imaginary part, leading to a subset of the latent signals.

The problem with the aforementioned approach was pointed out by Shekel in [32]. As an example, consider

$$x(t) = \Re\left\{a_0(t)e^{j\left[\int_{-\infty}^{t}\omega_0(\tau)d\tau+\phi_0\right]}\right\}. \tag{6.9}$$

There is an infinite set of pairs of $a_0(t)$ and $\omega_0(t)$ for which $x(t)$ may be equivalently described and hence an infinite set of IA/IF parameterizations. Shekel pointed this ambiguity "with the hope of banishing it [IF] forever from the dictionary of the communication engineer." Others such as Hupert, suggested that, despite this problem, the concept of instantaneous parametrization of real signals was still useful and could be applicable [196, 197].

In [59–61], Vakman demonstrated that this problem can be solved, and a unique complex extension to a real signal can be obtained by imposing constrains on the signal in which he believed to be physically-justified: (a) amplitude continuity, (b) phase independence on scale changes and homogeneity, and (c) harmonic correspondence

(detailed discussion to follow in later sections). Under these constraints, Vakman showed that the unique complex extension is given by estimating the imaginary part as $\hat{y}(t) = \mathcal{H}\{x(t)\}$ in (6.8).

More recently, other authors have proposed solving the problem using different signal constraints, such as bounded amplitude and bounded IF variation, leading to alternate IA/IF parameterizations [186, 193]. Vakman has shown that all these methods violate one or more of the constraints that he proposed and almost all retain harmonic correspondence [63, 198]. However, as demonstrated with examples in Section 6.6, Valkman's constraints do not always result to physically match IA/IF parameterizations; this can sometimes lead to computationally intensive representations or representations with incorrect interpretations.

In our research, we have found that the dogmatic use of the AS does not provide the necessary flexibility for modeling non-stationary signals. As will be discussed in further detail in this chapter, it is more advantageous to allow the assumptions of an underlying model of a real observed signal to *imply* the best matched complex extension to the signal. As a result, assumptions must be made based on the physical properties of the signal before deciding on a representation model.

6.4 Hilbert Transform and Analytic Signal

Although there exist several methods for estimating instantaneous parameters, the use of the HT dominates science and engineering. The HT of $x(t)$ is given by

$$\mathcal{H}\{x(t)\} \equiv -\frac{1}{\pi} \fint_{-\infty}^{\infty} \frac{x(\tau)}{\tau - t} d\tau \tag{6.10}$$

where $\fint$ indicates the Cauchy principle value integral [199, 200]. The three main motivations for use of the HT are: (a) Vakman's signal constraints, (b) the analytic signal is an analytic (holomorphic) function [201] when time is considered complex,

and (c) ease of computation via Gabor's method. The motivations for using the HT are discussed next.

6.4.1 Vakman's Signal Constraints

Vakman proposed the following signal constraints when extending a real signal to complex representation [59–63].

Constraint 1: Amplitude Continuity

Simply stated, amplitude continuity requires that the IA, $\rho(t)$ in (6.2) is a continuous function. This implies that the rule $\hat{y}(t) = \mathcal{L}\{x(t)\}$ in (6.4) must be continuous, i.e.,

$$\mathcal{L}\{x(t) + \epsilon w(t)\} \to \mathcal{L}\{x(t)\} \text{ when } ||\epsilon w(t)|| \to 0, \tag{6.11}$$

where $||\cdot||$ denotes the L^2-norm and $\epsilon w(t)$ is a small variation.

Constraint 2: Phase Independence of Scaling and Homogeneity

Let $x(t)$ have a complex extension, $z(t) = \rho(t)\exp[j\Theta(t)]$, as in (6.2). Then for a real constant $c > 0$, $c\ x(t)$ has associated complex extension $z_1(t) = [c\rho(t)]\exp[j\Theta(t)]$, i.e., only the IA of the complex representation is affected and $\Theta(t)$ and $\Omega(t)$, in (6.6), remain unchanged. This implies that the rule for performing the complex extension is scalable

$$\mathcal{L}\{cx(t)\} = c\mathcal{L}\{x(t)\}. \tag{6.12}$$

Constraints 1 and 2 force the operator $\mathcal{L}\{\cdot\}$ to be linear [63].

Constraint 3: Harmonic Correspondence

Let $x(t) = a_0\cos(\omega_0 t + \phi_0)$, then using harmonic correspondence, the complex extension is given by,

$$\hat{z}(t) = a_0 e^{j(\omega_0 t + \phi_0)}, \tag{6.13}$$

where it is noted that the IA and IF are constant. This implies that

$$\mathcal{L}\{a_0\cos(\omega_0 t + \phi_0)\} = a_0\sin(\omega_0 t + \phi_0) \tag{6.14}$$

and $\hat{z}(t)$ is a SHC with positive IF.

As Vakman showed [59], the HT is the only operator that satisfies the above constraints.

Constraint 4: Phase Continuity

In addition to real signals having a problem with instantaneous parametrization when extended to a complex representation, complex signals also face a problem with IA/IF parameterizations. Specifically, as $\hat{z}(t)$ is constructed using a rule for $y(t)$, the IA/IF pair $\rho(t)$ and $\Omega(t)$ are the variables of interest for signal analysis, and there could be a problem in the parametrization of this coordinate transformation [202]. This problem can be resolved in two possible ways, as follows:

- Constraint 4a: Positive IA, $\rho(t) \geq 0, \ \forall \ t$.
- Constraint 4b: Phase continuity, phase $\Theta(t)$ is a continuous function.

Although Constraint 4a is the traditional choice, Constraint 4b is chosen to ensure that the IF is well defined. This is apparent in (6.5), where the IA may be negative.

6.4.2 Analyticity of the Analytic Signal

Unfortunately, the word "analytic" has two distinct meanings when used in signal processing and, in addition, another meaning in mathematics. A signal is said to be *analytic* if it consists only of non-negative frequency components [3, 67]. *The analytic signal* refers to the complex extension of a real signal using the HT, i.e., Gabor's method [3, 67, 191, 203, 204]. In mathematics, if $z(\mathfrak{t})$, where $\mathfrak{t} \equiv t + j\tau$, is an *analytic function* then the real and imaginary parts of the complex function,

$$z(\mathfrak{t}) = u(t, \tau) + jv(t, \tau) \tag{6.15}$$

satisfy the Cauchy-Riemann (CR) conditions

$$\frac{\partial}{\partial t}u\left(t,\tau\right)=\frac{\partial}{\partial \tau}v\left(t,\tau\right)\quad\text{and}\quad\frac{\partial}{\partial \tau}u\left(t,\tau\right)=-\frac{\partial}{\partial t}v\left(t,\tau\right). \tag{6.16}$$

For the AS, $\hat{z}(t)=x(t)+j\mathcal{H}\{x(t)\}$, such that $t\leftarrow \mathfrak{t}\equiv t+j\tau$, the complex function $\hat{z}(\mathfrak{t})=\hat{u}(t,\tau)+j\hat{v}(t,\tau)$ is an analytic function [201]. Hence, the reason for calling a HT-extended signal, the AS [3, 191].

6.4.3 Gabor's Method

The HT is an ideal operator that in practice is not physically realizable. Gabor's method provides a frequency domain approach that, under certain conditions, is equivalent to the HT. The steps of Gabor's method are as follows.

1. the signal $x(t)$ is decomposed into SHCs, i.e., Fourier spectrum of $x(t)$ is obtained.

2. The magnitude of the non-negative frequency components is doubled.

3. The negative frequency components are set to zero.

Gabor's method results in a complex signal formulated in terms of non-negative spectral frequencies,

$$\hat{z}(t)=\sum_{k=0}^{K-1}a_k\exp\left\{j\left[\omega_k t+\phi_k\right]\right\} \tag{6.17}$$

i.e., the signal is decomposed into SHCs, each having constant IA, a_k, and (non-negative) constant IF, ω_k. By comparing the expressions for $\hat{z}(t)$ in (6.17) and (6.2), it is possible to effectively collapse K SHCs into a single Amplitude Modulation–Frequency Modulation (AM-FM) component and then obtain an IA/IF pair. Implementing Gabor's method using a Fourier Transform (FT) is very convenient and is one reason for the method's popularity.

6.5 Relaxing the Constraint of Harmonic Correspondence

6.5.1 Harmonic Correspondence

We can understand Vakman's motivation for attaching harmonic correspondence to physical signal properties by considering the differential equation

$$\frac{d^2}{dt^2}z(t) + \omega_0^2 z(t) = 0 \tag{6.18}$$

which describes many ideal systems, e.g., an inductor/capacitor (LC) circuit or mass/spring model. The solution to (6.18) is the SHC in (6.13). Any deviation from this ideal case, requires a completely different differential equation. For example, in order to describe a circuit with a resistor or motion with damping, the corresponding differential equation is given by

$$\frac{d^2}{dt^2}z(t) + c\frac{d}{dt}z(t) + \omega_0^2 z(t) = 0 \tag{6.19}$$

where c is a constant. In this case, the solution is not an SHC; it has the form

$$z(t) = a_0 e^{-\nu t} e^{j(\omega_d t + \phi_0)} \tag{6.20}$$

with the AM term $\rho(t) = a_0 e^{-\nu t}$, where ν is the damping coefficient and $\omega_d < \omega_0$ is the natural frequency [65]. As another example, when the differential equation includes time-varying coefficients, non-linearities, or partial derivatives with respect to τ or spacial variables, the resulting solution may include both AM and FM terms, which is also not an SHC [63, 205, 206].

While most authors believe harmonic correspondence to be a reasonable constraint to describe physical phenomena, the aformentoned differential equations describing real systems demonstrate that SHC representations can lead to incorrect interpretations. By not assuming harmonic correspondence, a degree of freedom in our analysis can be gained that allows us to construct other complex extensions that may be better

suited to describing the underlying physical phenomena associated with the signal. Any attempt to find a unique rule to infer $z(t)$ of the form in (6.2) from $x(t)$ in the form of (6.3) *is fundamentally flawed, as no universal rule can exist that can describe all possible physical phenomena.*

6.5.2 Analyticity of the Complex Extension

As will be demonstrated, the HT is not the only approach to provide a complex extension to a real signal. Furthermore, the term "analytic" has in most, cases become too restrictive in signal processing. Other complex extensions of real signals can be constructed that result in analytic functions. Although, choosing $\hat{y}(t) = \mathcal{H}\{x(t)\}$ to obtain the AS $\hat{z}(t)$ ensures $z(\text{t})$ is an analytic function, there are other choices for $\hat{y}(t) \neq \mathcal{H}\{x(t)\}$ such that $z(\text{t})$ is an analytic function.

Theorem 1 If harmonic correspondence is not assumed, there exists at least one choice for the imaginary part, $y(t) \neq \mathcal{H}\{x(t)\}$ that results in $z(\text{t})$ being an analytic function.

Proof. Let $x(t) = a_0 \cos(\omega_0 t)$ and choose the imaginary signal such that

$$z(t) = a_0 \cos(\omega_0 t) + j\alpha a_0 \sin(\beta\omega_0 t) \tag{6.21}$$

with real α and β. The complex function is given by

$$z(\text{t}) = a_0 \cos(\omega_0[t + j\tau]) + j\alpha a_0 \sin(\beta\omega_0[t + j\tau]) \tag{6.22}$$

where

$$u(t,\tau) = a_0 \cos(\omega_0 t)\cosh(\omega_0\tau) - \alpha a_0 \cos(\beta\omega_0 t)\sinh(\beta\omega_0\tau) \tag{6.23}$$

and

$$v(t,\tau) = \alpha a_0 \sin(\beta\omega_0 t)\cosh(\beta\omega_0\tau) - a_0 \sin(\omega_0 t)\sinh(\omega_0\tau). \tag{6.24}$$

It can easily be shown that this choice leads to $z(\text{t})$ satisfying the CR conditions and hence is an analytic function. Any choice of $\alpha \neq 1$ or $\beta \neq 1$ does not imply harmonic correspondence. □

Although Gabor's method provides a simple rule to obtain an IA/IF pair that parameterizes $x(t)$, it may not address the problem of obtaining the latent signal from the observation. The heart of the problem is that no single rule can determine the latent signal from the observation for every possible latent signal because more than one complex signal can map to the same real signal under the real operator. Although a unique rule can be constructed to work for a particular $z(t)$, the same rule cannot generally be used for all signals.

Corollary 1 No unique rule for the imaginary part, $\hat{y}(t) = \mathcal{L}\{x(t)\}$, exists to obtain the latent signal, $z(t)$, from the observation, $x(t) = \Re\{z(t)\}$, for all $z(t)$.

Proof. Consider the latent signal, $z(t)$, of the form in (6.21). Assume a unique rule, $\hat{y}(t) = \mathcal{L}\{x(t)\} \equiv \alpha_0 a_0 \sin(\beta_0 \omega_0 t)$ exists to obtain $z(t)$. This implies a unique $\hat{z}(t) = a_0 \cos(\omega_0 t) + j\alpha_0 a_0 \sin(\beta_0 \omega_0 t)$. If $\alpha_0 \neq \alpha$ or $\beta_0 \neq \beta$, then $\hat{z}(t) \neq z(t)$ and $\mathcal{L}\{\cdot\}$ is not unique. □

6.5.3 Harmonic Conjugate Functions

If $z(\text{t}) = u(t,\tau) + jv(t,\tau)$ is an analytic function, then $u(t,\tau)$ and $v(t,\tau)$ are unique and are *harmonic conjugates* [201]. Even though $x(t) = u(t,0)$ and $y(t) = v(t,0)$ in (6.2), this does not imply that $u(t,\tau) = x(\text{t})$ and $v(t,\tau) = y(\text{t})$, but rather $u(t,\tau)$ and

$v(t,\tau)$ each contain terms from both $x(t)$ and $y(t)$:

$$\begin{aligned} z(t) &= x(t) + jy(t) \\ &= [x_R(t,\tau) + jx_I(t,\tau)] + j\,[y_R(t,\tau) + jy_I(t,\tau)] \\ &= [x_R(t,\tau) - y_I(t,\tau)] + j\,[x_I(t,\tau) + y_R(t,\tau)] \\ &= u(t,\tau) + jv(t,\tau) \end{aligned} \tag{6.25}$$

where R, I denote the real and imaginary parts, respectively. This is easily seen in (6.23), where α and β are present in $u(t,\tau)$ even though α and β appear only in $y(t)$ in (6.21). Thus the problem of finding $y(t)$ cannot be solved by finding a harmonic conjugate of $u(t,\tau)$ because $y(t)$ must be known to obtain $u(t,\tau)$.

6.5.4 Amplitude Modulation–Frequency Modulation Demodulation

The HT is widely used as a demodulation algorithm for Amplitude Modulation–Frequency Modulation (AM–FM) signals. This is typically justified as a valid approach because of Vakman and also by Bedrosian's theorem[2] . The HT can be used to determine the IA/IF for a small subset of latent signals, i.e., those with harmonic correspondence. The HT can also be used to closely approximate the IA/IF for latent signals whenever $\mathcal{H}\{x(t)\} \approx y(t)$ [209]. However, the HT cannot be used in general to obtain the IA/IF of *all* latent signals.

6.6 Example of a Latent Signal Analysis Problem

As an example, consider an LSA problem where the observation is $x(t) = \Re\{z(t)\}$, and $z(t) = \rho(t)\exp[j\Theta(t)]$. Thus, the IA and IF of $z(t)$ are $\rho(t)$ and $\Omega(t)$, respectively. Three solutions to the LSA problem are given next, where two of these solutions have $\hat{y}(t) \neq \mathcal{H}\{x(t)\}$, as illustrated in Figure 6.1(b).

[2]If the product $l(t)h(t)$ consists of a low-frequency factor, $l(t)$, and a high-frequency factor, $h(t)$, of non-overlapping spectra, then $\mathcal{H}\{l(t)h(t)\} = l(t)\mathcal{H}\{h(t)\}$ [63, 207, 208].

Assume three ideal systems: a frequency modulator, an amplitude modulator, and a Linear Time-Invariant (LTI) system in the steady state. All three systems have the same output that corresponds to a triangular waveform $x(t)$. The LSA problem is to find the corresponding latent signal $z(t)$ or more specifically, to find the instantaneous parameters, $\rho(t)$ and $\Omega(t)$.

We define the periodic, even-symmetric triangular waveform over one period T as

$$x(t) = \begin{cases} -2A\omega_0 t/\pi + A, & 0 \le t \le T/2 \\ 2A\omega_0 t/\pi + A, & -T/2 \le t \le 0 \end{cases} \tag{6.26}$$

with amplitude A, and fundamental frequency ω_0. The waveform is illustrated in Figure 6.3.

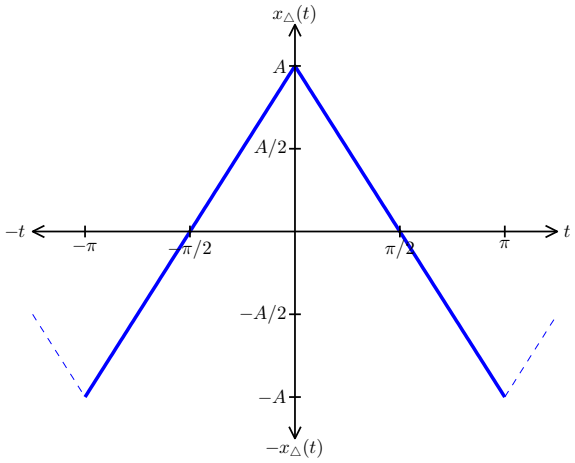


Figure 6.3: One period ($T = 2\pi$) of the triangle waveform, $x(t)$ in (6.26) with amplitude A and $\omega_0 = 1$ radian/s.

6.6.1 Solution Assuming Harmonic Correspondence

If we assume harmonic correspondence, then the complex extension is the analytic signal in (6.8) and the corresponding rule is the HT. The best matched physical system is the LTI system in a steady state. Specifically, the complex signal extension is given

by

$$\hat{z}_1(t) = x(t) + j\mathcal{H}\{x(t)\} \tag{6.27a}$$

$$= \sum_{k=0}^{\infty} \frac{8A}{\pi^2(2k+1)^2} \cos\left[(2k+1)\omega_0 t\right] + j \sum_{k=0}^{\infty} \frac{8A}{\pi^2(2k+1)^2} \sin\left[(2k+1)\omega_0 t\right] \tag{6.27b}$$

where the IA is

$$\hat{\rho}(t) = \frac{8A}{\pi^2}\tilde{a}_0(t) \tag{6.28}$$

and the IF is

$$\hat{\Omega}(t) = \omega_0 + \frac{d}{dt}M_0(t) \tag{6.29}$$

where $\tilde{a}_0(t)$ and $M_0(t)$ are related as

$$\tilde{a}_0(t)e^{jM_0(t)} = \sum_{k=0}^{\infty} \frac{1}{(2k+1)^2} e^{j2k\omega_0 t}. \tag{6.30}$$

6.6.2 Solution Assuming Constant IF

If we assume constant IF, $\hat{\Omega}(t) = \omega_0$, then the best matched physical system is the amplitude modulator. In this case, the complex signal extension is given by

$$\hat{z}_2(t) = \hat{\rho}(t)\cos(\omega_0 t) + j\hat{\rho}(t)\sin(\omega_0 t) \tag{6.31}$$

with the IA given by

$$\hat{\rho}(t) = x(t)/\cos(\omega_0 t). \tag{6.32}$$

Using this approach, the estimated imaginary part is given by, $\hat{y}(t) = \hat{\rho}(t)\sin(\omega_0 t)$. This solution is not the HT of $\hat{\rho}(t)\cos(\omega_0 t)$, because Bedrosian's theorem cannot be applied.

6.6.3 *Solution Assuming Constant IA*

If we assume constant IA, $\hat{\rho}(t) = A$, then the best matched system is the frequency modulator. In this case, the complex signal is given by

$$\hat{z}_3(t) = A\cos\left[\omega_0 t + M_0(t)\right] + jA\sin\left[\omega_0 t + M_0(t)\right] \tag{6.33}$$

with IF

$$\hat{\Omega}(t) = \omega_0 + \frac{d}{dt}M_0(t) \tag{6.34}$$

where

$$M_0(t) \quad = \quad \arccos\left[x(t)/A\right] - \omega_0 t. \tag{6.35}$$

The estimate of the imaginary part is given by, $\hat{y}(t) = A\sin\left[\omega_0 t + M_0(t)\right]$. This is no the HT of $A\cos\left[\omega_0 t + M_0(t)\right]$.

6.7 Discussion

We demonstrated that the harmonic correspondence constraint, that uses the HT to extend real signals to complex, is not necessary for analyticity. We proposed LSA that states that no unique rule exists for the complex extension exists to obtain the latent signal. Although the HT is widely used for demodulation of AM–FM signals, it cannot be used to obtain the IA/IF of all latent signals. In a strict sense, Gabor's method can only be used when the latent signal is a superposition of SHCs with non-negative IF and can approximate the latent signal for communication signals where Bedrosian's theorem is approximately satisfied.

Chapter 7

HILBERT SPECTRAL ANALYSIS AND THE AM–FM MODEL

7.1 Motivation for Hilbert Spectral Analysis

When defining the Instantaneous Frequency (IF) of an observed signal, there is one major drawback of the various approaches discussed in Chapter 6. This drawback is that the IF can only be interpreted as a single, albeit possibly averaged, frequency at any given time. Latent Signal Analysis (LSA) allows for a complex representation of an observed real signal that matches the signal's physical characteristics, in both IF and Instantaneous Amplitude (IA). This complex signal representation is in terms of a single latent component, with specific Amplitude Modulation (AM) and Frequency Modulation (FM).

We propose to expand the LSA framework to allow for a complex signal representation as a linear superposition of multiple latent components, such that each component has its corresponding AM and FM. This Amplitude Modulation–Frequency Modulation (AM–FM) model allows for each component in the superposition to have its own IF and IA. Note that other AM–FM models have also been considered that allow for a signal representation using multiple IFs [36–51]. However, these AM–FM models rely on a rigid narrowband assumption that limits their applicability. The proposed AM–FM model, that we call Hilbert Spectral Analysis (HSA), allows for both narrowband as well as wideband signal components. The general problem solved by HSA is to find a representation of a complex signal $z(t)$ in terms of multicomponent latent signal components. Specifically, instead of estimating a single IA/IF pair for $z(t)$, HSA looks for sets of IA/IF pairs, each associated with corresponding latent

signal components. Although time-frequency analysis has been extensively studied, the use of a generalized AM–FM model for this analysis, without the harmonic correspondence constraint, has never been proposed, to the best of our knowledge. Use of this model leads to non-unique signal expansions following Theorem 1 in Chapter 6. However, by imposing particular assumptions on the form of the AM–FM components, a unique parameterization in terms of IA and IF can be obtained for each component. This model enables us to analyze signals with very few restrictions, resulting in many signal model interpretations and possibly more useful expansions, especially for non-stationary signals.

7.2 The AM–FM Model

7.2.1 Definitions and Assumptions

Our goal in HSA is to expand a signal into complex AM–FM components, such that the estimated IA and IF of each component are matched to some criteria. The criteria could be based on some physical or perceptual signal properties or can depend on some prior knowledge on the underlying signal model. The HSA problem is illustrated in Figure 7.1 as an extension of the LSA problem shown in Figure 6.1. The figure shows models of the latent signal that, in general, are composed of a set of latent AM-FM components.

We propose to define the AM–FM model for a complex signal $z(t)$ as a superposition of $K \in \mathbb{Z}^+$ (possibly infinite) AM–FM components

$$z(t) \equiv \sum_{k=0}^{K-1} \psi_k \left(t; a_k(t), \omega_k(t), \phi_k\right) \tag{7.1}$$

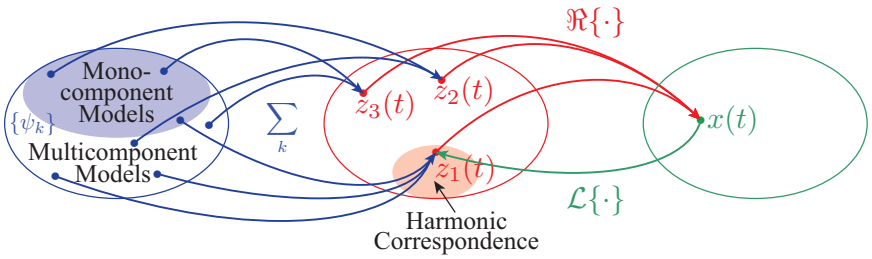


Figure 7.1: A set diagram for the HSA problem. In the LSA problem, many latent signals $z(t)$ map to the same observed signal $x(t)$ under the real operator. In the HSA problem, many component sets $\{\psi_k(t)\}$, $k = 0, 1, \ldots, K$ are superimposed to form the same latent signal $z(t)$. The goal in HSA is to properly choose $\{\psi_k(t)\}$ given $x(t)$.

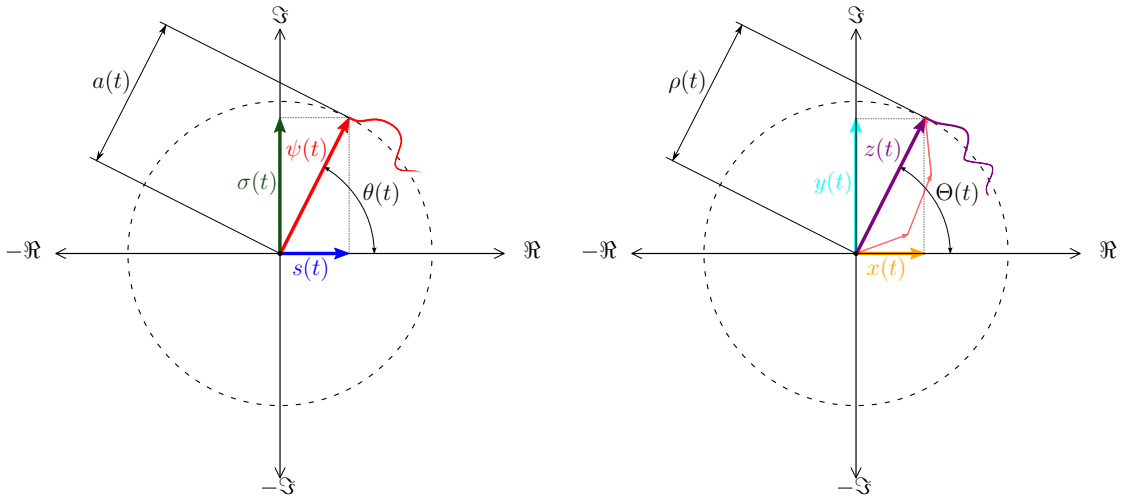
where the kth AM–FM component is defined as

$$\psi_k(t; a_k(t), \omega_k(t), \phi_k) \equiv a_k(t) \exp\left\{ j \left[\int_{-\infty}^{t} \omega_k(\tau) d\tau + \phi_k \right] \right\} \tag{7.2a}$$

$$= a_k(t) e^{j\theta_k(t)} \tag{7.2b}$$

$$= s_k(t) + j\sigma_k(t). \tag{7.2c}$$

The kth AM–FM component is parameterized by the IA $a_k(t)$, IF $\omega_k(t)$, and phase reference ϕ_k. We assume that the observed real signal $x(t)$ is related to the latent signal $z(t)$ according to (6.3). We note that most of the paradoxes related to IF are due to the various interpretations of $a_k(t)$, $\omega_k(t)$, and $\theta_k(t)$, $k = 0, 1, \ldots, K$ in the multicomponent case, and $\rho(t)$, $\Omega(t)$, and $\Theta(t)$ in (6.2) for the monocomponent case. This is because, for the monocomponent case, $K = 1$, the corresponding variables are equivalent (i.e., $a_0(t) = \rho(t)$, $\omega_0(t) = \Omega(t)$, and $\theta_0(t) = \Theta(t)$). The geometric interpretation of a single AM–FM component $\psi_k(t)$ in (7.2) is illustrated with the Argand diagram in Figure 7.2(a). The AM–FM component can be visually interpreted as a single rotating vector in the complex plane with time-varying length and time-varying angular velocity. The corresponding diagram for a multicomponent signal is shown in Figure 7.2(b). Note that, in Figure 7.2(a) and the remaining of the section, we drop the subscript k denoting the kth AM–FM component in (7.2) for notational simplicity.



(a) (b)

Figure 7.2: (a) Argand diagram of an AM–FM signal component $\psi(t)$ in (7.2) at some time instant. By interpreting $\psi(t)$ (—) as a vector, then the length of the vector is the signal component's IA $a(t)$, the vector's angular position is the signal component's phase function $\theta(t)$, and the vector's angular velocity is the signal component's IF $\omega(t)$; the phase reference ϕ is interpreted as an initial condition. The real part of the signal component $s(t)$ (—) and the imaginary part of the signal component $\sigma(t)$ (—) are interpreted as orthogonal projections of the vector $\psi(t)$. We have included an example path (—) taken by $\psi(t)$. (b) Argand diagram of the latent signal $z(t)$ in (6.2) at some time instant, from Figure 6.2, updated to show the latent signal composed of a superposition of three signal components shown in red. By interpreting the latent signal $z(t)$ (—) as a vector, then the length of the vector is the latent signal's IA $\rho(t)$ and the vector's angular position is the latent signal's phase function $\Theta(t)$. The real part of the latent signal $x(t)$ (—) and the imaginary part of the latent signal $y(t)$ (—) are interpreted as orthogonal projections of vector $z(t)$. We have included an example path (—) taken by $z(t)$.

In [190], Carson expressed the phase function $\theta(t)$ in terms of a carrier frequency ω_c and FM message $m(t)$ as

$$\theta(t) = \omega_c t + \lambda \int_0^t m(\tau) d\tau \tag{7.3}$$

assuming the modulation index satisfies $\lambda \leq \omega_c$ and $|m(t)| \leq 1$. We parameterize the phase function as

$$\theta(t) = \omega_c t + \int_{-\infty}^t m(\tau) d\tau + \phi \tag{7.4}$$

in order to avoid normalizing $m(t)$ and imposing an upper bound on the IF. The phase function can also be expressed in terms of ω_c and the Phase Modulation (PM) message $M(t)$ as

$$\theta(t) \;=\; \omega_c t + M(t) + \phi \tag{7.5}$$

where the relationship between the messages is given by

$$M(t) = \int_{-\infty}^{t} m(\tau) d\tau. \tag{7.6}$$

In the AM–FM model, we assume non-negative IF, thus

$$\omega(t) = \frac{d}{dt}\theta(t) = \omega_c + m(t) \geq 0, \;\forall\; t. \tag{7.7}$$

In Section 7.5, we discuss situations in which this assumption may be relaxed. Finally, we assume, without loss of generality, that the phase and carrier references are selected at $t = 0$. Specifically, we assume

$$\int_{-\infty}^{0} m(t) dt = M(0) = 0. \tag{7.8}$$

Thus, using (7.5),

$$\phi = \theta(0) - M(0) = \theta(0) \tag{7.9}$$

and using (7.7),

$$\omega_c = \omega(0) - m(0). \tag{7.10}$$

7.2.2 Monocomponents and Narrowband Components

The concept of a monocomponent signal (composed of one component) or multicomponent signal (composed of multiple components) and whether those components are narrowband or wideband has caused a lot of controversy in existing literature.

Although no agreed upon definition exists for a monocomponent signal [70], a few definitions have been published [3, 67, 210]. Some authors describe a monocomponent signal as a signal with a single 'ridge' in time and with harmonic frequency, corresponding to an elongated region of energy concentration [3, 189, 210]. Cohen agrees that if a signal component is well localized in frequency, the crest of the ridge corresponds to the IF and further states that the width of the ridge depends on the energy spread or instantaneous bandwidth of the component [3, 189]. Based on this definition for a monocomponent signal, a multicomponent signal may then be defined as any signal which is the sum of two or more monocomponents signals which can only occur if the separation between the ridges is large in comparison to the bandwidth of the individual components [3, 189, 210, 211]. It has also been noted that a signal may be monocomponent at some time instances and multicomponent at other time instances [189].

Taking this definition of a monocomponent signal into consideration, we point out that the definitions and concepts of IA and IF of a complex AM–FM component as defined by (7.2) are by no means justified only for narrowband signal components. Also, as pointed out by Cohen [3], a multicomponent signal is not defined by the ability to express a signal as a sum of parts, because there are an infinite number of ways to write a signal as a sum of parts, such as signal expansions into basis functions. Rather, it is the properties of the signal parts or components in relation to themselves and the signal which determines whether the decomposition is of interest [212]. Describing a signal component as inherently narrowband is unnecessarily restrictive. As seen in (7.2), the IA and IF for a complex AM–FM component are well-defined without imposing narrowband signal constraints.

Specifically, it is possible to represent a wideband signal component using the AM–FM component in (7.2), without following the restrictive narrowband definitions

of Carson [190], Ville [191], Boashash [34, 210], and Cohen [3]. Such a wideband component could consist of multiple ridges, each of which could be decomposed into narrowband components. However, for many problems, the wideband AM–FM signal component in (7.2) can better match the physical characteristics of the system that generated the signal than the multiple, narrowband signal components. Further, the appearance of structure in the Fourier spectrum can be viewed as an indication that a wideband component is present in the signal. For narrowband components, the Fourier and Hilbert spectra can be quite similar, but for wideband components, these can be quite different.

7.3 Hilbert Spectral Analysis

Huang's original definition of the Hilbert spectrum uses Empirical Mode Decomposition (EMD) to determine a set of Intrinsic Mode Functions (IMFs) which are individually demodulated with the Hilbert transform to obtain the IA $\{a_k(t)\}$ and IF $\{\omega_k(t)\}$ [70]. This analysis is also known as the Hilbert-Huang Transform (HHT) [70]. This definition can be generalized by recognizing that IMFs are a class (special case) of AM–FM components, and that other decomposition and demodulation methods exist for obtaining the instantaneous parameters $\{a_k(t)\}$ and $\{\omega_k(t)\}$.

We define the Hilbert spectrum as a representation or characterization of a signal by an *instantaneous spectrum* which is parameterized as a superposition of AM–FM components as in (7.2). With this definition, the Hilbert spectrum is not unique. In order to provide a solution that is matched to a particular application, we need to address the following problems:

P1: We need to identify and interpret the physical properties of the system that corresponds to the observed signal; using this as prior knowledge, we need to determine the appropriate complex extension to obtain a matched AM–FM

model.

P2: Once the appropriate AM–FM model is obtained, we need to estimate the IA and IF parameters $\{a_k(t)\}$ and $\{\omega_k(t)\}$, $0 \leq k \leq K-1$.

As the initial observation is real, we first extend the real signal $x(t)$ to the latent complex signal $z(t)$; the instantaneous parameters are defined on the complex signal. There are cases where the complex extension is straightforward; examples include simple harmonic motion, Fourier analysis, or a communication system where assumptions can be matched at the transmitter and receiver. However, in general, signal decomposition and component demodulation require us to identify the matched system assumptions in order to arrive at an appropriately matched solution.

The class of the solution obtained is dependent on the nature of the assumptions. Figure 7.3 illustrates the various classes of components under particular assumptions including the Simple Harmonic Component (SHC), AM, FM, and AM–FM components as well as IMFs (discussed in more detail in Chapter 8). The AM–FM component in (7.2) may be considered a generalization of all of these components. Next, we illustrate several forms of the AM–FM component and AM–FM model under various assumptions.

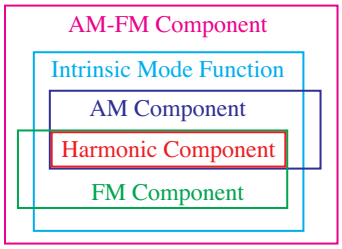


Figure 7.3: The AM–FM component in (7.2) may be considered a generalization of other well-known components. The harmonic component may be considered a special case of the AM component and FM component. Huang's IMF is a special case of the AM–FM component.

7.3.1 Two Conventional Ways to Relate a Real Observation to a Latent Signal

As noted by Gabor in [35], there are two conventional ways to relate a real signal $x(t)$ to a complex signal $z(t)$. The first convention is

$$x(t) = \Re\{z(t)\} \tag{7.11a}$$

$$= \Re\{x(t) + jy(t)\} \tag{7.11b}$$

which can be thought of as a single vector as in Figure 7.2. The second convention is

$$x(t) = [z(t) + z^*(t)]\,/2 \tag{7.12a}$$

$$= [x(t) + jy(t) + x(t) - jy(t)]/2 \tag{7.12b}$$

which can be thought of as two vectors. If $z(t)$ is a superposition of AM–FM components, the first convention can be viewed as a *single* rotating vector per component, while the second convention can be viewed as a *pair* of vectors per component rotating in opposite directions.

Equation (7.12) is implicit in standard Fourier analysis, with the interpretation that the latent signal is real-valued and its spectrum[1] consists of Hermitian symmetric, positive and negative frequency components. The convention in (7.11) is adopted for the AM–FM model in (7.1), with the interpretation that only the real part of the signal, $x(t)$, is observed. Inherent in both of these conventions is the ambiguity associated with the imaginary part $y(t)$, i.e., an infinite number of choices exist for $y(t)$ that lead to the same $x(t)$. Thus, we view the imaginary part $y(t)$ as a *free parameter* that can be *chosen* to better match different system observations and their properties.

[1] In this work, "spectrum" strictly means Fourier spectrum and not "Hilbert spectrum."

Using the vector characterization of a latent signal, the imaginary part $y(t)$ is most often chosen offset $\pi/2$ radians relative to the real signal $x(t)$. More specifically, if we assume a single SHC, $z(t) = \psi_0(t; a_0, \omega_0, \phi_0)$, then the AM–FM model is given by $z(t) = a_0 e^{j(\omega_0 t + \phi_0)}$ (see (6.13)). Although choosing the imaginary part given the real signal is trivial for the SHC, the same is not true for the component in (7.2) where the IA $a_k(t)$ and IF $\omega_k(t)$ are not constant; the IA and IF are (possibly) time-varying functions that are specified by the given application.

7.3.2 Simple Harmonic Component

The simplest form of the AM–FM model has one component ($K = 1$) and corresponding parameters $a_0(t) = a_0$ and $\omega_0(t) = \omega_0$ with $a_k, \omega_0 \in \mathbb{R}$; thus the AM–FM component in (7.2) becomes

$$\psi_0(t; a_0, \omega_0, \phi_0) = a_0 e^{j(\omega_0 t + \phi_0)}. \tag{7.13}$$

We recognize the significance of this form as representing simple harmonic motion, as we discussed in Chapter 6 following Vakman's Constraint 3.

7.3.3 Superposition of Simple Harmonic Components

One of the most common forms of the AM–FM model is the form with an infinite number of components ($K = \infty$) with parameters $a_k(t) = a_k$, and $\omega_k(t) = k\omega_0$

$$z(t) = \sum_{k=0}^{\infty} a_k e^{j(k\omega_0 t + \phi_k)}. \tag{7.14}$$

This model represents the latent signal that has real observation (also shown in (6.3))

$$x(t) = \Re\{z(t)\} \tag{7.15a}$$

$$= \sum_{k=0}^{\infty} a_k \cos(k\omega_0 t + \phi_k) \tag{7.15b}$$

$$= A_0 + \sum_{k=1}^{\infty} [A_k \cos(k\omega_0 t) + B_k \sin(k\omega_0 t)]. \tag{7.15c}$$

Equation (7.15) is recognized as a Fourier Series (FS) of $x(t)$ with the convention used for the complex extension given in (7.12) [213, 214]. Note that when $phi_k = 0$, this corresponds to the cosine series for even signals $x(t)$. The FS can be considered the simplest of AM–FM superposition models when SHCs are assumed.

The FT is the limiting form of the FS as the fundamental frequency tends to zero, $\omega_0 \to 0$ [215]. When viewing the FT as a special case of the AM–FM model, subtle and important observations can be made as follows:

- The AM–FM model corresponding to a separable solution,i.e., product of two functions each of which depend on only one variable, with $X(\omega)$ (not time-varying) and $e^{j\omega t}$ (constant frequency), *is* the FT, i.e., the inverse FT is a superposition (inner product) of $X(\omega)$ and $e^{j\omega t}$.

- The FT yields a solution that can be described as a superposition of non-time-varying components described by constant state parameters, $a_k(t) = a_k$, $\omega_k(t) = k\omega_0$, and ϕ_k which corresponds to a constant model state. However, this does not, in any practical way, constrain the real superposition $x(t)$. Although analysis of a time-varying system can be accomplished using a constant model, like the FT, in most cases this forces $K = \infty$, which may not accurately describe a physical system.

- HSA can be considered as a generalization of Fourier analysis [70]. However, rather than a generalization of Fourier analysis obtained by generalizing a kernel function (Gabor transform, Wigner-Ville distribution, wavelet transform, etc.) [3, 29, 216–220], an AM–FM component can be viewed as allowing for *additional* degrees of freedom in the basis functions of Fourier analysis. Although it may be convenient to think of (7.2) as a basis for HSA because it spans the entire Hilbert space, the AM–FM component does not meet the linear independence property of a basis representation [221].

7.3.4 The AM Component

One way to generalize the SHC in (7.13) is to relax the constraint of constant amplitude which leads to an AM component

$$\psi_0(t; a_0(t), \omega_0, \phi_0) \quad = \quad a_0(t) e^{j(\omega_0 t + \phi_0)}. \tag{7.16}$$

The simple AM component was originally developed in the context of communication theory [222–224].

7.3.5 Superposition of AM Components

Another form of the AM–FM model is a superposition of AM components that have frequencies that are integer multiples of a fundamental frequency ω_0,

$$z(t) = \sum_{k=0}^{\infty} a_k(t) e^{j(k\omega_0 t + \phi_k)} \tag{7.17}$$

with real observation given by (6.3).

The AM superposition was first conceived by Gabor, and it leads to the develop-

ment of the Short-Time Fourier Transform (STFT) [35]

$$X(t,\omega) = \int x(\tau)h(\tau - t)e^{-j\omega\tau}d\tau \tag{7.18a}$$

$$= \frac{e^{-j\omega t}}{2\pi}\int X(\omega)H(\omega - \omega)e^{j\omega t}d\omega. \tag{7.18b}$$

where $h(t)$ is a window function with FT $H(\omega)$, and $x(t)$ and $X(\omega)$ are a FT pair. There are two possible interpretations of the STFT in the context of the AM–FM model:

1. The window function $h(t)$ is multiplied by the signal $x(t)$ in order to use the FT, i.e., simple harmonic analysis at each time t. This interpretation couples $H(\omega)$ to $X(\omega)$.

2. The window function $h(t)$ is multiplied by $e^{-j\omega t}$ and is effectively an AM superposition analysis where the window is an assumption on the model and effectively defines the IAs, $a_k(t)$, in the AM–FM model. This interpretation couples $H(\omega)$ to $e^{j\omega t}$.

The second interpretation provides an important and alternate interpretation of the STFT. What is typically viewed as imprecision in the ability to compute time-*frequency* parameters in the first interpretation can alternatively be viewed in the second interpretation as a precise ability to compute time-*component* parameters using the modified basis, $h(t-\tau)e^{-j\omega t}$.

7.3.6 FM Component

Another way to generalize the SHC in (7.13) is to relax the constraint of constant frequency which leads to a simple FM component

$$\psi_0(t; a_0, \omega_0(t), \phi_0) \quad = \quad a_0 e^{j\left[\int_{-\infty}^{t}\omega_0(\tau)d\tau + \phi_0\right]}. \tag{7.19}$$

The FM component was also developed in the context of communication theory [190, 192]. The FM component has been used in signal synthesis where Chowning observed that a single, wideband FM component is perceived by the ear as spectrally rich, i.e., multiple SHCs [225]. This FM-synthesis method was used in commercial audio synthesizers [226, 227].

7.3.7 Superposition of FM Components

Signal analysis using a superposition of FM components is not usually considered due to the loss of linear independence of the components in the model and the restriction of constant amplitude. For this reason, signal analysis using a superposition of AM–FM components is more useful than a superposition of FM components.

7.3.8 Other AM–FM Models

Alternatives to the proposed AM–FM model in Section 7.2 have been considered. However, these alternative models have some restrictive assumptions which limit their usefulness, as discussed next. Previous AM–FM models for signal analysis/synthesis usually fall into one of three main groups: 1) Hilbert transform [36–39], 2) peak tracking/sinusoidal modeling [40–43], and 3) Teager energy operator [44–49]. However, some models exist that do not fall into any of these groups [46, 50]. A historical summary of AM–FM modeling is presented by Gianfelici [51]. A review of algorithms for estimating IF is presented by Boashash [228, 229].

The Hilbert transform permits the unique definition of IA, IF, and phase of any real signal (random or deterministic) [62]. Thus, it provides a direct means of performing AM–FM analysis, however, it requires that the imaginary part be defined in a consistent manner in all cases. Implicit in the Hilbert transform is harmonic correspondence. As a result, the more likely this assumption is true, the better the

Hilbert-transformed signal approximates the true imaginary part and the more likely the Hilbert transform based solution provides an accurate model of the underlying physical synthesis model [210]. Direct application of the Hilbert transform does not, however, address the signal decomposition problem. As a result, methods for using the Hilbert transform for decomposition were proposed [38, 39, 230].

In peak tracking, one accepts the narrowband definition of a component and as a result, each component appears in the time-frequency plane as a single ridge of energy concentration. Thus, a signal can be parameterized by tracking its ridges in location, intensity, and possibly bandwidth. The time-frequency distribution is usually derived from a STFT [40], but a generalized time-frequency distribution can also be used [41]. Regardless of the time-frequency distribution used, the narrowband assumption is inherent. Extensions of the sinusoidal model were proposed, such as the harmonic plus noise model and the adaptive quasi-harmonic model [42].

The Teager Energy Separation Algorithm (ESA) provides a method of estimating the IA and IF of an AM–FM component, assuming harmonic correspondence. The Teager Energy Operator (TEO) is defined as [28]

$$\Psi\left\{x(t)\right\} \equiv \left[\dot{x}(t)\right]^2 - x(t)\ddot{x}(t) \tag{7.20}$$

where $\dot{x}(t)$ and $\ddot{x}(t)$ are the first and second time derivatives of $x(t)$, respectively. The TEO applied to a single narrowband AM–FM component in (7.2) results in [28]

$$\Psi\left\{\psi_0(t)\right\} \approx \left[a_0(t)\omega_0(t)\right]^2. \tag{7.21}$$

Assuming the modulating signals $a_0(t)$ and $m_0(t)$ are bandlimited, it can be shown that

$$a_0(t) \approx \frac{\Psi\left\{\psi_0(t)\right\}}{\sqrt{\Psi\left\{\dot{\psi}_0(t)\right\}}} \tag{7.22}$$

and

$$\omega_0(t) \approx \sqrt{\frac{\Psi\left\{\dot{\psi}_0(t)\right\}}{\Psi\left\{\psi_0(t)\right\}}} \tag{7.23}$$

thus providing a method to estimate narrowband monocomponent IA/IF parameters [28, 231–233].

The AM–FM model in these groups *all* rely on a rigid narrowband component. Surprisingly, the use of wideband components is a well-known means of audio synthesis [225, 227, 234, 235]. However, *analysis* using the wideband components in the general form givin in (7.2), without the constraints such as harmonic correspondence has not been considered.

7.4 Frequency Domain View of Latent Signal Analysis

In Chapter 6, we introduced the LSA problem in the time domain. The frequency domain view of the LSA problem is to estimate the FT of $Z(\omega)$ of the latent signal $z(t)$ from the FT of $X(\omega)$ of the real observation $x(t)$. We call the FT of $Z(\omega)$, the latent spectrum. As $x(t) = \Re\{z(t)\}$, this imposes structure on the spectrum of $x(t)$ using the Hermitian symmetry FT property, $X(\omega) = [Z(\omega) + Z^*(-\omega)]/2$. As discussed in Chapter 6, the conventional way to estimate $z(t)$ from $x(t)$ is by using the Hilbert transform as in (6.8). Equivalently, in the frequency domain, the conventional way to estimate $Z(\omega)$ from $X(\omega)$ is by using Gabor's method. Using Gabor's method effectively forces the FT of (6.7) to become $\hat{Y}(\omega) = -j\text{sgn}(\omega)X(\omega)$, where $\text{sgn}(\cdot)$ is the sign function and the resulting estimate of the latent spectrum $\hat{Z}(\omega)$ is always spectrally one-sided [43]. In the frequency domain, by relaxing harmonic correspondence, we gain the freedom to choose $\hat{Y}(\omega)$ that is appropriately matched to the signal properties of the system under study.

As an example of this frequency domain view, consider the observed spectrum

$$X(\omega) = a_0\pi[\delta(\omega + \omega_0) + \delta(\omega - \omega_0)] \tag{7.24}$$

where $\delta(\cdot)$ is the Dirac delta function. If we assume harmonic correspondence,

$$\hat{Y}(\omega) = -ja_0\pi[-\delta(\omega + \omega_0) + \delta(\omega - \omega_0)] \tag{7.25}$$

and the corresponding latent spectrum is given by

$$\begin{aligned} \hat{Z}(\omega) &= X(\omega) + j\hat{Y}(\omega) \\ &= 2a_0\pi\delta(\omega - \omega_0). \end{aligned} \tag{7.26}$$

If we do not assume harmonic correspondence, there are many choices for $\hat{Y}(\omega)$. Three such choices for $\hat{Y}(\omega)$ lead to three FTs of latent signals

$$\hat{Z}_1(\omega) = 2a_0\pi\delta(\omega + \omega_0), \tag{7.27}$$

$$\hat{Z}_2(\omega) = a_0\pi[(1 - \alpha)\delta(\omega + \omega_0) + (1 + \alpha)\delta(\omega - \omega_0)], \tag{7.28}$$

and

$$\hat{Z}_3(\omega) = a_0\pi[\delta(\omega + \omega_0) + \delta(\omega - \omega_0) + \alpha\delta(\omega - \beta\omega_0) - \alpha\delta(\omega + \beta\omega_0)] \tag{7.29}$$

where $\alpha, \beta \in \mathbb{R}$ and (7.29) corresponds to the FT of (6.21). Because of the structure imposed on the spectrum $X(\omega) = [Z(\omega) + Z^*(-\omega)]/2$, the latent spectra in (7.26)-(7.29) all yield the same $X(\omega)$.

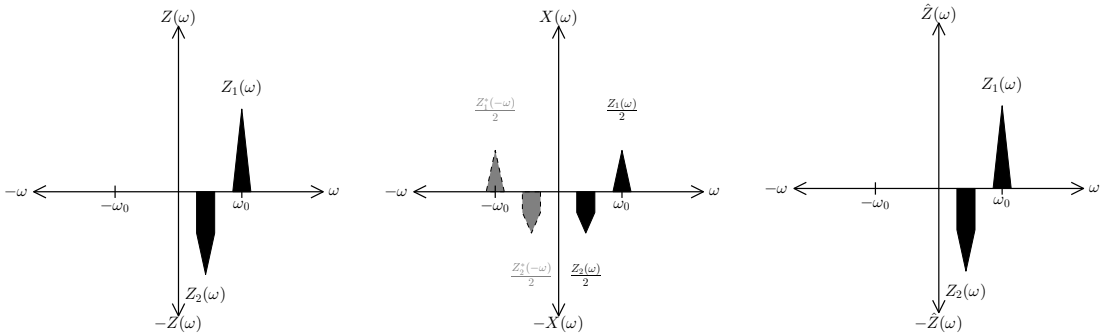
In Table 7.1, we summarize the spectral structure of a complex signal depending on whether or not harmonic correspondence is assumed. If $z(t)$ assumes only real values, then the spectrum has Hermitian symmetry (first row of Table 7.1). If $z(t)$ assumes complex values and has harmonic correspondence, then the Fourier spectrum is single-sided (second row of Table 7.1). If $z(t)$ assumes complex values and does not have harmonic correspondence, the spectrum does not have Hermitian symmetry.

Table 7.1: Structure of the Fourier spectrum under the assumption of a model composed of Simple Harmonic Components (SHCs).[2]

Class of $z(t)$	**FT Structure**
$z(t) \in \mathbb{R}$	$Z(\omega) = Z^*(-\omega)$
$z(t) \in \mathbb{C}$ with harmonic correspondence	$Z(\omega) = 0,\ \omega < 0$
$z(t) \in \mathbb{C}$ without harmonic correspondence	$Z(\omega) \neq Z^*(-\omega)$

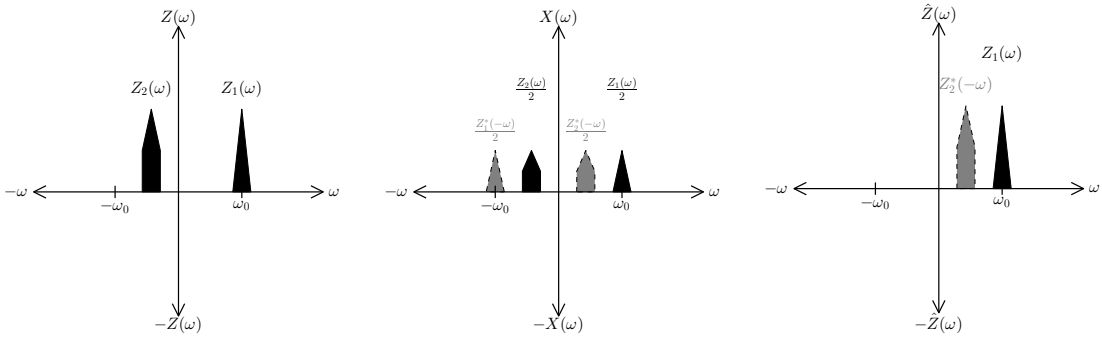
In the frequency domain view of LSA, Gabor's method can be interpreted as simply a method to distinguish $Z(\omega)$ from $Z^*(-\omega)$ which works only if the harmonic frequencies of $Z(\omega)$ are all non-negative. This concept can be generalized to the Hilbert spectrum. As an illustration, consider Figure 7.4(a) where the latent spectrum $Z(\omega) = Z_2(\omega) + Z_1(\omega)$ and the real signal under analysis has spectrum $[Z_1^*(-\omega) + Z_2^*(-\omega) + Z_2(\omega) + Z_1(\omega)]/2$. Applying Gabor's method, the complex signal has spectrum $Z(\omega) = Z_2(\omega) + Z_1(\omega)$. Next consider Figure 7.4(b) where the latent spectrum $Z(\omega) = Z_2(\omega)+Z_1(\omega)$ and the real signal under analysis has spectrum $[Z_1^*(-\omega)+Z_2(\omega)+Z_2^*(-\omega)+Z_1(\omega)]/2$. Applying Gabor's method, the complex signal has spectrum $Z_2^*(-\omega)+Z_1(\omega)$ which is the incorrect spectral grouping, i.e., conjugate-symmetric terms were confused for the terms of interest. Finally, consider Figure 7.4(c) where the latent spectrum $Z(\omega) = Z_3(\omega)+Z_2(\omega)+Z_1(\omega)$ and the real signal under analysis has spectrum $[Z_1^*(-\omega)+\cancel{Z_2^*(-\omega) + Z_3(\omega)}+\cancel{Z_3^*(-\omega) + Z_2(\omega)}+Z_1(\omega)]/2$; cancellation is due to the symmetry of the latent spectrum. Applying Gabor's method, the complex signal has spectrum $Z_1(\omega)$ and not only has Gabor's method failed to yield the latent spectrum but terms are missing entirely.

[2]For convenience, we omit the case $Z(\omega) = 0,\ \omega > 0$.



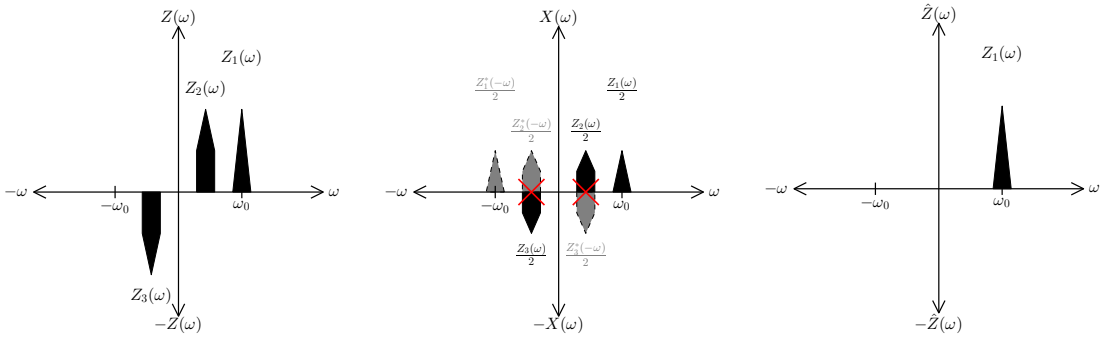
Latent Spectrum | Observed Spectrum | Gabor's AS Spectrum

(a)



Latent Spectrum | Observed Spectrum | Gabor's AS Spectrum

(b)



Latent Spectrum | Observed Spectrum | Gabor's AS Spectrum

(c)

Figure 7.4: Illustrations of when Gabor's method (a) can distinguish $Z(\omega) = Z_1(\omega) + Z_2(\omega)$ from $Z^*(-\omega) = Z_1^*(-\omega) + Z_2^*(-\omega)$ because $Z(\omega)$ has harmonic correspondence, (b) cannot distinguish $Z(\omega) = Z_1(\omega) + Z_2(\omega)$ from $Z^*(-\omega) = Z_1^*(-\omega) + Z_2^*(-\omega)$ because the latent spectrum is two-sided and Hermitian symmetry is imposed by the real observation, and (c) is incorrect because the structure of the latent spectrum along with the Hermitian symmetry imposed by the real observation has *concealed* terms.

7.5 Subtleties of the Hilbert Spectrum

In the previous section, we considered the LSA problem where we assumed SHCs but relaxed the harmonic correspondence assumption. By relaxing harmonic correspondence, we also relax the non-negative IF constraint associated with it. In this section, we discuss the implications of assuming non-negative IF and relaxing the assumption of SHCs.

First, consider making the assumption of non-negative IF and a model composed of SHCs. The second row of Table 7.1 shows that this implies harmonic correspondence on $z(t)$. Next, consider non-negative IF and a model not composed of SHCs. Contrary to rows two and three of Table 7.1 where SHCs are assumed, there can exist models with non-negative IF but without harmonic correspondence. We illustrate the existence of such models with the following example. Consider the complex signal,

$$z(t) = a_0 \cos(\omega_0 t) + j\alpha a_0 \sin(\omega_0 t) \tag{7.30}$$

and the spectrum given by (7.28) which is not one-sided. However in terms of a single AM–FM component, we can have

$$z(t) = a_0(t) e^{j\theta_0(t)} \tag{7.31}$$

where

$$a_0(t) = \sqrt{a_0^2 \cos^2(\omega_0 t) + \alpha^2 a_0^2 \sin^2(\omega_0 t)} \tag{7.32}$$

and

$$\theta_0(t) = \arctan\left[\frac{\alpha a_0 \sin(\omega_0 t)}{a_0 \cos(\omega_0 t)}\right]. \tag{7.33}$$

The IF, $\omega_0(t) = \dfrac{d}{dt}\theta_0(t)$, is strictly positive for any positive choice of α. Thus, the associated *Hilbert* spectrum is one-sided even though the spectrum is two-sided. In other words, saying that we assume positive IF is not the same as saying that we

assume a one-sided spectrum. Although signals can be designed such that $Z(\omega)$, for practical purposes, has a one-sided spectrum, e.g., communications signals, this does not mean that naturally-occurring signals have, in general, a one-sided spectrum. Therefore, making the one-sided spectrum assumption (and assuming harmonic correspondence) may lead to incorrect model parameters.

Another incorrect assumption that is often made is that if the signal is narrowband, the Hilbert transform has a meaningful complex extension. Unfortunately, this is not the case because narrowband signals exist that do not have Hermitian symmetry and hence use of the Hilbert transform yields incorrect results. This is demonstrated by (7.28) which could be considered narrowband, yet use of the Hilbert transform to extend its real observation would not yield the correct latent signal due to the absence of harmonic correspondence.

By changing from SHCs to AM–FM components and relaxing the assumption of harmonic correspondence, we have generalized the spectrum to a Hilbert spectrum and effectively changed the definition of IF for real signal to be component-dependent. By reverting back to Carson's definition and not using SHCs in the analysis, we move away from Gabor and Ville's specialized definition back towards the generalized definition of IF.

While using AM–FM components to define IF, which can change in time, we have to consider the possibility that a component's IF changes sign. Such a sign change is not possible under the assumption of SHCs which have constant IF. In this work, we have arbitrarily assumed that all components *must have non-negative IF for all time*, although there may be some signal classes for which this is not true. In these classes, the AM–FM parameters may not match the underlying signal model parameters, although the superposition will yield the correct real signal. This is illustrated in Figure 7.5, where a component of $z(t)$ with associated IF $\omega(t)$ and the component

of $z^*(t)$ with associated IF $-\omega(t)$ cannot be separated by assuming non-negative IF because each has both positive and negative instantaneous frequencies. Therefore, in these cases, we need to relax the assumption of non-negative IF at some time instances in order to properly estimate the latent signal.

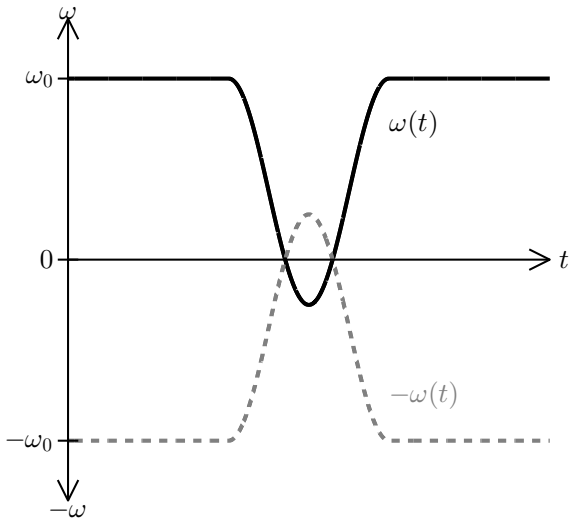


Figure 7.5: Illustration of when the assumption of non-negative IF cannot distinguish $z(t)$ with associated IF $\omega(t)$ (—) from $z^*(t)$ with associated IF $-\omega(t)$ (---) because each has both positive and negative values of IF at some time instances.

7.6 Examples of the Hilbert Spectral Analysis Problem

Similar to the example of the LSA problem presented in Section 6.6, some examples of the HSA problem are given in terms of IA/IF pairs, $\rho(t)$ and $\Omega(t)$, for latent components in the AM–FM model in (7.1). This is illustrated by the set diagram in Figure 7.1. In our closed-form solutions, we choose assumptions based on specific system observations and model interpretations. For convenience, it is assumed that the phase reference $\phi_k = 0$, where possible.

7.6.1 Periodic Triangle Waveform Example

As an example of HSA, we consider the triangle waveform formulated in (6.26) and a sinusoidal FM signal.

7.6.1.1 Simple harmonic components

If we assume the component in (7.2) has the SHC form given in (7.13), then the AM–FM model can be obtained using Fourier analysis. The triangle waveform in Section 6.6 can then be expressed as an infinite number of components

$$x(t) = \Re\left\{\sum_{k=0}^{\infty} \frac{8A}{\pi^2}\frac{1}{(2k+1)^2}e^{j(2k+1)\omega_0 t}\right\} \tag{7.34}$$

where the IA is given by

$$a_k(t) = \frac{8A}{\pi^2}\frac{1}{(2k+1)^2}, \tag{7.35}$$

the IF is given by

$$\omega_k(t) = (2k+1)\omega_0, \tag{7.36}$$

and the fundamental frequency ω_0 is constant.

7.6.1.2 Single AM–FM component with harmonic correspondence

The solution in (7.34) can be given as a single, wideband AM–FM component ($K = 1$), in terms of the AM–FM model as

$$x(t) = \Re\left\{a_0(t)e^{j[\omega_0 t + M_0(t)]}\right\} \tag{7.37}$$

where the IA $a_0(t)$ is given by $\hat{\rho}(t)$ in (6.28) and the IF $\omega(t)$ is given by $\hat{\Omega}(t)$ in (6.29).

7.6.1.3 Single AM component

If we assume a single component ($K = 1$) with constant frequency $\omega_0(t) = \omega_0$, the triangle waveform can be expressed in terms of the AM–FM model as

$$x(t) = \Re\left\{a_0(t)e^{j\omega_0 t}\right\} \tag{7.38}$$

where the IA $a_0(t)$ is given by $\hat{\rho}(t)$ in (6.32).

7.6.1.4 Single FM component

If we assume a single component ($K = 1$) with constant amplitude $a_0(t) = A$, the triangle waveform can be expressed in terms of the AM–FM model as

$$x(t) = \Re\left\{Ae^{j[\omega_0 t + M_0(t)]}\right\} \tag{7.39}$$

where the IF $\omega_0(t)$ is given by $\hat{\Omega}(t)$ in (6.34).

7.6.2 Sinusoidal FM Example

A sinusoidal FM signal with carrier frequency ω_c, modulation frequency ω_m, and constant $B \in \mathbb{R}$, is expressed as

$$x(t) = \Re\left\{e^{j[\omega_c t + B\sin(\omega_m t)]}\right\}. \tag{7.40}$$

7.6.2.1 Single FM component

Assuming a single component ($K = 1$), (7.40) is already in the form of the AM–FM model, with IA

$$a_0(t) = 1. \tag{7.41}$$

This leads to the IF

$$\omega_0(t) = \omega_c + \frac{d}{dt}B\sin(\omega_m t). \tag{7.42}$$

7.6.2.2 Simple harmonic components

The sinusoidal FM signal can be expressed in terms of the AM–FM model as

$$x(t) = \Re\left\{ \sum_{k=-\infty}^{\infty} J_k(2\pi B/\omega_m) e^{j[(\omega_c + k\omega_m)t + \phi_k]} \right\} \tag{7.43}$$

where $J_k(\cdot)$ denotes the kth-order Bessel function of the first kind [199]. This yields an infinite number of components with IA given by

$$a_k(t) = J_k(2\pi B/\omega_m) \tag{7.44}$$

and IF given by

$$\omega_k(t) = \omega_c + k\omega_m. \tag{7.45}$$

This is an example of a signal class where the assumption of non-negative IF components in (7.7) is relaxed, as the summation on k is double-sided. For most practical choices of B and ω_m, the term $J_k(2\pi B/\omega_m)$ associated with the negative frequencies becomes vanishingly small, therefore the assumption of non-negative IF is, for practical purposes, true.

7.6.3 Remarks

As demonstrated in the aforementioned examples, we obtain different parameterizations for the same signal based on the physical assumptions made during analysis. Some parameterizations, such as those obtained using Fourier analysis and the Hilbert transform, are closely related as they assume to be observed from similar physical systems. However, there exist many different types of parameterizations, based on different interpretations. As a result, without other information we cannot select one particular parameterization over another one. This highlights the importance of considering the hidden assumptions when using any particular signal representation and

interpreting meaning from the parameters. If assumptions in analysis do not match the underlying signal model, erroneous interpretations may be made.

The previous solutions make use of different imaginary parts, each corresponding to different signal models. This demonstrates the non-uniqueness of the imaginary part for the AM–FM component, and hence our view of the imaginary signal as a free parameter. Also, note that the solutions in (7.34), (7.37), (7.38), etc., each utilize a complex extension that is implied by the particular assumptions made, rather than utilizing an extension based on a single rigorously-defined procedure, like the Hilbert transform. In fact, for any practically-chosen real signals $x(t)$ and $y(t)$, there exist assumptions leading to instantaneous parameters in which $y(t)$ can be viewed as the imaginary part corresponding to $x(t)$.

The AM–FM model leads to exact solutions for the instantaneous parameters $a(t)$ and $\omega(t)$, so it might appear to violate the uncertainty principle, i.e., exceed the lower limit of the time-bandwidth product. However, when AM–FM modeling is viewed as a quantum mechanics problem, our casting of the problem is fundamentally different than Gabor's. A comparison of the formal correspondences between quantum mechanics and Fourier analysis, is given in Table 7.2. A comparison of the formal correspondences between quantum mechanics and HSA, is given in Table 7.3. In Gabor's casting, *time and frequency* are the complementary variables, i.e., (position, momentum) are complementary variables to $(t,\, \omega)$ [3] while for the AM–FM model, the *real and imaginary parts* are the complementary variables, i.e., (position, momentum) are complementary variables to $(x(t),\, y(t))$. Therefore, all uncertainty arises because only the real signal is observed. We believe that our casting is not only more appropriate, but also more useful and powerful than Gabor's casting.

Table 7.2: Formal correspondences between Fourier analysis and quantum mechanics (p. 197 in [3]).

Fourier Analysis			**Quantum Mechanics**	
Description	**Symbol**		**Description**	**Symbol**
Time	t		Position	q
Frequency	ω		Momentum	p
no correspondence		$\Leftrightarrow$	Time	t
Signal	$x(t)$		Wave Function	$\Psi(q,t)$
Uncertainty	$BT \geq \frac{1}{2}$		Uncertainty	$\sigma_p \sigma_q \geq \frac{1}{2}\hbar$

Table 7.3: Formal correspondences between HSA and quantum mechanics.

Hilbert Spectral Analysis		**Quantum Mechanics**	
Description	**Symbol**	**Description**	**Symbol**
Real Signal	$x(t)$	Position	q
Imaginary Signal	$y(t)$	Momentum	p
Time	t	Time	t
Latent Signal	$z(t)$	Wave Function	$\Psi(q,t)$
Uncertainty	$y(t)$ not observed	Uncertainty	$\sigma_p \sigma_q \geq \frac{1}{2}\hbar$
Instantaneous Frequency	$\omega_k(t) = \frac{d}{dt} \arg\{\psi_k(t)\}$		

7.7 Visualization of the Hilbert Spectrum

The ability to visualize and interpret model parameters is key to the adoption of any analysis method. Often complex AM–FM signals are plotted as a series of real two-dimensional (2-D) plots, i.e., $s_k(t)$ versus t, which for AM–FM components, would provide little insight into the underlying signal model. Alternatively, Argand diagrams may be used for AM–FM signal visualization, however, drawbacks include

the imaginary part possibly having no assigned meaning and the revolution rate not intuitively displayed.

It would be more appropriate to have a plot of the Hilbert spectrum in its entirety. Unfortunately, plots of the Hilbert spectrum are often crudely discretized because a clear distinction between instantaneous parameters and spectral parameters is not made [70]. We propose a method for visualizing the Hilbert spectrum, which is both complete, intuitive, and avoids the coarse discretization. The proposed visualization can be considered a (pseudo-) phase-space plot because every degree of freedom or parameter of each AM–FM component is represented.

By plotting $\omega_k(t)$ versus $s_k(t)$ versus time as a line in a three-dimensional (3-D) space and varying the color intensity the line with respect to $|a_k(t)|$ for each component, the simultaneous visualization of multiple parameters for each component is possible. Further, orthographic projections yield common plots: the real-time plane (the real signal waveforms), the time-frequency plane (Hilbert spectrum), and the real-frequency plane (analogous to the Fourier magnitude spectrum). We have found it beneficial to interpret each component as an "illuminated" particle moving in 3-D space. Each particle's motion in time and frequency is governed by $\psi_k(t)$. The proposed method allows one to visualize the assumed underlying signal model.

To illustrate visualization of the Hilbert spectrum, we plot the examples in Section 7.6. However, the sophisticated nature of the proposed 3-D visualization of the Hilbert spectrum is not well-accommodated with paper media, and as a result we provide associated MATLAB functions and additional visualizations online [8]. In the plots, we have utilized a perceptually-motivated colormap in order to improve interpretation over other colormaps [236, 237].

For the triangle waveform example, Figure 7.6(a), Figure 7.6(c), and Figure 7.6(e) illustrate the Hilbert spectrum plots for the assumptions of SHC, single AM–FM

component with harmonic correspondence, and single AM component, respectively. Figure 7.6(a) shows the first three SHCs (constant amplitude and constant frequency in (7.35)) at integer multiples of a fundamental frequency, 50π rads/s. Figure 7.6(c) shows the single AM–FM component with harmonic correspondence, where line color variation indicates a time-varying IA in (6.28) and clear time-varying IF in (6.29). Figure 7.6(e) shows the single AM component, where constant IF and color variation indicates time-varying IA in (6.32). Figure 7.6(b), Figure 7.6(d), and Figure 7.6(f) show the corresponding time-frequency planes of Figure 7.6(a), Figure 7.6(c), and Figure 7.6(e) by projecting out the $s_k(t)$ dimension.

For the sinusoidal FM example, Figure 7.7(a) and Figure 7.7(c) illustrate the Hilbert spectrum plots assuming a single FM component and SHCs, respectively. Figure 7.7(a) shows the constant IA in (7.41) indicated by a constant line color and time varying IF in (7.42). Figure 7.7(c) shows SHCs at integer multiples of a fundamental frequency 4π rads/s, where each component has constant IA in (7.44). Figure 7.7(b) and Figure 7.7(d) show the corresponding time-frequency planes of Figure 7.7(a) and Figure 7.7(c) by projecting out the $s_k(t)$ dimension.

7.8 Discussion

In Chapter 7, we presented the Hilbert spectrum as a generalized LSA problem. In the general problem, we seek a representation of the complex signal as $z(t)$ a superposition of latent signals, i.e., a multicomponent model consisting of a superposition of complex AM–FM components. We used the AM–FM model to parameterize the Hilbert spectrum as sets of IA/IF pairs, along with phase and frequency references. In the LSA problem, relaxation of the harmonic correspondence constraint allowed freedom in the choice of the signal's imaginary part and admitted many solutions for the latent signal. In the HSA problem, relaxation of the harmonic correspondence

constraint allows freedom in the choice of each component's imaginary part, thereby allowing even greater freedom in the construction of the signal's model. We illustrated examples of the AM–FM model by considering variations of constant amplitude and constant frequency, and time-varying amplitude and time-varying frequency, leading to known signal cases.

We discussed the frequency domain view of LSA, including the implications of relaxing harmonic correspondence. From there, we discussed the Hilbert spectrum view of LSA, specifically the implications of using a general AM–FM component with relaxed harmonic correspondence and assumed positive IF. Closed form HSA examples were provided in which we assume various forms of the AM-FM component and determined the unique corresponding instantaneous parameterization in terms of IA and IF. A novel 3-D visualization of the Hilbert spectrum was proposed by plotting $\omega(t)$ versus $s(t)$ versus time and coloring with respect to $|a(t)|$.

Finally, we discussed the analogy of the Hilbert spectrum to quantum mechanics in which our casting of time-frequency analysis is fundamentally different than Gabor's casting, where the uncertainty in our framework is in the imaginary part and not the frequency variable.

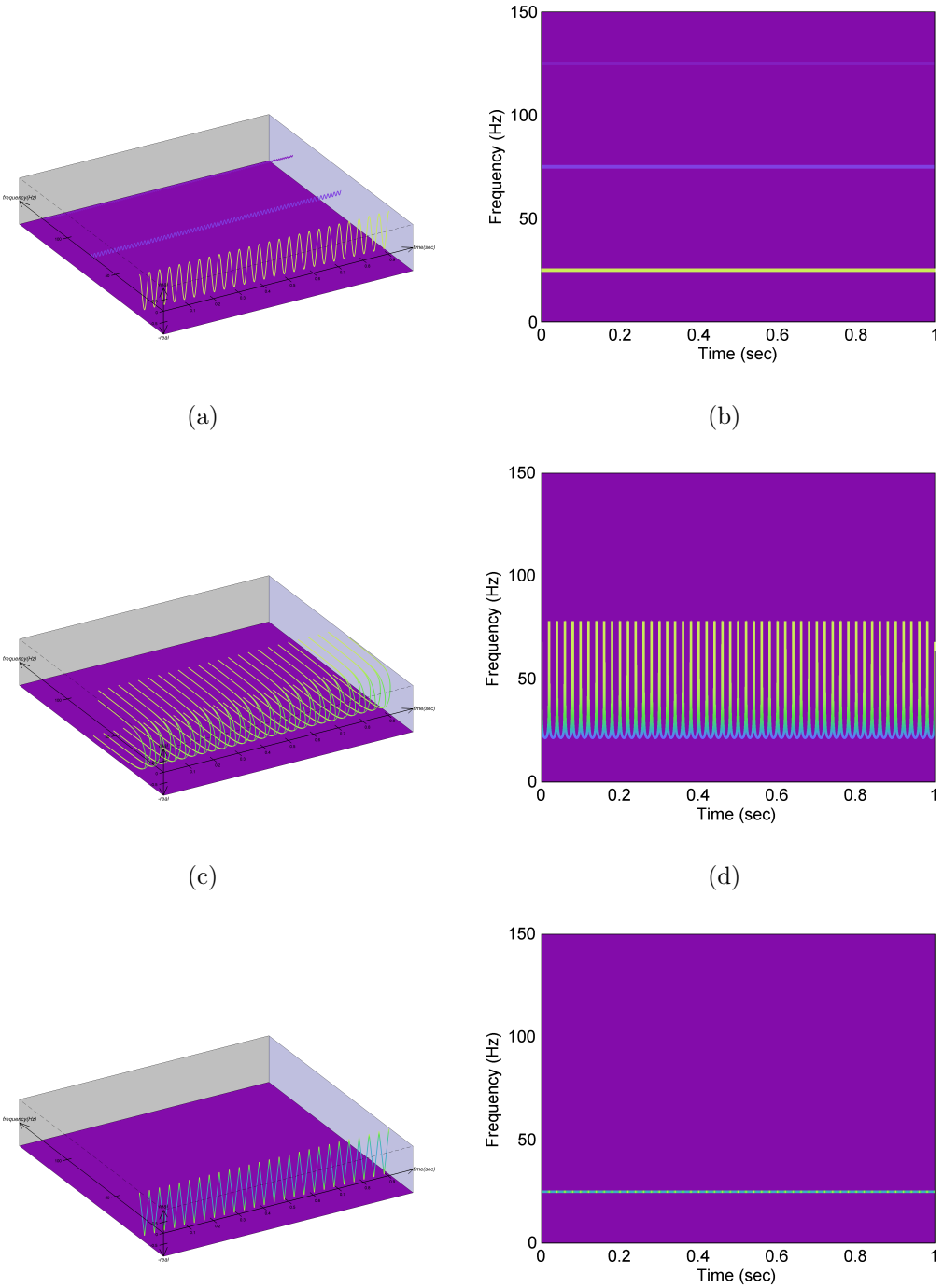
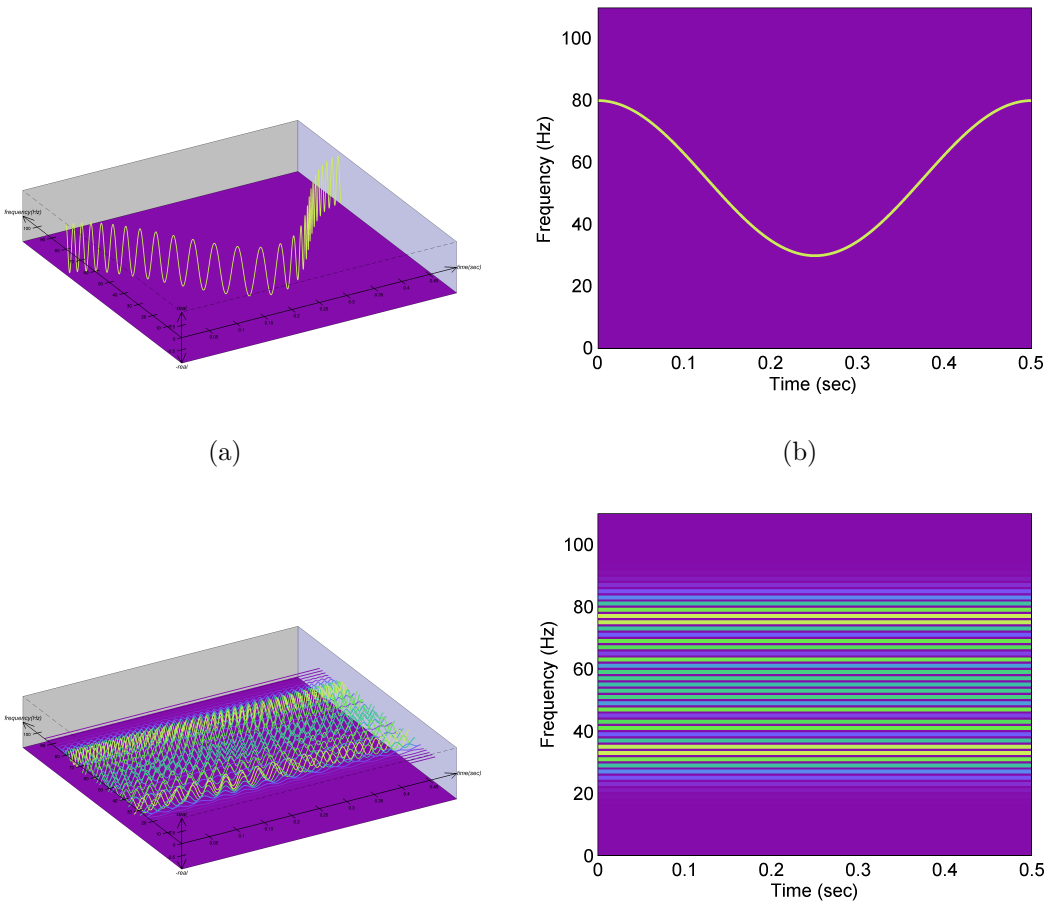


Figure 7.6: Hilbert spectrum for the triangular waveform $x(t)$ with $\omega_0 = 50\pi$ rads/s for the assumptions of (a) SHCs, (c) single AM–FM component with harmonic correspondence, and (e) single AM component. The corresponding time-frequency planes, obtained by projecting out the $s_k(t)$ dimension, are shown in (b),(d),(f).



(a) (b)

(c) (d)

Figure 7.7: Hilbert spectrum for the sinusoidal FM waveform $x(t)$ with $\omega_c = 110\pi$ rads/s, $\omega_m = 4\pi$ rads/s, and $B = 25$ for the assumptions of (a) a single FM component and (c) SHCs. The corresponding time-frequency planes, obtained by projecting out the $s_k(t)$ dimension, are shown in (b),(d).

Chapter 8

NUMERICAL HILBERT SPECTRAL ANALYSIS ASSUMING INTRINSIC MODE FUNCTIONS

8.1 Motivation for Proposed Numerical Methods for Hilbert Spectral Analysis

While the mathematical theory of stationary signals is well developed, mathematical analysis of non-stationary signals is almost nonexistent [31]. A major step toward the analysis of non-stationary signals was made by Huang [70] with the introduction of the Empirical Mode Decomposition (EMD) algorithm. According to Huang [70], "there are some crucial restrictions of the Fourier spectral analysis; the system must be linear; and the data must be strictly periodic or stationary; otherwise, the resulting spectrum will make little physical sense." The original EMD is an iterative method that decomposes a non-stationary signal into a finite set of complete and nearly orthogonal basis functions of the signal; these functions are called Intrinsic Mode Functions (IMFs). Although Fourier analysis and EMD are normally presented as two entirely different signal processing methods, we propose to associate them within the Hilbert Spectral Analysis (HSA), framework, presented in Chapter 7.

This work presents present numerical algorithms for estimating the Instantaneous Amplitude (IA) and the Instantaneous Frequency (IF) of a signal component in the Amplitude Modulation–Frequency Modulation (AM–FM) model, by first assuming that each signal component can be considered as an IMF. The numerical algorithms first decompose the signal into IMFs using an improved version of Huang's original EMD algorithm. Then, assuming IMF components, we propose and algorithm to compute the Hilbert spectrum by demodulating the IMFs to obtain the instantaneous

parameters. Unlike previous studies, special consideration is given to the assumptions made in the demodulation step. This avoids any ambiguity associated with obtaining the instantaneous parameters. It is important to note that while IMFs can be considered latent AM–FM components, there are other classes of AM–FM components that are not IMFs, as illustrated in Figure 7.3. Examples using the proposed algorithm are provided that highlight alternative decompositions compared to traditional Fourier analysis and demonstrate the advantages of using the HSA framework.

8.2 Numeric Hilbert Spectral Analysis

In Chapter 7, we defined the Hilbert spectrum using the AM–FM model in (7.1), representing the latent signal $z(t)$, corresponding to a real observation $x(t)$, as a superposition of AM–FM components. In particular, we defined an AM–FM component in (7.2), in terms of three parameters: IA $a_k(t)$, IF $\omega_k(t)$, and phase reference ϕ_k. in this chapter, we provide numerical methods to compute the IA and IF parameters of the AM–FM components assuming IMF components.

Many methods for computing the parameters of an AM–FM model already exist, each corresponding to a specific set of assumptions. As discussed in Chapter 7, a Fourier transform (FT) corresponds to the assumption of Simple Harmonic Components (SHCs) and a Short-Time Fourier Transform (STFT) corresponds to the assumption of AM components. Other AM–FM models have been proposed with alternative assumptions, but with restrictions which limit the utility of the model.

There is a two-step process when considering the practical estimation of the instantaneous parameters of the AM–FM model in (7.1). In the first step, the signal must be decomposed into a set of latent AM–FM components. Then, the components must be individually demodulated. To date, the most flexible decomposition method for AM–FM modeling is Huang's EMD and its variations [70]. This is because the

EMD does not assume that each component has constant amplitude or frequency and bandlimiting conditions are not strict. In [70], Huang proposed the original EMD algorithm which sequentially determines a set of AM–FM components, i.e., IMFs via an iterative sifting algorithm. The Ensemble Empirical Mode Decomposition (EEMD) algorithm [238] and tone masking algorithm [239] introduced ensemble averaging in order to address the mode mixing problem faced by the original EMD. The complete EEMD was proposed to address some of the undesirable features of EEMD by averaging at the component-level as each component is estimated rather than averaging after all components are obtained [240]. As will be further discussed, several improvements to the sifting algorithm have also been proposed [241].

In the second step of the two-step process, the instantaneous parameters $a_k(t)$ and $\omega_k(t)$ in (7.2) of the kth component, $k = 0, \ldots, K-1$, must be estimated by demodulating the kth IMF. Despite the numerous attempts to demodulate IMFs, most of the proposed methods were not successful as the assumptions made during decomposition were not maintained during demodulation. In this work, we develop a demodulation technique which adheres to the assumptions made during the decomposition step. We point out that Rato proposed an AM–FM demodulation procedure in which the IA estimation was consistent with the assumptions but the IF estimation was not [241]. Also, Huang examined numerous demodulation methods, including an iterative normalization to obtain an FM signal which was then demodulated to estimate the IF using an arctan approach [242]. However, although the iterative normalization method was consistent with the decomposition assumptions, it suffered from numerical instability. Utilizing both Rato's AM estimation and Huang's iterative normalization procedure, we propose a mathematically equivalent method to obtain the IF from the more numerically stable FM signal. We then incorporate the proposed demodulation and EMD into a single HSA–IMF algorithm which provides

very accurate estimates for the IA and IF parameters of the AM–FM model.

8.3 Empirical Mode Decomposition

EMD consists of an iterative procedure for decomposing a signal into a set of IMFs, $\{\varphi_k(t)\}$ [70]. In [70], an IMF is defined as any signal that satisfies two conditions:

C1: Given a signal segment, the number of extrema and the number of zero crossings must be either equal or differ at most by one.

C2: At any point, the mean value of the envelope defined by the local maxima, and the envelope defined by the local minima, is zero.

In the context of Latent Signal Analysis (LSA) and AM–FM modeling, an IMF can be considered as a latent *component*,

$$\varphi_k(t) = \Re\{\psi_k(t)\} = \Re\left\{a_k(t)e^{j\left[\int_{-\infty}^{t}\omega_k(\tau)d\tau+\phi_k\right]}\right\} \tag{8.1}$$

where $\Re\{\cdot\}$ denotes the real operator. Thus, we view an IMF as an inherently complex-valued component, which is not the conventional interpretation. Furthermore, we argue that the definition of an IMF forces the latent signal to have a unique imaginary part, that, in general, does not equal to the Hilbert transform of the real observation, $\mathcal{H}\{\varphi_k(t)\}$.

The EMD approach does not transform the signal but provides an algorithmic signal decomposition, which has both advantages and disadvantages. As a signal decomposition, the EMD has been primarily understood through experimentation [239]. Empirical experiments using white noise showed EMD to act as a dyadic filter bank [238, 243–246]. Using fractional Gaussian noise as a model for broadband noise, it was shown that the built-in adaptivity of EMD makes it behave spontaneously as a 'wavelet-like' filter, i.e., each iteration of the EMD results in a 'detail signal' and a 'trend signal' [244].

Efforts were made to replace the sifting algorithm with alternate formulations which are more mathematically based, such as techniques based on optimization [247–249], machine learning [250], partial differential equations [251–256], and Fourier analysis [257]. Other research on the EMD algorithm include the multivariate EMD [246, 258–263] and the complex EMD [264]. As a final note, not all variations of EMD utilize the IMF component or the proposed AM–FM model, and thus cannot be considered as a form of HSA. Examples include the Hilbert variational decomposition [39, 230, 265], the time-dependent intrinsic correlation [266], and synchrosqueezed wavelet transforms [31, 267, 268].

8.3.1 The Original EMD Algorithm

The original EMD and sifting algorithms proposed by Huang [70] are listed in Algorithms 1 and 2. The purpose of the sifting algorithm is to iteratively identify and remove the trend from the signal, effectively acting as a high pass filter. Step 5 of Algorithm 1 removes the high frequency component $\varphi_k(t)$ that is estimated during the sifting process. The process is then repeated to remove additional IMFs from the signal if they exist. The resulting decomposition is complete and sparse [70, 249, 269]. EMD is formulated with continuous-time signals, however, in practice, EMD is applied to discrete-time signals which may result in inaccurate decompositions. The effects of sampling in the context of EMD were considered by Rilling, and it is generally recommended to oversample but not resample before application of EMD, so that EMD effectively behaves like a continuous operator [270].

Algorithm 1 Empirical Mode Decomposition

procedure $\{\varphi_k(t)\} = \text{EMD}(\ x(t)\ )$
    initialize: $k = 0$ and $x_{-1}(t) = x(t)$
    while $x_{k-1}(t) \neq 0$ and $x_{k-1}(t)$ is not monotonic **do**
        $\varphi_k(t) = \text{SIFT}(\ x_{k-1}(t)\ )$ (see Algorithm 2)
        $x_k(t) = x_{k-1}(t) - \varphi_k(t)$
        replace k with $k + 1$
    end while
    $\varphi_k(t) = x_{k-1}(t)$
end procedure

Algorithm 2 Sifting Algorithm

procedure $\varphi(t) = \text{SIFT}(\ r(t)\ )$
    initialize: $e(t) \neq 0$
    while $e(t) \neq 0$ **do**
        find all local maxima: $u_p = r(t_p),\ p = 1, 2, \ldots$
        find all local minima: $l_q = r(t_q),\ q = 1, 2, \ldots$
        interpolate $u(t)$ using cubic spline and $\{t_p, u_p\},\ p = 1, 2, \ldots$
        interpolate $l(t)$ using cubic spline and $\{t_q, u_q\},\ q = 1, 2, \ldots$
        $e(t) = [u(t) + l(t)]/2.$
        replace $r(t)$ with $r(t) - e(t)$.
    end while
    $\varphi(t) = r(t)$
end procedure

8.3.2 Improving the Sifting Algorithm

If EMD is viewed as an AM–FM decomposition technique, then the sifting algorithm is an iterative way of removing the asymmetry between the upper and lower envelopes (i.e., the signals obtained by cubic spline interpolation of the local maxima and minima, respectively) in order to transform the input to the sifting algorithm in Algorithm 2, $r(t)$, into an IMF [241]. By doing so, low frequency content is discarded at every sifting iteration, effectively making the sifting algorithm behave as a high frequency filter or high frequency component tracker. Due to the doubly iterative nature of EMD and termination conditions, numerical imprecision and differing implementations can lead to very different IMFs. To achieve consistency when using EMD, Rato proposed the following constraints [241]:

Scale: IMFs should scale with the signal.

Bias: Any signal bias should only be reflected in the trend.

Identity: The EMD of an IMF should be the IMF itself (although this constraint could be relaxed in some cases [241]).

Time-reversal: Time-reversal of the signal should correspond to time-reversal of the IMFs.

Several improvements were made to the sifting algorithm (Algorithm 2) to improve the decomposition accuracy [241, 271]. Specifically, in Step 3 of Algorithm 2, the stop criterion was improved for robustness; a consistent method for identifying extrema was identified in Steps 4 and 5; the interpolation end effects were addressed in 6 and 7; and the mean envelope removal was scaled in Step 9.

Although several stopping criteria were proposed [241, 272, 273], Rato suggested the use of a resolution factor, which is the ratio between the energy of the signal at

the beginning of sifting $r(t)$ and the energy of the average of the envelopes $e(t)$ in Algorithm 2. If this ratio increases above a predetermined threshold, then the IMF computation terminates. We have found that Rato's use of a parabolic interpolator [241] to identify the extrema in Steps 4 and 5 works well. Furthermore, the parabolic interpolator inherently determines if a particular sample is a maxima, minima, or neither. End effects appear due to the fact that a given interpolator may not be a good extrapolator [241]. In order to deal with this problem, Rato suggested to insert artificial minima and maxima in order to control the behavior of the interpolator. In addition, in Step 9 of Algorithm 2, $r(t)$ is replaced with

$$r(t) - \alpha e(t); \tag{8.2}$$

specifically, the mean envelope is scaled and the step-size $0 < \alpha \leq 1$ increases the number of sifting iterations but improves stability and robustness of the resulting IMFs [241].

We illustrate a single iteration of the sifting algorithm (Algorithm 2) in Figure 8.1. In Figure 8.1(a), we show two unknown components, $\varphi_0(t)$ (—) and $\varphi_1(t)$ (—) and in Figure 8.1(b), we show the signal under analysis $r(t) = \varphi_0(t) + \varphi_1(t)$ (—), which is the input to the sifting algorithm. Figure 8.1(c) and Figure 8.1(d) show the location and interpolation of the extrema, as given in Steps 4 - 7. Interpolations lead to estimates of the upper envelope $u(t)$ (—) and lower envelope $l(t)$ (—) of $r(t)$. The average of the upper and lower envelopes $e(t) \approx \varphi_1(t)$ is shown in Figure 8.1(e). At the end of the first iteration of sifting, $r(t) - e(t) \approx \varphi_0(t)$, shown in Figure 8.1(f).

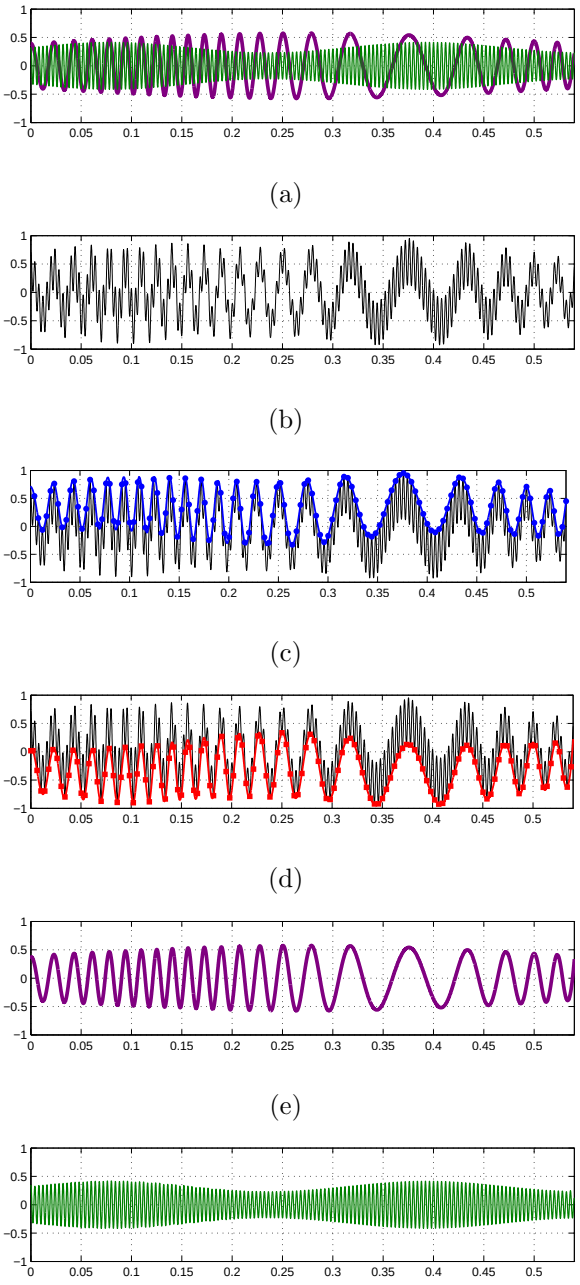
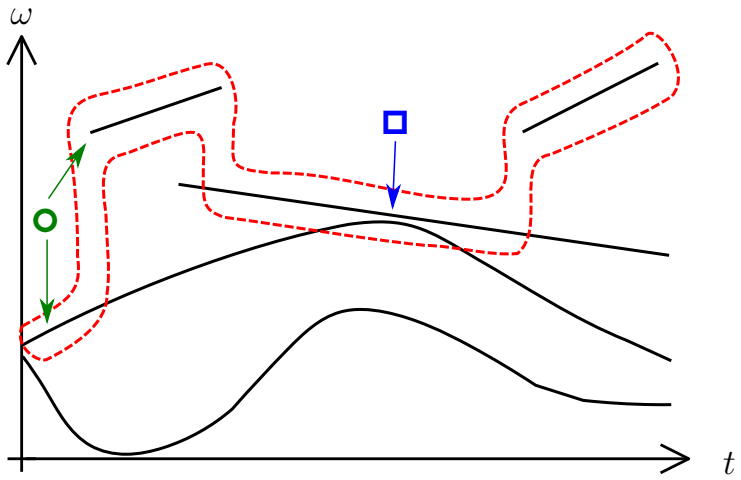


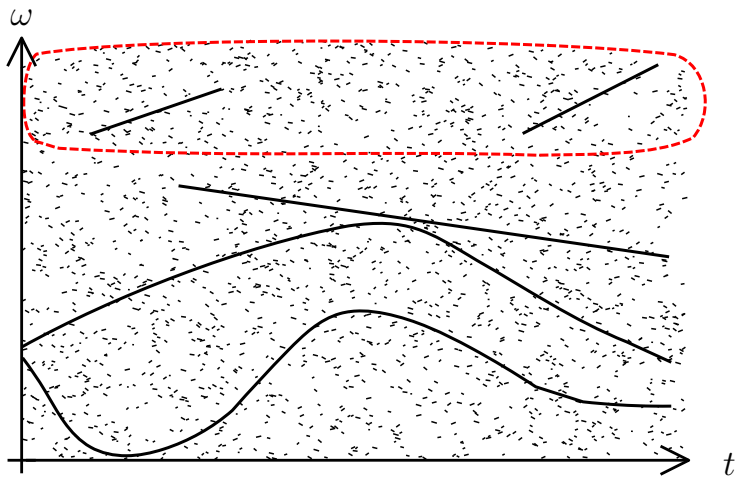
Figure 8.1: This sequence of plots illustrates the steps of a first iteration of the sifting algorithm in Algorithm 2. (a) The example signal composed of the components, $\varphi_0(t)$ (—) and $\varphi_1(t)$ (—); (b) the superposition of the components $r(t)$ (—) and the input to the sifting algorithm; (c)-(d) the upper envelope $u(t)$ (—) and lower envelope $l(t)$ (—) of $r(t)$; (e) average of the upper and lower envelopes $e(t) \approx \varphi_1(t)$; and (f) IMF estimate at first iteration $r(t) - e(t) \approx \varphi_0(t)$.

8.3.3 Improving the EMD Algorithm

The major problem in the EMD algorithm is mode mixing, which occurs when a single IMF consists of components of different scales or when components of similar scale are present in the same IMF [238]. Mode mixing is a consequence of signal intermittency, or more specifically, relative component intermittency, i.e., the number of components as well as the signal component's relative IAs and IFs may change over the duration of the signal. As a result, the particular component(s) tracked by the sifting algorithm in a particular IMF at any instant may change as intermittent components begin or end [239]. This is illustrated in Figure 8.2(a), where in the left part of the figure, we show components of different scales in the same IMF (denoted by ○), while in the center part of the figure, we show two components of similar scale in the same IMF (denoted by □). The ability of EMD to resolve two components, considering both the relative IAs and IFs of the components, was examined and quantified by Rilling [274]. However, as was noted, resolving closely-spaced components may not be the ultimate goal, provided that the application warrants such a decomposition [274].



(a)



(b)

Figure 8.2: In (a) and (b), the assumed components are indicated with — and the first component or high frequency IMF, identified with the sifting algorithm, is indicated within the --- frame. In (a), the mode mixing problem is apparent where we see components of disparate scales being in the same IMF (indicated by O) and components of similar scale in the same IMF (indicated by □). In (b), adding noise and ensemble averaging may assist in resolving mode mixing.

Two commonly used methods of mitigating mode mixing are EEMD [238] and tone masking [239]. EEMD (summarized in Algorithm 3), utilizes zero-mean white noise, $w^{(i)}(t)$, to perturb the signal so that a component can be tracked properly over an ensemble average. As the illustration in Figure 8.2(b) shows, noise can be used to assist the sifting algorithm. Inserting noise with high enough power gives the sifting algorithm something to track when the highest frequency component is intermittent, then vanishes in the ensemble average. Although injecting noise can help to track components properly, a carefully designed masking signal can result in better performance. A situation where a carefully designed masking signal may be beneficial is illustrated in Figure 8.2(a) (denoted by ◻).

Algorithm 3 Ensemble Empirical Mode Decomposition

1: **procedure** $\{\bar{\varphi}_k(t)\} = \text{EEMD}(\ x(t)\ )$
2: initialize: $i = 1, \ldots, I$, where I is the number of trials
3: $\varphi_k^{(i)}(t) = \text{EMD}\left(\ x(t) + w^{(i)}(t)\ \right)$ (see Algorithm 1)
4: $\bar{\varphi}_k(t) = \mathbb{E}[\varphi_k^{(i)}(t)]$, where $\mathbb{E}[\cdot]$ is statistical expectation
5: **end procedure**

EEMD is not without its disadvantages, as it is more computationally complex the EMD, loses the perfect reconstruction property, propagates IMF estimation error, results in inconsistent numbers of IMFs across the trials, and the resulting set of averaged IMFs $\{\bar{\varphi}_k(t)\}$ are not necessarily IMFs [240]. Torres proposed the "complete EEMD" summarized in Algorithm 4, to address some of these issues [240]. Complete EEMD defines a procedure $\text{EMD}_k(\cdot)$ which returns the kth IMF obtained using EMD [240]. This method of EEMD requires fewer sifting iterations, a smaller ensemble size, and recovers the completeness property of the original EMD algorithm to within the numerical precision of the computer [240].

Algorithm 4 Complete EEMD

1: **procedure** $\{\bar{\varphi}_k(t)\} = \text{CEEMD}(\ x(t)\ )$

2: initialize: $k = 1$, β_k is a Signal-to-Noise Ratio (SNR) factor, and $i = 1, \ldots, I$, where I is the number of trials

3: $\bar{\varphi}_0(t) = \frac{1}{I}\sum_{i=1}^{I} \text{SIFT}(\ x(t) + \beta_0 w^{(i)}(t)\ )$ (see Algorithm 2)

4: $x_0(t) = x(t) - \bar{\varphi}_0(t)$

5: **while** $x_{k-1}(t) \neq 0$ and $x_{k-1}(t)$ is not monotonic **do**

6: $\bar{\varphi}_k(t) = \frac{1}{I}\sum_{i=1}^{I} \text{SIFT}(\ x_{k-1}(t) + \beta_k \text{EMD}_k(\ w^{(i)}(t)\ )$ ($\text{EMD}_k(\cdot)$ uses the EMD in Algorithm 1 to provide the kth IMF)

7: $x_k(t) = x_{k-1}(t) - \bar{\varphi}_k(t)$

8: replace k with $k + 1$

9: **end while**

10: $\bar{\varphi}_k(t) = x_{k-1}(t)$

11: **end procedure**

The tone masking method, described in Algorithm 5, uses a deterministic signal $v(t)$ instead of noise as a perturbation and then removes it after IMF estimation. The tone masking technique in Algorithm 5 can be used, for example, in place of Step 4 in Algorithm 1 [239]. Advanced forms of tone masking were also proposed, termed Signal-Assisted EMD (SA-EMD) [245, 275–277]. These methods have additional advantages in their ability to track closely-spaced components, but require careful selection of the masking signal.

Algorithm 5 Tone Masking

1: **procedure** $\bar{\varphi}(t) = \text{TM}(\ x(t), v(t)\ )$

2: $x^+(t) = x(t) + v(t)$ and $x^-(t) = x(t) - v(t)$

3: $\varphi^+(t) = \text{SIFT}(\ x^+(t)\ )$ and $\varphi^-(t) = \text{SIFT}(\ x^-(t)\ )$ (see Algorithm 2)

4: $\bar{\varphi}(t) = \dfrac{\varphi^+(t) + \varphi^-(t)}{2}$

5: **end procedure**

8.3.4 IMF Demodulation

In order to obtain the IA/IF parameters in the AM–FM model, we are required to demodulate the IMFs returned by EMD. One approach, used by Huang, is to apply the Hilbert transform to each IMF in order to obtain estimates of IA and IF [70]. This analysis is often referred to as the Hilbert-Huang Transform (HHT) [241]. Using the Hilbert transform for IMF demodulation, however, only applied when SHCs and harmonic correspondence are assumed, and thus not for the general AM–FM model. A similar argument can be made of many of the other demodulation methods that were considered [242]. A second approach, discussed in Chapter 7, is to let the assumed form of the AM–FM component imply a complex extension. When viewed in the context of LSA, the definition of the IMF in Section 8.3 forces a unique complex extension to the IMF—justifying our view of the IMF as a latent component specified by a real signal.

Inherent in condition C2 of the IMF definition are the following assumptions on the IA $a(t)$:

A1: $a(t_p) = |s(t_p)|$, where $|s(t_p)|$ are the extrema of $s(t)$

A2: $a(t), t \notin \{t_p\}$ is inferred by cubic spline interpolation.

The first assumption, Assumption A1, can be viewed as forcing the imaginary part

to be zero at $t = t_p,\ p = 1, 2, \ldots$. In order to demonstrate this, note that from (7.2),

$$|a(t)| = \sqrt{s^2(t) + \sigma^2(t)} \tag{8.3}$$

and thus, $\sigma(t_p) = 0$ and $a(t_p) = |s(t_p)|$. Assumption A1 also implies non-negativity of $a(t_p)$; however, non-negativity of $a(t)$ is not guaranteed for all time values due to the interpolation.

The second assumption, A2 can be viewed as a relative bandlimiting condition on $a(t)$ controlled by two factors: the choice of interpolator and the density of the extrema points. To demonstrate this, note that $a(t)$ between extrema of $s(t)$ is defined by an interpolator. This implies that $a(t)$ only has as much variation between extrema as the interpolator allows. However, with dense extrema (high IF), much less restrictive constraints are imposed on $a(t)$ than if extrema are sparse (low IF)—thus our view as a relative bandlimiting condition on $a(t)$. Also note that using an interpolator other than cubic spline is likely to change the IA, effectively changing the resulting IMFs [241, 278].

In the IMF definition, Condition C1 requires that the number of extrema and the number of zero crossings must be either equal or differ at most by one. A general AM–FM component may not satisfy this condition. For example, this may occur due to sign reversal of $\omega(t)$ or when assumptions A1 and A2 are not satisfied. If the IF is assumed positive on the AM–FM component as in in Chapter 7, this eliminates the possibility of sign reversal of the IF. However, this assumption cannot be made for all signal classes.

Rato proposed an AM demodulation approach, described in Algorithm 6, that is consistent with the decomposition assumptions in Algorithm 2 [241]. Starting with an IMF estimate $\hat{\varphi}(t)$, we obtain an estimate for IA $\hat{a}(t)$ which can then be used to

estimate the IF via the FM signal

$$\hat{s}_{\mathrm{FM}}(t) = \frac{\hat{\varphi}(t)}{\hat{a}(t)}. \tag{8.4}$$

However, the estimate in (8.4) may result in $|\hat{s}_{\mathrm{FM}}(t)| > 1$. Thus, Huang proposed an iterative normalization procedure, described in Algorithm 7, which removes the AM from the signal to obtain a more accurate $\hat{s}_{\mathrm{FM}}(t)$ [242]. Although the iterative procedure improves demodulation accuracy, we noticed that it can be susceptible to oscillating artifacts introduced by overfitting of the cubic spline interpolator. As a result, we found that these artifacts can be minimized by limiting the number of iterations to three.

Algorithm 6 IMF IA Estimation

procedure $\hat{a}(t) = \mathrm{IAest}(\ \hat{\varphi}(t)\ )$
    $r(t) = |\hat{\varphi}(t)|$
    find all local maxima: $u_p = r(t_p),\ p = 1, 2, \ldots$
    interpolate $\hat{a}(t)$ using cubic spline and $\{t_p, u_p\},\ p = 1, 2, \ldots$
end procedure

Algorithm 7 Obtaining a real FM signal from an IMF

procedure $\hat{s}_{\mathrm{FM}}(t) = \mathrm{iterAMremoval}(\ \hat{\varphi}(t)\ )$
    initialize: $g(t) = \hat{\varphi}(t)$, $b(t) \neq 1$, and $n = 0$
    while $b(t) \neq 1$ and $n < 3$ **do**
        $b(t) = \mathrm{IAest}(\ g(t)\ )$ (see Algorithm 6)
        $g(t) \leftarrow g(t)/b(t)$
        replace n with $n + 1$
    end while
    $\hat{s}_{\mathrm{FM}}(t) = g(t)$
end procedure

We can directly obtain two estimates of the IF, $\pm\hat{\omega}(t)$, by substituting $a(t) = 1$ in (7.2b) and $s(t) = \hat{s}_{\rm FM}(t)$ in (7.2c), and then by computing (7.7). Direct FM demodulation is not straightforward because different approaches exist for obtaining the IF from $\hat{s}_{\rm FM}(t)$. These approaches, although mathematically equivalent, may differ in numerical stability [242].

We propose an FM demodulation method to address the numerical stability issues. We begin by estimating the imaginary part in (7.2c) using the estimate of $\hat{s}_{\rm FM}(t)$ as

$$\hat{\sigma}_{\rm FM}(t) = -\text{sgn}\left[\frac{d}{dt}\hat{s}_{\rm FM}(t)\right]\sqrt{1^2 - \hat{s}^2_{\rm FM}(t)} \tag{8.5}$$

where $-\text{sgn}\left[\frac{d}{dt}\hat{s}_{\rm FM}(t)\right]$ is required to obtain an appropriate four quadrant estimate with assumed positive IF. Figure 8.3 provides an illustration to help explain (8.5). Note that an animated version of this figure can be found at internet site at [8] or in the electronic version of this report. The computationally unstable points, $\{t_0\}$, occur where $\hat{\sigma}_{\rm FM}(t_0) = 0$, thus we can replace a small range around these points $(t_0 - \epsilon, t_0 + \epsilon)$ with interpolated values. The resulting estimated phase function is then given by

$$\hat{\theta}(t) = \arg\left[\hat{s}_{\rm FM}(t) + j\hat{\sigma}_{\rm FM}(t)\right] \tag{8.6}$$

and the IF is obtained using (7.7). We can optionally smooth the resulting IF estimate. Our complete IMF demodulation algorithm is summarized in Algorithm 8.

Note that the proposed IMF demodulation procedure, which does not involve the Hilbert transform or FT, can return the same results for some signals; however, this is not true in general. For example, demodulation of $\cos(\omega_0 t)$ with either the IMF or Hilbert transform demodulation provides an SHC with harmonic correspondence. On the other hand, Hilbert transform demodulation of the triangle waveform provided in (6.26) returns a single AM-FM component with harmonic correspondence as given

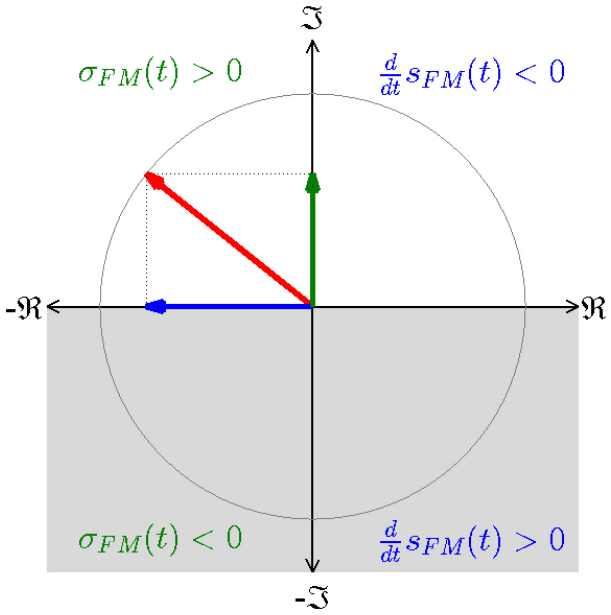


Figure 8.3: In this figure (animated in electronic version of this report or found at [8]), $\hat{s}_{\mathrm{FM}}(t)$, $\hat{\sigma}_{\mathrm{FM}}(t)$ are represented by the blue, and green vectors, respectively. The amplitude-normalized $\hat{\psi}_{FM}(t)$, is represented by the red vector. The magnitude of $\hat{\sigma}_{\mathrm{FM}}(t)$ is easily calculated. The sign of $\hat{\sigma}_{\mathrm{FM}}(t)$ is obtained as follows. We note that when $\hat{s}_{\mathrm{FM}}(t)$ is decreasing (blue vector moves left), $\hat{\sigma}_{\mathrm{FM}}(t)$ is always positive (green vector is in the upper half plane) and when $\hat{s}_{\mathrm{FM}}(t)$ is increasing (blue vector moves right), $\hat{\sigma}_{\mathrm{FM}}(t)$ is always decreasing. Thus, by reversing the sign of the derivative of $\hat{s}_{\mathrm{FM}}(t)$, we obtain the sign of $\hat{\sigma}_{\mathrm{FM}}(t)$.

in (6.29), while IMF demodulation returns the IF in (6.34) and constant IA.

8.3.5 Proposed Algorithm for HSA Assuming IMFs

To the best of our knowledge, the proposed algorithm for HSA assuming IMFs presented in this section has not been previously considered. In particular, our resulting proposed HSA–IMF algorithm incorporates the most desirable features of the complete EEMD and tone masking to address the mode-mixing problem, Rato's improvements to the sifting algorithm, and our proposed demodulation method. The HSA–IMF algorithm is summarized in Algorithm 9.

Algorithm 8 IMF Demodulation

procedure $[\hat{a}(t), \hat{\omega}(t)] = \text{IMFdemod}(\ \hat{\varphi}(t)\ )$

  $\hat{a}(t) = \text{IAest}(\ \hat{\varphi}(t)\ )$ (see Algorithm 6)

  $\hat{s}_{\text{FM}}(t) = \text{iterAMremoval}(\ \hat{\varphi}(t)\ )$ (see Algorithm 7)

  $\hat{\sigma}_{\text{FM}}(t) = -\text{sgn}\left[\dfrac{d}{dt}\hat{s}_{\text{FM}}(t)\right]\sqrt{1^2 - \hat{s}^2_{\text{FM}}(t)}$

  Find $\{t_0\}$ such that $\hat{\sigma}_{\text{FM}}(t_0) = 0$

  For each t_0, replace $(\hat{\sigma}_{\text{FM}}(t_0 - \epsilon), \hat{\sigma}_{\text{FM}}(t_0 + \epsilon))$ with interpolated points

  $\hat{\omega}(t) = \dfrac{d}{dt}\arg\left[\hat{s}_{\text{FM}}(t) + j\hat{\sigma}_{\text{FM}}(t)\right]$

end procedure

Algorithm 9 HSA–IMF Algorithm

procedure $\{\bar{\varphi}_k(t), \hat{a}_k(t), \hat{\omega}_k(t)\} = \text{HSA} - \text{IMF}(\ x(t)\ )$

  initialize: $x_{-1}(t) = x(t)$, $k = 0$, β_k is an SNR factor, ε is an energy threshold, and I is the number of trials

  while $\int |x_{k-1}(t)|^2 dt > \varepsilon$

   and $x_{k-1}(t)$ is not monotonic **do**

    $\bar{\varphi}_k(t) = \dfrac{1}{I}\displaystyle\sum_{i=1}^{I} \text{TM}(\ x_{k-1}(t),\ \beta_k v^{(i,k)}(t)\ )$ (see Algorithm 5)

    $[\hat{a}_k(t), \hat{\omega}_k(t)] = \text{IMFdemod}(\ \bar{\varphi}_k(t)\ )$ (see Algorithm 8)

    $x_k(t) = x_{k-1}(t) - \bar{\varphi}_k(t)$

    $k \leftarrow k + 1$

  end while

  $\bar{\varphi}_k(t) = x_{k-1}(t)$

end procedure

Although the masking signal $v^{(i,k)}(t)$ in Step 4 of Algorithm 9 is usually a carefully-selected real-valued AM–FM signal [275, 276], we choose $v^{(i,k)}(t)$ as white, Gaussian, lowpass-filtered noise with cutoff frequency below the amplitude-weighted IF [279]

of the previously found IMF. This replaces the need to use sifted white noise in complete EEMD. Note, however, that we do allow for more sophisticated approaches for choosing the masking signal. Note that this introduces a *feedback loop* between decomposition and demodulation. This changes EMD from simply a decomposition method to a full HSA method because the IA and IF estimates of the kth component are inherently computed and then used to design the masking signal for the $(k+1)$th component.

8.4 Examples using the HSA–IMF Algorithm

In this section, we demonstrate the HSA–IMF algorithm for signal analysis and compare it to conventional STFT analysis using both synthetic and real-world signals. In these examples, we use the proposed visualization method presented in Section 7.7, orthographically projected onto the time-frequency plane, in order to plot the instantaneous parameters resulting from HSA. We also plot the STFT magnitude (STFTM) using the same colormap to facilitate comparisons [237]. The STFT is computed with a Hamming window that is 4,906 samples long and uses 4,095 sample overlapping windows. The MATLAB functions used for these examples can be found at [8].

8.4.1 Synthetic Signals

We provide two examples using synthetic signals with known underlying signal models. The synthetic signal examples demonstrate the proposed algorithm for a single component in a noiseless environment. In this case, mode mixing is not a problem, and we initialize $I = 1$, $\alpha = 0.95$, and $\beta_k = 0$ in Algorithm 9.

In the first example, we analyze a signal with a slow-varying AM and fast-varying

FM given by

$$x(t) = \Re\left\{ a(t) \exp\left[j\left(6000\pi t + \int_{-\infty}^{t} m(\tau) d\tau \right) \right] \right\},\ 0 \le t \le 1 \tag{8.7}$$

where the IA is

$$a(t) = e^{-(t-0.5)^2/25} \tag{8.8}$$

and the FM message is

$$m(t) = 250 \sin(140\pi t) + 2000[\exp(-4t) - 1]. \tag{8.9}$$

Figure 8.4(a) shows the STFTM, where we observe classic harmonic structure due to the inherent assumption of SHCs even though (8.7) shows only one component. Figure 8.4(b) shows a single component in the Hilbert spectrum with a fast FM oscillation consistent with the underlying signal model in (8.7). The IA in (8.8), consisting of a Gaussian envelope centered at $t = 0.5$, is visible as variation in color intensity of the component shown in Figure 8.4(a). The FM message in (8.9) consists of a 70 Hz oscillation superimposed on a decaying exponential; this FM message is offset by a 3,000 Hz carrier. The FM message is reflected by the vertical behavior in Figure 8.4(b) where the IF is swept from 3,000 Hz to 1,000 Hz with a 70 Hz oscillation. Due to more appropriate assumptions of an underlying signal structure, this example shows that the HSA–IMF parametrization can more closely match the underlying signal model than traditional methods.

In the second example, we analyze a signal with a fast-varying AM and slow-varying FM given by

$$x(t) = \Re\left\{ a(t) \exp\left[j\left(1000\pi t + \int_{-\infty}^{t} m(\tau) d\tau \right) \right] \right\} \tag{8.10}$$

where the IA is

$$a(t) = \frac{1}{2} + \frac{1}{3}\sin(100\pi t) + \frac{1}{5}\sin(200\pi t) \tag{8.11}$$

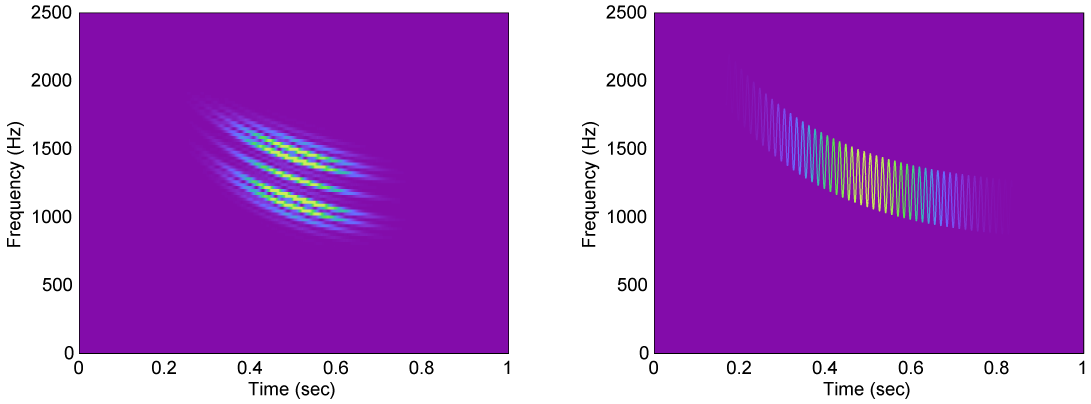
and the FM message is

$$m(t) = 150 \sin(2\pi t). \tag{8.12}$$

Figure 8.5(a) shows the STFTM, where we again observe classic harmonic structure resulting from the inherent assumption of SHCs even though (8.10) consists of only a single component. Figure 8.5(b) shows a single component in the Hilbert spectrum with a fast AM oscillation, captured by variation in color intensity of the component shown, and is consistent with the underlying signal model in (8.10).

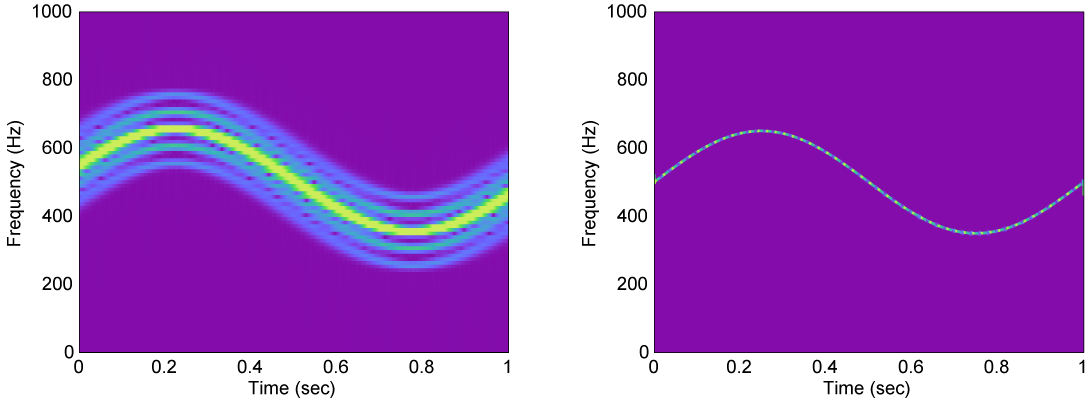
As described in Chapter 7, signals composed of narrowband components can have similar Fourier and Hilbert spectra, however, signals composed of wideband components can have spectra which are quite different. Fourier analysis of wideband signals may result in multiple ridges in terms of spectral frequencies, which may be further decomposed into narrowband components. However, for many real-world signals, a wideband component, such as the AM–FM component in (7.2), is a more appropriate choice rather than multiple, narrowband components. The appearance of structure in the Fourier spectrum may indicate the presence of a wideband component(s) in the signal as demonstrated in Figure 8.4(a) for Example 1 (fast-varying IF) and in Figure 8.5(a) for Example 2 (fast-varying IA).

Both the STFT and HSA–IMF are equally valid models for the real signal $x(t)$—the resulting signal representations simply match the underlying assumptions on the IA and IF of the components. The complex extensions assumed in STFT and implied in HSA–IMF are fundamentally different, and because of this, the imaginary part $y(t)$ can be different for the two analysis methods. This ultimately results in a different $z(t)$ for each analysis and consequently different instantaneous parameterizations even though both models produce the same $x(t)$ in (6.3).



(a) (b)

Figure 8.4: (a) STFTM and (b) Hilbert spectrum for the fast-varying FM and slow-varying AM synthetic signal given in (8.7)-(8.9) in Example 1. The wideband FM message results in harmonic structure under Fourier analysis in (a) and a fast-frequency-varying component under HSA in (b).



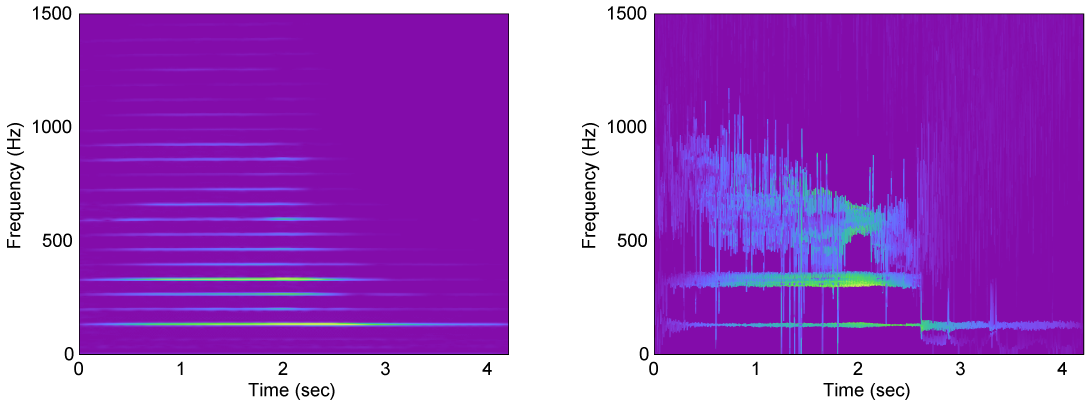
(a) (b)

Figure 8.5: (a) STFTM and (b) Hilbert spectrum for the fast-varying AM and slow-varying FM synthetic signal given in (8.10)-(8.12) in Example 2. The wideband IA results in harmonic structure under Fourier analysis in (a) and a fast-amplitude-varying component under HSA in (b).

8.4.2 *Real-World Signals*

We provide two examples using real-world audio signals. In Algorithm 9, we initialize $I = 200$, $\alpha = 0.95$, and through experimentation, we select $\beta_k = 4$ for all k. In addition, we apply a 1 ms moving-average filter to smooth the IF estimate.

In the first example, we analyze a recording of a single note played on a cello using a sampling frequecy of 22.050 kHz. Figure 8.6(a) shows the STFTM where we again see classic harmonic structure resulting from the inherent assumption of SHCs. We note that the fundamental frequency is approximately 67 Hz (15 spectral lines evenly-spaced over 1,000 Hz) and two dominant spectral lines at the second harmonic (about 133 Hz) and at the fifth harmonic (about 333 Hz). In addition, there is a brief dominant spectral line at $t = 2$ s corresponding to the ninth harmonic at about 600 Hz.



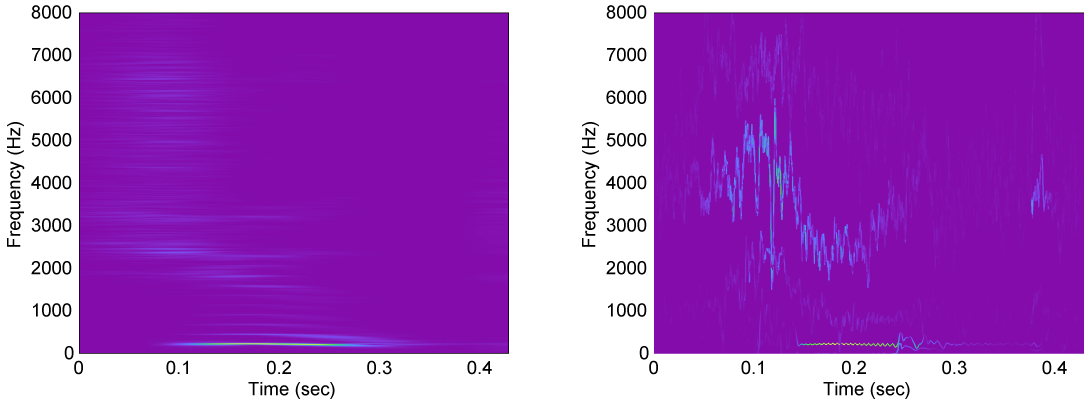
(a) (b)

Figure 8.6: (a) STFTM and (b) Hilbert spectrum for the cello recording in Example 1. The lower two components in (b) range in IF from 120-140 Hz and from 300-360 Hz corresponding to the dominant spectral lines in the Fourier spectrum at 133 Hz and 333 Hz. The harmonics above 500 Hz in (a) and the upper component in (b) with IF ranging from 500-1,000 Hz partially accounts for the spectral richness of this instrument's note.

Figure 8.6(b) shows the five components resulting from the HSA–IMF algorithm,

where we see three dominant components. The lower two components, ranging in IF between about 120 to 140 Hz and between about 300 to 360 Hz, correspond to the dominant spectral lines in the Fourier spectrum. The upper component also exhibits significant energy at $t = 2$ s between about 550 and 750 Hz, corresponding to the brief dominant spectral line, i.e., ninth harmonic in the Fourier spectrum.

In the second example, we analyze a recording of the word "shoot" sampled at 44.1 kHz. Figure 8.7(a) shows the STFTM, where we see the spectral energy of the fricative "SH" over $0 \leq t \leq 0.15$ s, scattered over the range 0 to 8 kHz. The spectral energy for the vowel "UW" over $0.15 \leq t \leq 0.25$ s is concentrated at a fundamental of about 230 Hz. The spectral energy for the stop "T" over $0.37 \leq t \leq 0.4$ s is very weak and spread across the 0 to about 8 kHz band and hence, not visible in the plot.



(a) (b)

Figure 8.7: (a) STFTM and (b) Hilbert spectrum for the speech recording "shoot" in Example 2. The "SH" fricative appears in three components with IF ranges of 6,000-7,000 Hz, 2500-5000 Hz, and 1,000-2,500 Hz. The vowel "UW" is clearly captured in a single component near 230 Hz. Unlike in the Fourier spectrum, the stop "T" is clearly captured in the Hilbert spectrum by the first component near $t = 0.4$ s.

Figure 8.7(b) plots the five components returned by the HSA–IMF algorithm. The "SH" fricative appears in three components with IF ranging between about 6 to 7 kHz (zeroth component), between about 2.5 to 5 kHz (first component), and between

about 1 to 2.5 kHz (second component); however, it is mostly captured in the second component with rapidly varying AM and FM. The vowel "UW" is clearly captured in a single component near 230 Hz, exhibiting some FM variation conjectured to be natural jitter. Unlike in the Fourier spectrum, the stop "T" is clearly captured in the Hilbert spectrum by the first component near $t = 0.4$ s.

The IA/IF parameterizaton of the components provides an alternative and very simple method of estimating a formant frequency F, via an IA-weighted average of the IF [279]

$$F = \frac{\int \omega_k(t) a_k(t) dt}{\int a_k(t) dt}. \tag{8.13}$$

Thus HSA of speech provides a unique method for automatic formant estimation.

One of the main advantages of our proposed HSA is the ability to analyze and quantify fine spectral structure that exists in speech [72]. We have performed HSA for the following twelve vowels and three diphthongs in /hVd/ context: heed, hid, hayed, head, had, hod, hawed, hoed, hood, who'd, hud, herd, hoyed, hide, and how'd [7, 26]. This analysis includes the /hVd/ utterances from a female speaker and two male speakers. The resulting Hilbert spectral plots and spectrograms are collected into contact sheets to facilitate comparison and can be found online at [280]. In the online Hilbert spectral plots, we have used a Savitzky-Golay filter to smooth the IF while preserving the fine structure necessary for speech analysis [281–283]. We used one of two Savitzky-Golay filters depending on the level of smoothing desired. The filter parameters are order $\kappa = 1$ and frame length $f = 255$ for aggressive smoothing and $\kappa = 9$ and $f = 65$ for reserved smoothing.

8.5 Comments on HSA–IMF Algorithm

8.5.1 Resolving Closely-Spaced Components

EMD was criticized for its inability to resolve closely-spaced components, and there were numerous studies on its resolving ability [239, 274, 284]. Rilling has investigated EMD's ability to resolve two tones as a function of a relative amplitude parameter and a relative frequency spacing parameter. This analysis describes regions in this parameter space where EMD returns one or two components [274]. As Rilling noted, the goal of EMD is not to resolve closely-spaced components but rather to resolve components that are suitably matched to an underlying signal model or components that are compatible with assumptions made on the signal model [274].

As an example, consider two infinite duration tones, $\cos(\omega_a t)$ and $\cos(\omega_b t)$, with $\omega_b > \omega_a$. We can express the sum of these tones as

$$x(t) = \Re\left\{\exp(j\omega_a t) + \exp(j\omega_b t)\right\} \tag{8.14a}$$

$$= \Re\left\{2\cos\left[(\omega_b - \omega_a)\, t/2\right]\exp\left[j\,(\omega_b + \omega_a)\, t/2\right]\right\}. \tag{8.14b}$$

If ω_a and ω_b are sufficiently far apart, both Fourier analysis and EMD can resolve two SHCs as in (8.14a). On the other hand, if ω_a and ω_b are not sufficiently far apart, EMD can only resolve a single IMF as in (8.14b). As is well known, when ω_a and ω_b are closely-spaced, the signal exhibits a beat effect. In the human auditory system, these closely-spaced tones are not perceived as two distinct tones but rather a single AM tone [27]. As noted by Deering, EMD may correspond to the psychoacoustics of human hearing [239]. As Rilling pointed, a decomposition into SHCs may not be an appropriate solution "...if the aim is to get a representation matched to physics (and/or perception) rather than to mathematics" [274].

A generalized example of this beat effect was given in Example 2 in Subsection

8.4.1 in Figure 8.4. When the signal in this example is analyzed with the STFT, five closely-spaced tones are shown as in Figure 8.4(a). However, if these tones are closely spaced, then as noted, they may be perceived as a single tone with AM variation. This is demonstrated with an analysis using HSA–IMF in Figure 8.4(b).

A similar example regarding an FM signal was given in Example 1 in Subsection 8.4.1, and it is essentially the same signal used in FM synthesis pioneered by Chowning [227]. In Chowning's work, the FM signal was expressed as a superposition of SHCs weighted by Bessel functions which showed rich spectra associated with this signal. Figure 8.5(a) illustrates this rich spectra via the presence of multiple harmonics. In the AM–FM model, such rich spectra may be encapsulated in a single component as illustrated in Figure 8.5(b); a decomposition into SHCs may not be an appropriate solution to describe the underlying signal model.

8.5.2 Computational Complexity of HSA–IMF

The HSA–IMF algorithm consists of a triple-nested loop that incurs significant computation, depending on the signal and parameter choices. If the signal has many underlying components, EMD can require more computation due to the extrema searches and interpolations. Unfortunately, there is no way to predict the number of components ahead of time.

The outermost loop in Algorithm 9 described in Step 3 iteratively removes the IMF estimated by tone masking until termination conditions are reached. Hence, there is no way to predict the number of iterations required for terminating the loop. The middle loop ensemble averages the IMFs returned from tone masking in Step 4 of Algorithm 9. This average is controlled by a fixed number of trials, I. The inner loop results from the tone masking procedure in Step 4 of Algorithm 9 calling the sifting algorithm twice, in Step 3 of Algorithm 5. This in turn calls Algorithm 2, in which

Step 3 iteratively estimates an IMF. Within this inner loop, the step-size, introduced in (8.2) and used in Step 9, also controls the speed at which termination conditions are reached. Of these loops, the innermost loop requires the most computation due to the search for extrema and interpolation. Considered together, the iterative nature of the outer and inner loops, coupled with the computationally complex inner loop, can require significant computation depending on the signal length and sample rate. As pointed out, signal oversampling is required for robust estimates of the IMFs, that further increases computation. Finally, IMF demodulation in Step 5 of Algorithm 9 occurs in the outermost loop and does not add significant computational burden. Much of this computation can occur in parallel [285–288]; an example of the parallel processing is the ensemble averaging in Step 4 of Algorithm 9, and the search for extrema in Steps 4 and 5 of Algorithm 2.

We implemented Algorithm 9 in MATLAB, where the trials are computed in parallel using a `parfor` loop, and we timed the computation for the synthetic and real-world signal examples. Our PC consists of an eight-core AMD FX at 4.01 GHz with 32 GB RAM. The computation time results are given in Table 8.1, where we see that for relatively short audio signals, decomposition may require relatively large β and I, leading to long computation times.

8.5.3 HSA–IMF Algorithm Robustness

The goal in HSA is to obtain an AM–FM decomposition with meaningful interpretation. Thus, proper identification of the underlying components is required from Algorithm 9. Two parameters must therefore be carefully chosen: the Signal-to-Noise Ratio (SNR) factor β_k and the number of trials I. As a reminder, β_k weighs the additive masking signal used to mitigate mode mixing and I minimizes, through ensemble averaging, the influence of the additive masking signal on the resulting IMF; both of

Table 8.1: Benchmarks for computing the AM–FM model parameters using Algorithm 9.

Example	Sampling Frequency (kHz)	Signal Duration (s)	SNR β	Trials I	Computation Time (s)
Synthetic 1	44.1	1	0	1	1.3
Synthetic 2	44.1	1	0	1	0.9
Cello	22.05	4.21	4	200	320.1
Speech	44.1	0.43	4	200	52.4

these parameters appear in Step 4 of Algorithm 9.

In our experience with real-world signals, where the underlying signal model is unknown, we begin by selecting $\beta_k = 0$ and $I = 1$ and visualizing the resulting IMFs by plotting the real-time plane or time-frequency plane. Mode mixing is evident in the real-time plane when the frequency of the waveform changes abruptly. Mode mixing is also evident in the time-frequency plane when the IMF is similar to the illustrated IMF within the red-dashed (---) frame in Figure 8.2(a). If mode mixing is present, we increase β_k and I. This process is repeated until reasonable IMFs are obtained, keeping in mind the associated computational load for large I. Although this process for decomposition is heuristic, such refinement is typically present in many tine-frequency analysis methods[3], e.g., choice of window function in the STFT analysis and wavelet selection in wavelet analysis.

The step-size parameter α introduced in (8.2) and used in Step 9 of Algorithm 2, is of secondary importance and merely scales the trend which is removed from the sifted signal. This scaling is used to minimize the impact of possible overshoot in the trend. In our experience, selecting $\alpha = 0.95$ gives satisfactory performance, noting

that lower values lead to additional iterations and more computation.

The termination condition in the sifting algorithm can be selected to be too restrictive, hence preventing convergence. In our implementation, we include a maximum number of iterations, typically 50, to guarantee algorithm convergence. As a final point, IMFs with excessively large values are omitted from the ensemble.

8.6 Discussion

In this chapter, we presented an end-to-end solution for the estimation of instantaneous parameters of the AM–FM model assuming IMF components. By leveraging the theory developed in Chapter 7 to interpret and demodulate the results returned by EMD, we provided a complete numerical method for HSA. We provided examples of HSA–IMF using synthetic signals and argued that the resulting decompositions were more representative of the underlying signal models as compared to conventional Fourier analysis. Examples of HSA–IMF on real-world signals were shown to allow for alternative, and possibly more useful interpretation, of the underlying signal model. Finally, we discussed computational aspects of the proposed algorithm.

Chapter 9

CONCLUSIONS AND FUTURE WORK

9.1 Conclusions

9.1.1 Speech Assessment

The assessment of speech intelligibility is the cornerstone of clinical practice in speech-language pathology, as it indexes a patient's communicative handicap. There has been a desire to develop efficient, objective, and reliable measures that can be added to the clinical repertoire. Given the relationship of Vowel Space Area (VSA) and intelligibility decrements [17–19, 21, 289], it is critical to have a sensitive and efficient assessment of VSA; this includes the exploration of a more complete assessment of the vowel space by including the complete repertoire of vowels in spoken language. In the current investigation, an automated assessment of the VSA demonstrated a strong relationship with the traditional methods of VSA derivation. Moreover, the proposed method is fully automated and was demonstrated to capture a more complete assessment of the VSA by allowing for arbitrary VSA shapes, rather than only triangle or quadrilateral shaped VSAs. Moving forward, the relationship between the proposed calculation of VSA will be related to intelligibility ratings to understand its relationship with intelligibility decrements. The success with which the automated procedure estimates the VSA along with the ease of computation, makes the proposed approach an attractive metric for characterizing speech motor control.

9.1.2 Mean Formant Trajectories

Since the introduction of formant analysis, Peterson and Barney [26] found that increased crowding of vowels in the static $F1-F2$ formant space is not accompanied by an increase in perceptual confusions among vowels. Hillenbrand et al. [7] speculated that this could be explained by spectral change, which further supports the idea that formant trajectories are important for phoneme perception. This is further elucidated as we consider the formant trajectories of phonemes other than vowels. Even though phonemes can appear considerably crowded in the $F1-F2$ space, most have distinct trajectories across this space during the duration of the production, presumably lending to accurate perception.

Although the effectiveness of spectral information provided by the first two formant frequencies in vowel identification is indisputable, it was also shown that temporal information provides additional cues [7, 148, 150, 150, 151, 151**?** –157]. Static measurements alone do not explain why vowels are perceived correctly despite having similar temporal midpoints; however, it is the change across time that provides insight into this perceptual process. In this work, we illustrated that this holds true for other phoneme types as well. Furthermore, the use of duration as an additional feature to differentiate between phoneme with closely spaced $F1-F2$ values is common. Although the increase in classification performance when including duration cannot be denied, this does not imply that duration is the best or even the most relevant way to discriminate between these sound types. For example, vowel duration is sensitive to speaking rate, where as a formant trajectory computed relative to vowel duration, as performed in this study, is not. This suggests that formant trajectories may be a measurement robust to dialectical variations or pathological changes in rate, while still capturing variations in phoneme productions. Furthermore, because $F1-F2$ val-

ues roughly relate to jaw/tongue excursion, the use of formant trajectories as a proxy for kinematic movement may be useful as a means to track improvement of therapy or progression of disease for pathological speakers. This could further be validated in experiments to determine how trajectories with reduced/increased variation relate to perceptual errors and the communication disorder associated with such changes in formant trajectories.

It should be noted that the automated formant extraction was not hand verified or individually optimized due to the vast amount to speech material and speakers utilized. Also, raw formant values in Hertz were directly averaged rather than averaging the value subsequent to formant normalization. Nevertheless, the reported values are well representative of formant trajectories in American English, and can serve as a basis for progressing the investigation of formant trajectories in acoustical phonetics.

9.1.3 Hilbert Spectral Analysis

Although the concept of frequency, defined as the number of cycles undergone during one unit of time, is well understood, the concept of instantaneous frequency (IF) is often controversial [32–34]. A major step toward the analysis of non-stationary signals was made by Huang [70] with the introduction of the Empirical Mode Decomposition (EMD) algorithm. In order to resolve some of the controversy surrounding the definition of IF, we reframed the classic complex extension problem as a Latent Signal Analysis (LSA) problem where the objective is to determine the complex-valued latent signal, $z(t)$ from the real-valued observation, $x(t) = \Re\{z(t)\}$; here $\Re\{\cdot\}$ is the operator that gives the real part of $z(t)$. We used this framework to demonstrate that not all signal assumptions are matched to harmonic correspondence, and hence Gabor's method and the Hilbert transform for complex extension are not always the best matched signal analysis methods. By relaxing the harmonic correspondence as-

sumption, there are many choices for the imaginary part and furthermore, many of these complex extensions can be useful in modeling physical phenomena.

We presented a theory for the Hilbert spectrum as the mathematical framework to generalize the LSA problem in which we seek a representation of a complex signal, $z(t)$, consisting of a superposition of latent AM–FM components parameterized by a set of Instantaneous Amplitude (IA) and IF pairs. Using the AM–FM signal model as the superposition of latent components allows for many possible forms for the IA and IF of each component, thus allowing for considerable freedom in the signal model. We presented the analogue of the LSA problem in the frequency domain, where we showed that a latent spectrum cannot be uniquely matched to the spectrum of the real observation because of the structure imposed by the real operator. We showed that, without the harmonic correspondence assumption, there are many choices for the latent spectrum. We proposed a novel three-dimensional (3-D) visualization of the Hilbert spectrum which plots the IF $\omega(t)$ versus the real signal $s(t)$ versus time and varies the intensity of color with respect to magnitude of the IA $|a(t)|$ for each latent component, and allows for the visualization of instantaneous parameters. We recast time-frequency analysis and Gabor's analogies to quantum mechanics to an analysis method where uncertainty is in the imaginary part and not in frequency. Further, by moving away from Simple Harmonic Components (SHCs) with harmonic correspondence, we allow for a new and powerful way to analyze non-stationary signals.

We recognized that an Intrinsic Mode Function (IMF) of the EMD algorithm can be considered a latent AM–FM component and leveraged Hilbert Spectral Analysis (HSA) theory to interpret both the IMF and EMD. With this interpretation, we showed that the definition of an IMF unambiguously forces a unique complex signal extension to a the IMF. We showed that the Hilbert transform cannot be used for

IMF demodulation and proposed an IMF demodulation method that is compatible with the IMF definition. Finally, we applied modifications to the EMD algorithm and integrated it with IMF demodulation to calculate the IA/IF parameters of the real observation $x(t)$, thus providing a numerical method for HSA.

We applied the proposed HSA–IMF algorithm to compute and also visualize the Hilbert spectrum of speech. We compared the Hilbert spectrum of vowel to their corresponding Short-Time Fourier Transform Magnitude (STFTM) to illustrate advantages of using HSA. One of the advantages is revealing spectral fine structure on small time-scales such as within a single glottal pulse, which may not be apparent in the STFTM. We also leveraged the IA/IF parameterization of the AM–FM components to provide a simple formulation for computing formant frequencies. Although the HSA–IMF algorithm is iterative and requires more computation than the fast Fourier transform used for Fourier analysis, we demonstrated that the Hilbert spectra of speech sounds may be computed in a few seconds on an ordinary PC.

9.2 Future Work

9.2.1 Mean Formant Trajectories

The current study utilized two different databases, with varying contexts: isolated vowels from the Hillenbrand database and vowels taken from productions of sentences in the TIMIT database. It is possible that the differences seen between the trajectories from the two databases are due to contextual differences or coarticulation effects, or due to the different methods used to compute the formants. The significance of these differences is uncertain and should be explored in further research to ascertain true differences of formant trajectories among children and adults of different genders. Other topics include a close examination of the influence of regional dialect on the

vowel trajectories, and a comparison of individual trajectories to an average population trajectory. Finally, the use of format trajectories as a proxy for the kinematic movement associated with speech production should be further explored.

9.2.2 Computation of the Hilbert Spectrum

There are a few more approaches that could be investigated to further improve the HSA-IMF algorithm. Some of the approached include the use alternative masking signals, alternative interpolators, and improvements to IMF demodulation. In the HSA–IMF algorithm, we use lowpass filtered noise as the masking signal. However, for certain signal analysis problems, more sophisticated masking signals were proposed [275, 276, 290]. Ideally, the masking signals would have properties similar to the underlying components, but this would would likely require prior knowledge of the signal model. As the HSA–IMF is based on the EMD, interpolation is needed for the iterative estimation of the IMFs. As noted earlier, the cubic spline interpolator may be susceptible to overfitting which can lead to inaccurate IMF estimates. Alternatives to cubic spline interpolation have been investigated including the B-spline and Akima interpolators [241, 291–296], however, these interpolators do not appear to offer significant improvements. Nevertheless, improvements in interpolation may lead to more robust decomposition and demodulation. Finally, IMF demodulation requires estimation of the IF which uses Huang's iterative normalization procedure. Unfortunately, the required interpolation in the normalization procedure can also result in an overfitting of the cubic spline interpolation. As noted earlier, changing the cubic spline interpolator in Algorithm 8 changes the IMF. Thus, a change of the interpolator in the IA estimator requires the same change in the sifting algorithm.

9.2.3 Speech Evaluation using the Hilbert Spectrum

We propose to apply our HSA–IMF algorithm to the problem of speech evaluation and compare our results to those obtained when analysing speech with both narrowband and wideband spectra. We intend to demonstrate how the AM–FM components, assumed to be IMFs, align well with the energy concentrations of the STFTM and furthermore highlight the fine structure present in the Hilbert spectrum. We intend to show never before seen intra-glottal pulse phenomena that are not readily apparent when using other analysis methods. Such fine-scale analysis may have application in speech-based medical diagnosis and automatic speech recognition for pathological speakers.

Our work has demonstrated that there is potential in utilizing the spectral fine structure obtained through HSA for evaluating aspects of speech that have traditionally been difficult, such as evaluation of vocal quality. For example, measures similar to jitter and shimmer, which have proven useful in the detection of vocal tremor and vocal flutter, may be accessible from the fine-grained analysis obtainable though HSA. Finally, we are currently investigating the efficacy of features extracted from the Hilbert spectrum for classification of dysarthic speech, with the goal of providing new methods for speech-based medical diagnosis and monitoring.

There also exist several practical limitations that must be overcome before efficient and widespread use of the HSA–IMF algorithm for speech analysis can be achieved. Due to the triple iterative nature of HSA–IMF, termination conditions, numerical imprecision, and different implementation approaches can lead to very different IMFs. If the HSA–IMF algorithm is to be used as the front-end of a feature extraction algorithm, care must be taken to ensure consistency of the decomposition. Also, we have utilized filtered white Gaussian noise as the masking signal for the HSA–IMF

algorithm. While this provides a simple method for masking signal design, given no other information about the latent signal, it may not be optimal once we know the latent signal consists of speech. As a result, an investigation into alternative masking signals for speech analysis should be performed.

REFERENCES

[1] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*, Prentice Hall, 1988.

[2] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[3] L. Cohen, *Time-Frequency Analysis*, Prentice Hall, 1995.

[4] P. Loizou, *Speech Enhancement Theory and Practice*, CRC Press, 2007.

[5] L. Malfait, J. Berger, and M. Kastner, "P.563 - The ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 1924–1934, Nov. 2006.

[6] "Full IPA Chart," `https://www.internationalphoneticassociation.org/content/full-ipa-chart`, 2016.

[7] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.*, vol. 97, no. 5, pp. 3099–3111, 1995.

[8] "Hilbert Spectral Analysis," `http://www.HilbertSpectrum.com`, 2015.

[9] P. Flipsen and S. Lee, "Reference data for the American English acoustic vowel space," *Clin. Linguist. Phon.*, vol. 26, no. 11-12, pp. 926–933, 2012.

[10] H. K. Vorperian and R. D. Kent, "Vowel acoustic space development in children: a synthesis of acoustic and anatomic data," *J. Speech Lang. Hear. Res.*, vol. 50, no. 6, pp. 1510, 2007.

[11] A. T. Neel, "Vowel space characteristics and vowel identification accuracy," *J. Speech Lang. Hear. Res.*, vol. 51, no. 3, pp. 574–585, 2008.

[12] S. Skodda, W. Grönheit, and U. Schlegel, "Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease," *PloS one*, vol. 7, no. 2, pp. e32132, 2012.

[13] L. B. Leonard, E. S. Weismer, C. A. Miller, D. J. Francis, J. B. Tomblin, and R. V. Kail, "Speed of processing, working memory, and language impairment in children," *J. Speech Lang. Hear. Res.*, vol. 50, no. 2, pp. 408, 2007.

[14] H. M. Liu, F. M. Tsao, and P. K. Kuhl, "The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy," *J. Acoust. Soc. Am.*, vol. 117, pp. 3879, 2005.

[15] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech," *J. Speech Lang. Hear. Res.*, vol. 53, no. 1, pp. 114, 2010.

[16] Y. I. Bang, K. Min, Y. H. Sohn, and S. R. Cho, "Acoustic characteristics of vowel sounds in patients with Parkinson disease.," *NeuroRehabilitation*, vol. 32, no. 3, pp. 649–654, 2013.

[17] K. Tjaden and G. E. Wilding, "Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings," *J. Speech Lang. Hear. Res.*, vol. 47, no. 4, pp. 766–783, Aug. 2004.

[18] P. A. McRae, K. Tjaden, and B. Schoonings, "Acoustic and perceptual consequences of articulatory rate change in Parkinson disease," *J. Speech Lang. Hear. Res.*, vol. 45, no. 1, pp. 35–50, Feb. 2002.

[19] G. S. Turner, K. Tjaden, and G. Weismer, "The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis," *J. Speech Hear. Res.*, vol. 38, no. 5, pp. 1001–1013, 1995.

[20] G. Weismer, J. Y. Jeng, J. Laures, R. D. Kent, and J. F. Kent, "Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders," *Folia Phoniatrica et Logopaedica*, vol. 53, pp. 1–18, 2001.

[21] C. M. Higgins and M. M. Hodge, "Vowel area and intelligibility in children with and without dysarthria," *J. Med. Speech Lang. Pathol.*, vol. 10, no. 4, pp. 271–278, 2002.

[22] S. Sapir, J. L. Spielman, L. O. Ramig, B. H. Story, and C. Fox, "Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on ataxic dysarthria: A case study," *J. Speech Lang. Hear. Res.*, vol. 50, no. 4, pp. 899–912, Aug. 2007.

[23] K. L. Lansford and J. M. Liss, "Vowel acoustics in dysarthria: Speech disorder diagnosis and classification," *J. Speech Lang. Hear. Res.*, vol. 57, no. 1, pp. 57–67, 2014.

[24] K. L. Lansford and J. M. Liss, "Vowel acoustics in dysarthria: Mapping to perception," *J. Speech Lang. Hear. Res.*, vol. 57, no. 1, pp. 68–80, 2014.

[25] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task," *J. Acoust. Soc. Am.*, vol. 134, pp. 2171, 2013.

[26] G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," *J. Acoust. Soc. Am.*, vol. 24, no. 2, pp. 175–184, Mar. 1952.

[27] D. O'Shaughnessy, *Speech Communication: Human and Machine*, Addison-Wesley Pub. Co., 1987.

[28] T. F. Quatieri, *Discrete-Time Speech Signal Processing*, Prentice Hall, 2002.

[29] J. B. Allen and L. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, Nov. 1977.

[30] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.

[31] C. K. Chui and H. K. Mhaskar, "Signal decomposition and analysis via extraction of frequencies," *Appl. Comput. Harmonic Anal.*, vol. 40, no. 1, pp. 97–136, 2016.

[32] J. Shekel, " 'Instantaneous' frequency," *Proc. IRE*, vol. 41, no. 4, pp. 548–548, Apr. 1953.

[33] L. M. Fink, "Relations between the spectrum and instantaneous frequency of a signal," *Problemy Peredachi Informatsii*, vol. 2, no. 4, pp. 26–38, 1966.

[34] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. I: Fundamentals," *Proc. IEEE*, vol. 80, no. 4, pp. 520–538, Apr. 1992.

[35] D. Gabor, "Theory of communication. Part 1: The analysis of information," *J. Inst. Electr. Eng. 3*, vol. 93, no. 26, pp. 429–441, Nov. 1946.

[36] M. Feldman, "Non-linear system vibration analysis using Hilbert transform–I. Free vibration analysis method 'FREEVIB'," *Mech. Syst. and Signal Process.*, vol. 8, no. 2, pp. 119–127, Mar. 1994.

[37] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 240–254, May 2000.

[38] F. Gianfelici, G. Biagetti, P. Crippa, and C. Turchetti, "Multicomponent AM–FM representations: An asymptotically exact approach," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 3, pp. 823–837, Mar. 2007.

[39] M. Feldman, *Hilbert Transform Applications in Mechanical Vibration*, Wiley, 2011.

[40] R. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[41] P. Rao and F. J. Taylor, "Estimation of instantaneous frequency using the discrete Wigner distribution," *Electron. Lett.*, vol. 26, no. 4, pp. 246–248, Feb. 1990.

[42] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM–FM signal decomposition with application to speech analysis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 2, pp. 290–300, Feb. 2011.

[43] B. Boashash, G. Azemi, and J. O'Toole, "Time-frequency processing of nonstationary signals: Advanced TFD design to aid diagnosis with highlights from medical applications," *IEEE Signal Process. Mag.*, vol. 30, no. 6, pp. 108–119, Nov. 2013.

[44] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993.

[45] A. C. Bovik, P. Maragos, and T. F. Quatieri, "AM–FM energy detection and separation in noise using multiband energy operators," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3245–3265, Dec. 1993.

[46] L. B. Fertig and J. H. McClellan, "Instantaneous frequency estimation using linear prediction with comparisons to the DESAs," *IEEE Signal Process. Lett.*, vol. 3, no. 2, pp. 54–56, Feb. 1996.

[47] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an AM–FM modulation model," *Speech Commun.*, vol. 28, no. 3, pp. 195–209, Jul. 1999.

[48] A. O. Boudraa, J. C. Cexus, F. Salzenstein, and L. Guillon, "IF estimation using empirical mode decomposition and nonlinear Teager energy operator," in *Int. Symp. Cont., Comm. Signal Process.*, 2004, pp. 45–48.

[49] A. O. Boudraa, "Instantaneous frequency estimation of FM signals by ψ B-energy operator," *Electron. Lett.*, vol. 47, no. 10, pp. 623–624, 2011.

[50] T. F. Quatieri, T. E. Hanna, and G. C. O'Leary, "AM–FM separation using auditory-motivated filters," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 465–480, Sep. 1997.

[51] F. Gianfelici, C. Turchetti, and P. Crippa, "Multicomponent AM–FM demodulation: the state of the art after the development of the iterated Hilbert transform," in *Proc. Int. Conf. Sig. Process. and Commun.*, Nov. 2007, pp. 1471–1474.

[52] S. Sandoval, V. Berisha, R. Utianski, J. Liss, and A. Spanias, "Automatic assessment of vowel space area," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. EL477–EL483, 2013.

[53] S. Sandoval, R. Utianski, V. Berisha, J. Liss, and A. Spanias, "Feature divergence of pathological speech," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. 4133–4133, 2013.

[54] R. Utianski, S. Sandoval, N. Lehrer, V. Berisha, and J. Liss, "Speech assist: An augmentative tool for practice in speech-language pathology.," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. 4134–4134, 2013.

[55] R. Utianski, S. Sandoval, V. Berisha, and J. Liss, "The effects of speech compression algorithms on the intelligibility of dysarthric speech," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. 4132–4132, 2013.

[56] V. Berisha, S. Sandoval, R. Utianski, J. Liss, and A. Spanias, "Selecting disorder-specific features for speech pathology fingerprinting," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 7562–7566.

[57] V. Berisha, S. Sandoval, R. Utianski, J. Liss, and A. Spanias, "Characterizing the distribution of the quadrilateral vowel space area," *J. Acoust. Soc. Am.*, vol. 135, no. 1, pp. 421–427, 2014.

[58] S. Sandoval and R. L. Utianski, "Average formant trajectories," in review.

[59] D. E. Vakman, "On the definition of concepts of amplitude phase and instantaneous frequency," *Radio Eng. and Electron. Phys.*, vol. 17, pp. 754–759, 1972.

[60] D. E. Vakman, "Do we know what are the instantaneous frequency and instantaneous amplitude of a signal," *Radio Eng. and Electron. Phys.*, vol. 21, pp. 95–100, 1976.

[61] D. E. Vakman, "Measuring the frequency of an analytical signal," *Radio Eng. and Electron. Phys.*, vol. 24, pp. 63–69, 1979.

[62] D. E. Vakman and L. A. Vainshtein, "Amplitude, phase, frequency–fundamental concepts in the theory of oscillations," *Uspekhi Fizicheskikh Nauk*, vol. 123, pp. 657–682, 1977.

[63] D. Vakman, *Signals, Oscillations and Waves*, Artech House, 1998.

[64] R. Shankar, *Fundamentals of Physics: Mechanics, Relativity and Thermodynamics*, Yale University Press, 2014.

[65] L. Kinsler, A. Frey, A. Coppens, and J. Sanders, *Fundamentals of Acoustics*, Wiley Publishing, 3rd edition, 1982.

[66] M. B. Priestley, *Non-linear and Non-stationary Time Series Analysis*, Academic Press, 1988.

[67] B. Boashash, *Time Frequency Signal Analysis and Processing*, Elsevier, 2003.

[68] A. Papandreou-Suppappola, *Applications in Time-Frequency Signal Processing*, CRC press, 2002.

[69] R. L. Allen and D. Mills, *Signal Analysis: Time, Frequency, Scale, and Structure*, John Wiley & Sons, 2004.

[70] 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, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. London Ser. A*, vol. 454, no. 1971, pp. 903–995, Mar. 1998.

[71] S. Sandoval and P. L. De Leon, "Theory of the Hilbert spectrum," *arXiv*, Apr. 2015, math.cv/1504.07554.

[72] S. Sandoval, P. L. De Leon, and J. M. Liss, "Hilbert spectral analysis of vowel using intrinsic mode functions," in *IEEE Workshop Automatic Speech Recognition and Understanding Workshop*, 2011, pp. 1–5.

[73] "ITU-T P.800. Methods for subjective determination of transmission quality - Series P: telephone transmission quality; methods for objective and subjective assessment of quality," Aug. 1996.

[74] "International Telecommunication Union," `http://www.itu.int/`, 2016.

[75] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Recommendation P.863, International Telecommunication Union, Telecommunication Standardization Sector, 2001.

[76] "PESQ," `http://www.pesq.org/`, 2010.

[77] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II psychoacoustic model," *J. Audio Eng. Soc.*, 1998.

[78] J. Liss, S. LeGendre, and A. Lotto, "Discriminating dysarthria type from envelope modulation spectra," *J. Speech Lang. Hear. Res.*, vol. 53, no. 5, pp. 1246–1255, 2010.

[79] T. Falk and W. Chan and F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Commun.*, vol. 54, no. 5, pp. 622–631, 2012.

[80] M. De Bodt and M. Huici and P. Van De Heyning, "Intelligibility as a linear combination of dimensions in dysarthric speech," *J. Commun. Disord.*, vol. 35, no. 3, pp. 283–292, May-Jun.

[81] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 1982, vol. 7, pp. 1278–1281.

[82] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of the modified bark spectral distortion as an objective speech quality measure," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 1998, vol. 1, pp. 541–544.

[83] S. Voran, "Objective estimation of perceived speech quality. II. Evaluation of the measuring normalizing block technique," *IEEE Trans. Audio Speech Lang. Process.*, vol. 7, no. 4, pp. 383–390, Jul. 1999.

[84] "Single-sided speech quality measure," Recommendation P.563, International Telecommunication Union, Telecommunication Standardization Sector, 2004.

[85] J. Liss, M. Spitzer, J. Caviness, and C. Adler, "The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria," *J. Acoust. Soc. Am.*, vol. 112, pp. 3022–3030, 2002.

[86] S. Borri and M. McAuliffe and J. Liss, "Perceptual learning of dysarthric speech: A review of experimental studies," *J. Speech Lang. Hear. Res.*, pp. 290–305, 2012.

[87] M. McHenry, "An exploration of listener variability in intelligibility judgments," *Am. J. Speech Lang. Pathol.*, vol. 20, no. 2, pp. 119–123, 2011.

[88] C. Sheard, R. Adams, and P. Davis, "Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility," *J. Speech Hear. Res.*, vol. 34, no. 2, pp. 285–293, 1991.

[89] P. Enderby, "Frenchay dysarthria assessment," *Int. J. Lang. Commun. Disord.*, vol. 15, no. 3, pp. 165–173, 1980.

[90] A. Tsanas, M. Little, P. McSharry, and L. Ramig, "Using the cellular mobile telephone network to remotely monitor Parkinsons disease symptom severity," *IEEE Trans. Biomed. Eng.*, (submitted) 2012.

[91] "CMU Sphinx," `http://cmusphinx.sourceforge.net`, 2013.

[92] "HTK," `http://htk.eng.cam.ac.uk`, 2013.

[93] "Kaldi," `http://http://kaldi-asr.org/`, 2015.

[94] B. P. Bogert, J. R. Healy, and J. W. Tukey, "The quefrency analysis of time series for echoes: Cepstrum, pseudo autocovariance, cross-cepstrum and saphe cracking," in *Proc. Symp. Time Series Analysis*, 1963.

[95] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[96] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[97] S. S. Stevens and J. Volkmann, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Am.*, vol. 8, pp. 185–190, Jan. 1937.

[98] L. E. Boucheron, P. L. De Leon, and S. Sandoval, "Low bit-rate speech coding through quantization of Mel-frequency cepstral coefficients," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 2, pp. 610–619, Feb. 2012.

[99] L. E. Boucheron, P. L. De Leon, and S. Sandoval, "Hybrid scalar/vector quantization of Mel-frequency cepstral coefficients for low bit-rate coding of speech," in *Proc. Data Compress. Conf.*, Mar. 2011, pp. 103–112.

[100] E. Mendoza, N. Valencia, J. Muoz, and H. Trujillo, "Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS)," *J. Voice*, vol. 10, no. 1, pp. 59–66, 1996.

[101] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High quality speech at very low bit rates," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1985, pp. 937–940.

[102] A. V. McCree and T. P. Bamwell, "Mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 242–250, Jun. 1995.

[103] "Analog-to-digital conversion of voice by 2400 bits per second mixed excitation linear prediction," US MIL-STD-3005, United States Military, Dec. 1999.

[104] "The 1200 and 2400 bit/s NATO interoperable narrow band voice coder," STANAG 4591 Ratification Draft 1, North Atlantic Treaty Organization, Dec. 1999.

[105] M. Chamberlain, "A 600 bps MELP vocoder for use on HF channels," in *Proc. IEEE MILCOMM Conf.*, 2001.

[106] P. L. De Leon, *EE589 Class Lecures*, New Mexico State University, 2008.

[107] "Coding of speech at 16 kbit/s using low-delay code excited linear prediction," ITU - Recommendation G.728, European Telecommunications Standards Institute, Sep. 1992.

[108] J. P. Campbell, T. E. Tremain, and V. C. Welch, "The federal standard 1016 4800 bps CELP voice coder," *Digit. Signal Process.*, vol. 1, no. 3, pp. 145–155, 1991.

[109] G. Fant, *Speech Sounds and Features*, The MIT Press, 1973.

[110] A. Bladon, "Two-formant models of vowel perception: Shortcomings and enhancement," *Speech Commun.*, vol. 2, no. 4, pp. 305–313, 1983.

[111] R. L. Diehl, B. Lindblom, K. A. Hoemeke, and R. P. Fahey, "On explaining certain male-female differences in the phonetic realization of vowel categories," *J. Phon.*, vol. 24, no. 2, pp. 187–208, 1996.

[112] R. A. Fox, "Perceptual structure of monophthongs and diphthongs in English," *Lang. and Speech*, vol. 26, no. 1, pp. 21–60, 1983.

[113] E. Jacewicz and R. A. Fox, "Dialectal and age-related acoustic variation in vowels in spontaneous speech.," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 2002, 2012.

[114] J. Lam, K. Tjaden, and G. Wilding, "Acoustics of clear speech: Effect of instruction," *J. Speech Lang. Hear. Res.*, vol. 55, no. 6, pp. 1807, 2012.

[115] C. G. Clopper, D. B. Pisoni, and K. de Jong, "Acoustic characteristics of the vowel systems of six regional varieties of American English," *J. Acoust. Soc. Am.*, vol. 118, no. 3 Pt 1, pp. 1661–76, 2005.

[116] P. Flipsen and S. Lee, "Reference data for the American English acoustic vowel space," *Clin. Linguist. Phon.*, vol. 26, no. 11-12, pp. 926–933, 2012.

[117] N. Flynn, "Comparing vowel formant normalisation procedures," *York Papers in Linguistics Series*, vol. 2, no. 11, pp. 1–28, 2011.

[118] J. Lee and S. Shaiman, "Relationship between articulatory acoustic vowel space and articulatory kinematic vowel space.," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 2003, 2012.

[119] A. R. Bradlow and T. Bent, "The clear speech effect for non-native listeners.," *J. Acoust. Soc. Am.*, vol. 112, no. 1, pp. 272–284, Jul. 2002.

[120] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proceedings of DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[121] P. Boersma, "Praat, a system for doing phonetics by computer.," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[122] D. G. Childers and S. B. Kesler, *Modern Spectrum Analysis*, vol. 331, IEEE Press New York, 1978.

[123] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, New York, NY, USA, 2nd edition, 1992.

[124] C. Bishop, *Pattern recognition and machine learning*, springer, 2006.

[125] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Berkeley Symp. Math. Stat. Prob.*, L. M. Le Cam and J. Neyman, Eds. 1967, vol. 1, pp. 281–297, University of California Press.

[126] F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*, Springer, 1985.

[127] MATLAB, *version 8.0.0.783 (R2012b)*, The MathWorks Inc., Natick, Massachusetts, 2012.

[128] Carol Jean Colby, Rex Wallace, and Catherine Jolly, Eds., *Language Files*, Ohio State University Press, 1982.

[129] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proc. R. Soc. Lond.*, vol. 58, pp. 240–242, 1895.

[130] V. Berisha, R. Utianski, and J. Liss, "Towards a clinical tool for automatic intelligibility assessment," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 2825–2828.

[131] C. I. Watson and J. Harrington, "Acoustic evidence for dynamic formant trajectories in Australian English vowels," *J. Acoust. Soc. Am.*, vol. 106, no. 1, pp. 458–468, 1999.

[132] T. M. Nearey, *Phonetic Feature Systems for Vowels*, vol. 77, Indiana University Linguistics Club, 1978.

[133] T. M. Nearey, J. Hogan, and A. Rozsypal, "Speech signals, cues and features," *Perspect. Exp. Linguist.*, pp. 73–96, 1979.

[134] A. K. Syrdal, "Aspects of a model of the auditory representation of American English vowels," *Speech Commun.*, vol. 4, no. 1, pp. 121–135, 1985.

[135] A. K. Syrdal and H. S. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *J. Acoust. Soc. Am.*, vol. 79, no. 4, pp. 1086–1100, 1986.

[136] R. P. Lippmann, "Review of neural networks for speech recognition," *Neural Comput.*, vol. 1, no. 1, pp. 1–38, 1989.

[137] J. D. Miller, "Auditory-perceptual interpretation of the vowel," *J. Acoust. Soc. Am.*, vol. 85, no. 5, pp. 2114–2134, 1989.

[138] T. M. Nearey, "Applications of generalized linear modeling to vowel data," in *Int. Conf. Spoken Lang. Process.*, 1992.

[139] J. Hillenbrand and R. T. Gayvert, "Vowel classification based on fundamental frequency and formant frequencies," *J. Speech Lang. Hear. Res.*, vol. 36, no. 4, pp. 694–700, 1993.

[140] K. McDougall and F. Nolan, "Discrimination of speakers using the formant dynamics of /u:/ in British English," in *Proc. Int. Congr. Phonetic Sci.*, 2007, pp. 1825–1828.

[141] P. Delattre, A. M. Liberman, F. S. Cooper, and L. J. Gerstman, "An experimental study of the acoustic determinants of vowel color; observations on one-and two-formant vowels synthesized from spectrographic patterns," *Word*, 1952.

[142] W. Klein, R. Plomp, and L. C. Pols, "Vowel spectra, vowel spaces, and vowel identification," *J. Acoust. Soc. Am.*, vol. 48, no. 4, pp. 999–1009, 1970.

[143] K. N. Stevens, S. Kasowski, and G. Fant, "An electrical analog of the vocal tract," *J. Acoust. Soc. Am.*, vol. 25, no. 4, pp. 734–742, 1953.

[144] G. Fant, *Acoustic Theory of Speech Production*, Gravenhage: Mouton and Co, 1960.

[145] P. Ladefoged, *Three Areas of Experimental Phonetics: Stress and Respiratory Activity, the Nature of Vowel Quality, Units in the Perception and Production of Speech*, vol. 15, Oxford University Press, 1972.

[146] C. Essner, "Recherche sur la structure des voyelles orales," *Archives Néerlandaises de Phonétique Expérimentale*, vol. 20, pp. 40–77, 1947.

[147] M. Joos, "Acoustic phonetics," *Lang.*, vol. 24, no. 2, pp. 5–136, 1948.

[148] D. C. Bennett, "Spectral form and duration as cues in the recognition of English and German vowels," *Lang. Speech*, vol. 11, no. 2, pp. 65–85, 1968.

[149] W. A. Ainsworth, "Duration as a cue in the recognition of synthetic vowels," *J. Acoust. Soc. Am.*, vol. 51, no. 2B, pp. 648–651, 1972.

[150] J. J. Jenkins, W. Strange, and T. R. Edman, "Identification of vowels in "vowelless" syllables," *Percept. Psychophys.*, vol. 34, no. 5, pp. 441–450, 1983.

[151] T. M. Nearey, "Static, dynamic, and relational properties in vowel perception," *J. Acoust. Soc. Am.*, vol. 85, no. 5, pp. 2088–2113, 1989.

[152] W. Strange, J. J. Jenkins, and T. L. Johnson, "Dynamic specification of coarticulated vowels," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 695–705, 1983.

[153] T. M. Nearey and P. F. Assmann, "Modeling the role of inherent spectral change in vowel identification," *J. Acoust. Soc. Am.*, vol. 80, no. 5, pp. 1297–1308, 1986.

[154] M.-G. Di Benedetto, "Vowel representation: Some observations on temporal and spectral properties of the first formant frequency," *J. Acoust. Soc. Am.*, vol. 86, no. 1, pp. 55–66, 1989.

[155] W. Strange, "Dynamic specification of coarticulated vowels spoken in sentence context," *J. Acoust. Soc. Am.*, vol. 85, no. 5, pp. 2135–2153, 1989.

[156] D. H. Whalen, "Vowel and consonant judgments are not independent when cued by the same information," *Percept. Psychophys.*, vol. 46, no. 3, pp. 284–292, 1989.

[157] J. Hillenbrand and R. T. Gayvert, "Identification of steady-state vowels synthesized from the peterson and barney measurements," *J. Acoust. Soc. Am.*, vol. 94, no. 2, pp. 668–674, 1993.

[158] I. Lehiste and G. Peterson, "Some basic considerations in the analysis of intonation," *J. Acoust. Soc. Am.*, vol. 33, no. 4, pp. 419–425, 1961.

[159] C. B. Huang, "The effect of formant trajectory and spectral shape on the tense/lax distinction in American vowels," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 1986, vol. 11, pp. 893–896.

[160] W. Strange, "Evolving theories of vowel perception," *J. Acoust. Soc. Am.*, vol. 85, no. 5, pp. 2081–2087, 1989.

[161] J. R. Bernard, "Australian pronunciation," *The Macquarie Dictionary*, vol. 18, pp. 27, 1981.

[162] F. Cox, *An Acoustic Study of Vowel Variation in Australian English*, Ph.D. thesis, Macquarie University, 1996.

[163] F. Cox, "The Bernard data revisited," *Aust. J. Linguis.*, vol. 18, no. 1, pp. 29–55, 1998.

[164] J. Harrington and S. Cassidy, "Dynamic and target theories of vowel classification: Evidence from monophthongs and diphthongs in Australian English," *Lang. Speech*, vol. 37, no. 4, pp. 357–373, 1994.

[165] J. Harrington, F. Cox, and Z. Evans, "An acoustic analysis of cultivated, general and broad Australian English speech," *Aust. J. Linguis.*, vol. 17, pp. 155–84, 1997.

[166] C. B. Huang, "Modelling human vowel identification using aspects of formant trajectory and context," *Speech perception, production and linguistic structure*, pp. 43–61, 1992.

[167] S. A. Zahorian and J. A. Jagharghi, "Spectral-shape features versus formants as acoustic correlates for vowels," *J. Acoust. Soc. Am.*, vol. 94, no. 4, pp. 1966–1982, 1993.

[168] A. T. Neel, "Formant detail needed for vowel identification," *Acoust. Res. Lett. Online*, vol. 5, no. 4, pp. 125–131, 2004.

[169] J. M. Hillenbrand, "Static and dynamic approaches to vowel perception," in *Vowel inherent spectral change*, pp. 9–30. Springer, 2013.

[170] T. R. William, "Vowel recognition as a function of duration, frequency modulation and phonetic context," *J. Speech Hear. Disord.*, vol. 18, no. 3, pp. 289–301, 1953.

[171] G. S. Morrison and P. F. Assmann, *Vowel inherent spectral change*, Springer Science & Business Media, 2012.

[172] T. M. Nearey, "Vowel inherent spectral change in the vowels of North American English," in *Vowel Inherent Spectral Change*, pp. 49–85. Springer, 2013.

[173] G. S. Morrison, "Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs," *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2387–2397, 2009.

[174] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, no. 3, pp. 971–995, 1980.

[175] P. F. Assmann and W. F. Katz, "Time-varying spectral change in the vowels of children and adults," *J. Acoust. Soc. Am.*, vol. 108, no. 4, pp. 1856–1866, 2000.

[176] D. J. Broad and F. Clermont, "Linear scaling of Vowel-Formant Ensembles (VFEs) in consonantal contexts," *Speech Commun.*, vol. 37, no. 3, pp. 175–195, 2002.

[177] D. Kewley-Port and A. Neel, "Perception of dynamic properties of speech: Peripheral and central processes," *Listening to Speech: An Auditory Perspective*, p. 49, 2006.

[178] D. J. Broad and F. Clermont, "Target–locus scaling methods for modeling families of formant transitions," *J. Phon.*, vol. 38, no. 3, pp. 337–359, 2010.

[179] R. A. Fox and E. Jacewicz, "Cross-dialectal variation in formant dynamics of American English vowels," *J. Acoust. Soc. Am.*, vol. 126, no. 5, pp. 2603–2618, 2009.

[180] D. J. Broad and R. H. Fertig, "Formant-frequency trajectories in selected CVC-syllable nuclei," *J. Acoust. Soc. Am.*, vol. 47, no. 6B, pp. 1572–1582, 1970.

[181] D. J. Broad and F. Clermont, "A methodology for modeling vowel formant contours in CVC context," *J. Acoust. Soc. Am.*, vol. 81, no. 1, pp. 155–165, 1987.

[182] MATLAB, *version 8.3.0.532 (R2014a)*, The MathWorks, Inc., Natick, Massachusetts, 2014.

[183] M. Pettinato, O. Tuomainen, S. Granlund, and V. Hazan, "Vowel space area in later childhood and adolescence: Effects of age, sex and ease of communication," *J. Phonetics*, vol. 54, pp. 1–14, 2016.

[184] "Steven Sandoval's Homepage," `http://StevenSandoval.info`, 2016.

[185] M. S. Gupta, "Definition of instantaneous frequency and frequency measurability," *Am. J. of Phys.*, vol. 43, no. 12, pp. 1087–1088, 1975.

[186] P. J. Loughlin and B. Tacer, "On the amplitude- and frequency-modulation decomposition of signals," *J. Acoust. Soc. Am.*, vol. 100, no. 3, pp. 1594–1601, 1996.

[187] P. J. Loughlin and B. Tacer, "Comments on the interpretation of instantaneous frequency," *IEEE Signal Process. Lett.*, vol. 4, no. 5, pp. 123–125, 1997.

[188] W. Nho and P. J. Loughlin, "When is instantaneous frequency the average frequency at each time?," *IEEE Signal Process. Lett.*, vol. 6, no. 4, pp. 78–80, 1999.

[189] L. Cohen, "What is a multicomponent signal?," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1992, vol. 5, pp. 113–116.

[190] J. R. Carson and T. C. Fry, "Variable frequency electric circuit theory," *Bell System Technical Journal*, vol. 16, pp. 513–540, Oct. 1937.

[191] J. Ville, "Theorie et applications de la notion de signal analytique," *Cables et Transmission*, vol. 2a, pp. 61–74, 1948.

[192] J. R. Carson, "Notes on the theory of modulation," *Proc. IRE*, vol. 10, no. 1, pp. 57–64, Feb. 1922.

[193] P. J. Loughlin, "Do bounded signals have bounded amplitudes?," *Multidim. Syst. Sig. Proc.*, vol. 9, no. 4, pp. 419–424, Oct. 1998.

[194] L. Mandel, "Interpretation of instantaneous frequencies," *Am. J. of Phys.*, vol. 42, no. 10, pp. 840–846, 1974.

[195] B. Van Der Pol, "The fundamental principles of frequency modulation," *J. Inst. Electr. Eng. 3*, vol. 93, no. 23, pp. 153–158, May 1946.

[196] J. Hurpert, "Instantaneous frequency," *Proc. IRE*, vol. 41, pp. 1188–1188, Sep. 1953.

[197] A. W. Rihaczek and E. Bedrosian, "Hilbert transforms and the complex representation of real signals," *Proc. IEEE*, vol. 54, no. 3, pp. 434–435, Mar. 1966.

[198] D. Vakman, "On the analytic signal, the Teager–Kaiser energy algorithm and other methods for defining amplitude and frequency," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 791–797, 1996.

[199] E. C. Titchmarch, *Theory of Fourier Integrals*, Oxford University Press, 2nd edition, 1948.

[200] C. W. Therrien, "The Lee-Wiener legacy [statistical theory of communication]," *IEEE Signal Process. Mag.*, vol. 19, no. 6, pp. 33–34, Nov. 2002.

[201] J. W. Brown and R. V. Churchill, *Complex Variables and Applications*, McGraw-Hill Higher Education, 2009.

[202] L. Cohen, P. Loughlin, and D. Vakman, "On an ambiguity in the definition of the amplitude and phase of a signal," *Signal Process.*, vol. 79, no. 3, pp. 301–307, 1999.

[203] E. Bedrosian, "The analytic signal representation of modulated waveforms," *Proc. IRE*, vol. 50, no. 10, pp. 2071–2076, 1962.

[204] X. G. Xia and L. Cohen, "On analytic signals with nonnegative instantaneous frequency," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1999, pp. 1329–1332.

[205] B. Van der Pol, "The nonlinear theory of electric oscillations," *Proc. IRE*, vol. 22, no. 9, pp. 1051–1086, 1934.

[206] S. Farlow, *Partial Differential Equations for Scientists and Engineers*, Courier Corporation, 2012.

[207] E. Bedrosian, "A product theorem for Hilbert transforms," *Proc. IEEE*, vol. 51, no. 5, pp. 868–869, May 1963.

[208] A. H. Nuttall and E. Bedrosian, "On the quadrature approximation to the Hilbert transform of modulated signals," *Proc. IEEE*, vol. 54, no. 10, pp. 1458–1459, Oct. 1966.

[209] B. Picinbono, "On instantaneous amplitude and phase of signals," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 552–560, Mar. 1997.

[210] B. Boashash, *Time-Frequency Signal Analysis*, Longman Publishing Group, 1992.

[211] D. Wei and A. C. Bovik, "On the instantaneous frequencies of multicomponent AM–FM signals," *IEEE Signal Process. Lett.*, vol. 5, no. 4, pp. 84–86, 1998.

[212] L. Cohen and C. Lee, "Instantaneous frequency, its standard deviation and multicomponent signals," in *Proc. SPIE Int. Soc. Opt. Eng.*, Feb. 1988, pp. 186–208.

[213] J. Fourier, "On the propagation of heat in solid bodies," 1807.

[214] J. Fourier, *The Analytical Theory of Heat*, Firmin Didot, 1822.

[215] A. V. Oppenheim, A.S. Willsky, and S. H. Nawab, *Signals and Systems*, Prentice Hall, 2nd edition, 1997.

[216] L. Cohen, "Time-frequency distributions–a review," *Proc. IEEE*, vol. 77, no. 7, pp. 941–981, 1989.

[217] G. Jones and B. Boashash, "Instantaneous frequency, instantaneous bandwidth and the analysis of multicomponent signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1990, pp. 2467–2470.

[218] L. Cohen, "Instantaneous 'anything' ," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1993, pp. 105–108.

[219] B. C. Lovell, R. C. Williamson, and B. Boashash, "The relationship between instantaneous frequency and time-frequency representations," *IEEE Trans. Signal Process.*, vol. 41, no. 3, pp. 1458–1461, 1993.

[220] S. Qian and C. Dapang, "Joint time-frequency analysis," *IEEE Signal Process. Mag.*, vol. 16, no. 2, pp. 52–67, Mar. 1999.

[221] G. Strang, *Introduction to Linear Algebra*, Wellesley-Cambridge Press, 2009.

[222] E. H. Armstrong, "Some recent developments in the audion receiver," *Proc. IRE*, vol. 3, no. 3, pp. 215–238, 1915.

[223] H. B. Voelcker, "Toward a unified theory of modulation. I: Phase-envelope relationships," *Proc. IEEE*, vol. 54, no. 3, pp. 340–353, 1966.

[224] H. B. Voelcker, "Toward a unified theory of modulation. II: Zero manipulation," *Proc. IEEE*, vol. 54, no. 5, pp. 735–755, 1966.

[225] J. M. Chowning, "Origins of FM Synthesis," `http://www.youtube.com/watch?v=w4g92vX1YF4`, 2012.

[226] R. Bedaux, "Micro frequency modulation in sound synthesis," *J. New Music Res.*, vol. 3, no. 2, pp. 89–108, 1974.

[227] J. M. Chowning, "The synthesis of complex audio spectra by means of frequency modulation," *J. Audio Eng. Soc.*, vol. 21, no. 7, pp. 526–534, Sep. 1977.

[228] B. Boashash, P. O'Shea, and M. J. Arnold, "Algorithms for instantaneous frequency estimation: A comparative study," in *Proc. SPIE Int. Soc. Opt. Eng.*, 1990, pp. 126–148.

[229] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. II: Algorithms and applications," *Proc. IEEE*, vol. 80, no. 4, pp. 540–568, Apr. 1992.

[230] M. Feldman, "Theoretical analysis and comparison of the Hilbert transform decomposition methods," *Mech. Syst. and Signal Process.*, vol. 22, no. 3, pp. 509–519, Apr. 2008.

[231] A. Potamianos and P. Maragos, "A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation," *Signal Process.*, vol. 37, no. 1, pp. 95–120, 1994.

[232] E. S. Diop, A. O. Boudraa, and F. Salzenstein, "A joint 2D AM–FM estimation based on higher order Teager–Kaiser energy operators," *Signal, Image and Video Process.*, vol. 5, no. 1, pp. 61–68, 2011.

[233] B. Santhanam and P. Maragos, "Multicomponent AM–FM demodulation via periodicity-based algebraic separation and energy-based demodulation," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 473–490, 2000.

[234] J. A. Moorer, "The synthesis of complex audio spectra by means of discrete summation formulas," *J. Audio Eng. Soc.*, vol. 24, no. 9, pp. 717–727, Nov. 1976.

[235] C. Dodge and T. A. Jerse, *Computer Music: Synthesis, Composition and Performance*, Macmillan Library Reference, 1997.

[236] D. Borland and R. M. Taylor, "Rainbow color map (still) considered harmful.," *IEEE Trans. Visual. Comput. Graphics*, vol. 27, no. 2, pp. 14–17, Mar. 2007.

[237] M. Niccoli and S. Lynch, "A more perceptual color palette for structure maps," in *Proc. GeoConvention*, May 2012.

[238] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Adv. Adapt. Data Anal.*, vol. 1, no. 01, pp. 1–41, 2009.

[239] R. Deering and J. F. Kaiser, "The use of a masking signal to improve empirical mode decomposition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2005, pp. 485–488.

[240] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 4144–4147.

[241] R. Rato, M. D. Ortigueira, and A. Batista, "On the HHT, its problems and some solutions," *Mech. Syst. and Signal Process.*, vol. 22, no. 6, pp. 1374–1394, 2008.

[242] N. E. Huang, Z. Wu, S. R. Long, K. C. Arnold, X. Chen, and K. Blank, "On instantaneous frequency," *Adv. Adapt. Data Anal.*, vol. 1, no. 02, pp. 177–229, 2009.

[243] Z. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method," *Proc. R. Soc. London Ser. A*, vol. 460, no. 2046, pp. 1597–1611, 2004.

[244] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 112–114, 2004.

[245] P. Flandrin, P. Gonçalves, and G. Rilling, "EMD equivalent filter banks, from interpretation to applications," *Hilbert-Huang transform and its applications*, pp. 57–74, 2005.

[246] D. Mandic et al., "Filter bank property of multivariate empirical mode decomposition," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2421–2426, 2011.

[247] S. Meignen and V. Perrier, "A new formulation for empirical mode decomposition based on constrained optimization," *IEEE Signal Process. Lett.*, vol. 14, no. 12, pp. 932–935, 2007.

[248] Y. Kopsinis and S. McLaughlin, "Investigation and performance enhancement of the empirical mode decomposition method based on a heuristic search optimization approach," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 1–13, 2008.

[249] T. Y. Hou and Z. Shi, "Data-driven time–frequency analysis," *Appl. Comput. Harmonic Anal.*, vol. 35, no. 2, pp. 284–308, 2013.

[250] and D. P. Mandic D. Looney, "A machine learning enhanced empirical mode decomposition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2008, pp. 1897–1900.

[251] E. Deléchelle, J. Lemoine, and O. Niang, "Empirical mode decomposition: an analytical approach for sifting process," *IEEE Signal Process. Lett.*, vol. 12, no. 11, pp. 764–767, 2005.

[252] R. Sharpley and V. Vatchev, "Analysis of the intrinsic mode functions," *Constr. Approx.*, vol. 24, no. 1, pp. 17–47, 2006.

[253] V. Vatchev and R. Sharpley, "Decomposition of functions into pairs of intrinsic mode functions," *Proc. R. Soc. Lond. A Math. Phys. Sci.*, vol. 464, no. 2097, pp. 2265–2280, 2008.

[254] E. H. Diop, R. Alexandre, and A. O. Boudraa, "A PDE characterization of the intrinsic mode functions," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2009, pp. 3429–3432.

[255] E. S. Diop, R. Alexandre, and A. O. Boudraa, "Analysis of intrinsic mode functions: a PDE approach," *IEEE Signal Process. Lett.*, vol. 17, no. 4, pp. 398–401, 2010.

[256] S. D. El Hadji, R. Alexandre, and V. Perrier, "A PDE based and interpolation-free framework for modeling the sifting process in a continuous domain," *Adv. Comput. Math.*, vol. 38, no. 4, pp. 801–835, 2013.

[257] O. Niang, E. Deléchelle, and J. Lemoine, "A spectral approach for sifting process in empirical mode decomposition," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5612–5623, 2010.

[258] C. Damerval, S. Meignen, and V. Perrier, "A fast algorithm for bidimensional EMD," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 701–704, 2005.

[259] G. Rilling, P. Flandrin, P. Gonçalves, and J. M. Lilly, "Bivariate empirical mode decomposition," *IEEE Signal Process. Lett.*, vol. 14, no. 12, pp. 936–939, 2007.

[260] Z. Wu, N. E. Huang, and X. Chen, "The multi-dimensional ensemble empirical mode decomposition method," *Adv. Adapt. Data Anal.*, vol. 1, no. 03, pp. 339–372, 2009.

[261] A. Linderhed, "Image empirical mode decomposition: A new tool for image processing," *Adv. Adapt. Data Anal.*, vol. 1, no. 02, pp. 265–294, 2009.

[262] J. C. Nunes and E. Deléchelle, "Empirical mode decomposition: Applications on signal and image processing," *Adv. Adapt. Data Anal.*, vol. 1, no. 01, pp. 125–175, 2009.

[263] D. P. Mandic, N. U. Rehman, W. Zhaohua, and N. E. Huang, "Empirical mode decomposition-based time-frequency analysis of multivariate signals: The power of adaptive data analysis," *IEEE Signal Process. Mag.*, vol. 30, no. 6, pp. 74–86, Nov 2013.

[264] T. Tanaka and D. P. Mandic, "Complex empirical mode decomposition," *IEEE Signal Process. Lett.*, vol. 14, no. 2, pp. 101–104, 2007.

[265] M. Feldman, "Time-varying vibration decomposition and analysis based on the Hilbert transform," *J. Sound Vib.*, vol. 295, no. 3, pp. 518–530, 2006.

[266] X. Chen, Z. Wu, and N. E. Huang, "The time-dependent intrinsic correlation based on the empirical mode decomposition," *Adv. Adapt. Data Anal.*, vol. 2, no. 02, pp. 233–265, 2010.

[267] Daubechies I, J. Lu, and H. T. Wu, "Synchrosqueezed wavelet transforms: an empirical mode decomposition-like tool," *Appl. Comput. Harmonic Anal.*, vol. 30, no. 2, pp. 243–261, 2011.

[268] H. T. Wu, P. Flandrin, and I. Daubechies, "One or two frequencies? The synchrosqueezing answers," *Adv. Adapt. Data Anal.*, vol. 3, no. 01n02, pp. 29–39, 2011.

[269] T. Y. Hou and Z. Shi, "Adaptive data analysis via sparse time-frequency representation," *Adv. Adapt. Data Anal.*, vol. 3, no. 01n02, pp. 1–28, 2011.

[270] G. Rilling and P. Flandrin, "On the influence of sampling on the empirical mode decomposition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2006, pp. 444–447.

[271] R. Rato and M. Ortigueira, "A modified EMD algorithm for application in biomedical signal processing," in *Int. Conf. Comput. Intell. Med. Healthcare*, Jul. 2005.

[272] N. E. Huang, M. L. C. Wu, S. R. Long, S. S. P. Shen, W. Qu, P. Gloersen, and K. L. Fan, "A confidence limit for the empirical mode decomposition and Hilbert spectral analysis," *Proc. R. Soc. London Ser. A*, vol. 459, no. 2037, pp. 2317–2345, 2003.

[273] G. Rilling, P. Flandrin, P. Goncalves, et al., "On empirical mode decomposition and its algorithms," in *IEEE-EURASIP Workshop on Nonlinear Signal and Image Process.*, 2003, vol. 3, pp. 8–11.

[274] G. Rilling and P. Flandrin, "One or two frequencies? The empirical mode decomposition answers," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 85–95, 2008.

[275] N. Senroy and S. Suryanarayanan, "Two techniques to enhance empirical mode decomposition for power quality applications," in *Proc. IEEE Power Eng. Soc. General Meeting*, 2007, pp. 1–6.

[276] X. Guanle, W. Xiaotong, and X. Xiaogang, "Time-varying frequency-shifting signal-assisted empirical mode decomposition method for AM–FM signals," *Mech. Syst. and Signal Process.*, vol. 23, no. 8, pp. 2458–2469, 2009.

[277] X. M. Li, C. C. Bao, and M. S. Jia, "A sinusoidal audio and speech analysis/synthesis model based on improved EMD by adding pure tone," in *IEEE Int. Workshop Mach. Learn. Signal Process.*, 2011, pp. 1–5.

[278] S. D. Hawley, L. E. Atlas, and H. J. Chizeck, "Some properties of an empirical mode type signal decomposition algorithm," *IEEE Signal Process. Lett.*, vol. 17, no. 1, pp. 24–27, 2010.

[279] T. Strom, "On amplitude-weighted instantaneous frequencies," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 4, pp. 351–353, 1977.

[280] "Hilbert Spectral Analysis of Speech," `http://asru2015.HilbertSpectrum.com`, 2015.

[281] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures.," *J. Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, 1964.

[282] S. J. Orfanidis, *Introduction to Signal Processing*, Prentice-Hall, 1995.

[283] R. W. Schafer, "What is a Savitzky-Golay filter?," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 111–117, 2011.

[284] M. Feldman, "Analytical basics of the EMD: Two harmonics decomposition," *Mech. Syst. and Signal Process.*, vol. 23, no. 7, pp. 2059–2071, 2009.

[285] P. Waskito, S. Miwa, Y. Mitsukura, and H. Nakajo, "Parallelizing Hilbert-huang transform on a GPU," in *Networking and Computing (ICNC), 2010 First International Conference on*, 2010, pp. 184–190.

[286] D. Chen, D. Li, M. Xiong, H. Bao, and X. Li, "GPGPU-aided ensemble empirical-mode decomposition for EEG analysis during anesthesia," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 6, pp. 1417–1427, 2010.

[287] L. W. Chang, M. T. Lo, N. Anssari, K. H. Hsu, N. E. Huang, and W. M. Hwu, "Parallel implementation of multi-dimensional ensemble empirical mode decomposition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2011, pp. 1621–1624.

[288] P. Waskito, Y. Mitsukura, and H. Nakajo, "Evaluation of GPU-Based empirical mode decomposition for off-line analysis," *IEICE Trans. on Inf. and Syst.*, vol. 94, no. 12, pp. 2328–2337, 2011.

[289] K. Bunton and G. Weismer, "The relationship between perception and acoustics for a high-low vowel contrast produced by speakers with dysarthria," *J. Speech Lang. Hear. Res.*, vol. 44, no. 6, pp. 1215–1228, 2001.

[290] N. Senroy, S. Suryanarayanan, and P. F. Ribeiro, "An improved Hilbert–Huang method for analysis of time-varying waveforms in power quality," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1843–1850, 2007.

[291] S. Riemenschneider, B. Liu, Y. Xu, and N. E. Huang, "B-spline based empirical mode decomposition," *Interdisciplinary Math. Sciences*, vol. 5, pp. 27–56, 2005.

[292] S. Qin and Y. M. Zhong, "A new envelope algorithm of Hilbert–Huang transform," *Mech. Syst. and Signal Process.*, vol. 20, no. 8, pp. 1941–1952, 2006.

[293] Q. Chen, N. Huang, S. Riemenschneider, and Y. Xu, "A B-spline approach for empirical mode decompositions," *Adv. Comput. Math.*, vol. 24, no. 1-4, pp. 171–195, 2006.

[294] Y. Kopsinis and S. McLaughlin, "Improved EMD using doubly-iterative sifting and high order spline interpolation," *J. Adv. Signal Process.*, vol. 2008, pp. 120, 2008.

[295] F. Wang, X. Y. CHEN, F. L. Qiao, Z. Wu, and N. E. Huang, "On intrinsic mode function," *Adv. Adapt. Data Anal.*, vol. 2, no. 03, pp. 277–293, 2010.

[296] A. Bouchikhi and A. O. Boudraa, "Multicomponent AM–FM signals analysis based on EMD–B-splines ESA," *Signal Process.*, vol. 92, no. 9, pp. 2214–2228, 2012.