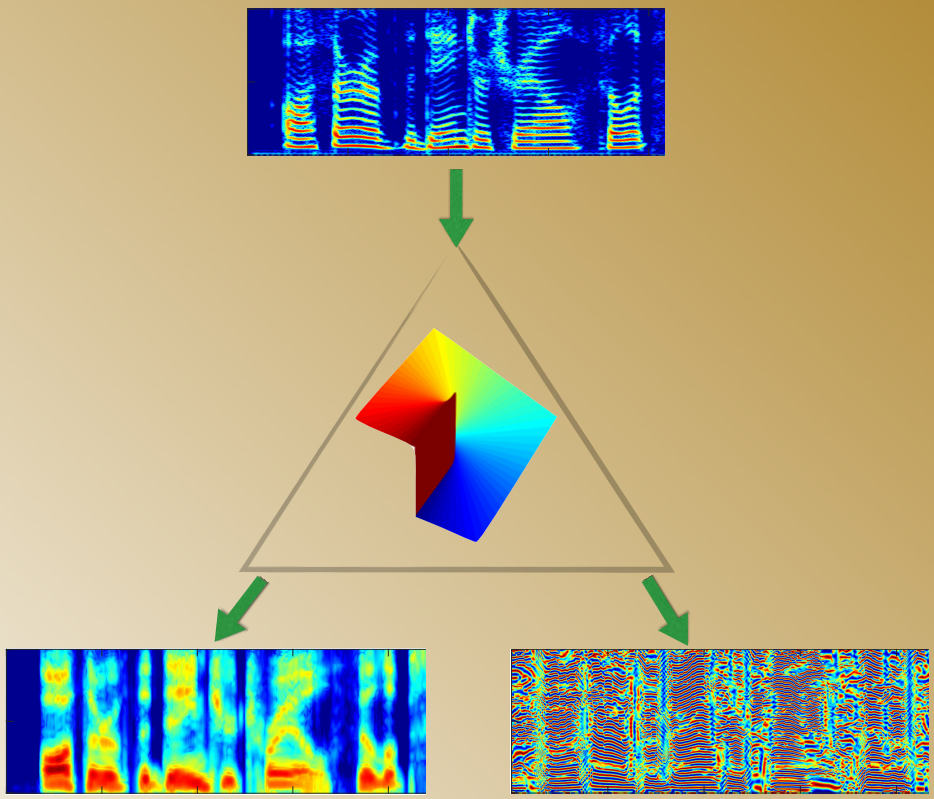
Spectrotemporal Processing of Speech Signals Using the Riesz Transform

PhD dissertation submitted to the Indian Institute of Science, Bangalore



Jitendra Kumar Dhiman

IISc

INDIAN INSTITUTE OF SCIENCE

भारतीय विज्ञान संस्थान

Abstract

In this thesis, we consider 2-D analysis of speech spectrograms. We consider a spectrotemporal patch and model it as a 2-D amplitude-modulated and frequency-modulated (AM-FM) sinusoid. Demodulation of the spectrogram yields the 2-D AM and FM, which correspond to the slowly varying vocal-tract envelope and the excitation, respectively. For solving the demodulation problem, we rely on the complex Riesz transform, which is a 2-D extension of the 1-D Hilbert transform. The demodulation viewpoint brings forth many interesting properties of the speech signal. The spectrotemporal carrier helps us identify time-frequency regions that are coherent and those that are not. Based on this idea, we introduce the coherencegram corresponding to a given spectrogram. The temporal evolution of the pitch harmonics can also be characterized by the orientation at each time-frequency coordinate resulting in the orientationgram. We show that these features collectively enable solutions for the important problems of voiced/unvoiced segmentation, time-frequency aperiodicity estimation, periodic/aperiodic component separation, and pitch tracking. The spectrotemporal amplitude characterizes the time-varying magnitude response of the vocal-tract filter. We use the spectrotemporal amplitude, pitch, aperiodicity, and voiced/unvoiced decisions for the task of speech reconstruction in a spectral synthesis model and WaveNet, which is a neural vocoder. The quality of the synthesized speech is assessed using both objective and subjective measures. We show that conditioning WaveNet on the spectrotemporal features results in high-quality speech synthesis that is on par with state-of-the-art vocoders, namely WORLD and STRAIGHT.

Author

Jitendra Kumar Dhiman received a B.Tech. degree in Electronics and Telecommunication Engineering from the Institution of Electronics and Telecommunication Engineering (IETE), Delhi in 2010, and M.Tech. degree in Signal Processing from Indian Institute of Technology Hyderabad in 2013. Subsequently, he worked as a project assistant in Spectrum Lab, Department of Electrical Engineering, Indian Institute of Science, Bangalore, where he worked on a Department of Electronics and Information Technology project on prosody modification of speech.

Subsequently, he enrolled for a Ph.D. in the same lab. His research interests include speech signal processing and machine learning. He is a recipient of Best Poster Award at the IISc EECS Divisional Symposium in 2018 and Pratiksha Trust Travel Fellowship in 2019.

SPECTRUM LAB

Department of Electrical Engineering
Indian Institute of Science, Bangalore

Spectrotemporal Processing of Speech Signals Using the Riesz Transform Jitendra K. Dhiman

Spectrotemporal Processing of Speech Signals Using the Riesz Transform

A dissertation submitted
in partial fulfillment of the
requirements for the Degree of

Doctor of Philosophy

by

Jitendra Kumar Dhiman

Department of Electrical Engineering
INDIAN INSTITUTE OF SCIENCE
Bangalore - 560012

July 2021

THESIS CERTIFICATE

This is to certify that the thesis entitled, "Spectrotemporal Processing of Speech Signals Using the Riesz Transform," submitted by me, Jitendra Kumar Dhiman, to the Indian Institute of Science, Bangalore, for the award of the degree of Doctor of Philosophy, is a bona fide record of the research done by me under the supervision of Prof. Chandra Sekhar Seelamantula. The contents of this thesis, in full or in parts, have not been submitted to any other institute or university for the award of any degree or diploma.

Jitendra Kumar Dhiman
(PhD Candidate)

Chandra Sekhar Seelamantula
(Thesis Supervisor)

Date: July 26, 2021
Place: Bangalore, India

To my parents

Acknowledgments

Today when I look back, I feel fortunate to be part of the IISc community during my Ph.D. study. The day I landed in IISc, I knew that I am going to meet amazing people here. The long research journey would not have been possible without the people whom I met here. I take this opportunity to thank them all for their unwavering support and belief in me.

First and foremost I would like to express my sincere gratitude to my advisor Prof. Chandra Sekhar Seelamantula for his invaluable supervision, support and tutelage during the course of my Ph.D. degree. I would like to thank him for his continuous support, plentiful experience, and patience, which have encouraged me in all the time of my academic research and daily life. His true understanding of my personal situation during the last phase of my Ph.D. has meant a lot to me. Without his consistent support and encouragement in the past few years, it would be impossible for me to complete my study. My gratitude extends to the Faculties of Electrical Engineering department Dr. Sriram Ganapathy, Dr. Prasanta Kumar Ghosh, and the retired faculty Prof. T.V. Sreenivas from the department of Electrical and Communication Engineering (ECE) for having several fruitful interactions with me.

I would like to thank my lab-mates and colleagues from the spectrum lab – Jishnu Sadasivan, Satish Mulleti, Shubhadeep Mukharji, Ajay Shenoy, Anirudth Adiga, Haricharan Aragonda, Sudarshan, Sriram Menon, Kishore, Abhishek, R. Sunil, Debabrata Mahapatra, Dhruv Jawali, Praveen Kumar Pokala, Amol Mahulkar, Dibakar Sil, Vinayaka Killadhar, for a cherished time spent together in the lab, and in social settings. They all have been very nice and kind to me during my stay in the lab. I share cheerful moments with Jishnu, Praveen, Amol, Vinayaka, and Debabrata. Going with them to the institute canteen for coffee breaks used to be quite refreshing, they always made time for short walks around the campus and discussions. I cherish great moments with Debabrata Mahapatra who used to stay in a nearby hostel block. Often we used to cycle together from hostel to the lab after breakfast while having some productive discussions. He was always in for weekend parties :). Amol was my closest neighbour in the lab which implicitly means that he was always available for any silly doubt, he often used to assist me for technical

writing and proofreading my documents. It is their kind help and support that have made my study and life at IISc a wonderful time. Furthermore, I would like to thank people from the EE department - Nisha Meenakshi, Vidyadhar Upadhya, Arvind Illa, Achut Rao, and Chiranjeevi Yarra. I had great pleasure to work with my colleagues and research team - Jishnu, Harshavardhan Sunder, Karthika Vijayan, and Nagaraj Adiga, Praneet, Harshawin, and Pavan Kulkarni. The technical discussions with them have always been illuminating. I would like to recognize the assistance and help received from Jishnu whose invaluable insights into various matters shaped my abilities to think differently and execute my research plans meticulously during my study. Outside the spectrum lab, I cherish unforgettable memories with several groups in IISc. Closest to me is the Speech and Audio Lab group (SAG lab) from the Department of ECE and the members of Fata-phat group - Neeraj Kumar Sharma, Sai Gunaranjan Pelluri, Arthi Subramaniam, Srikanth Raj, Ranjani, Vijay Girish, Shivashankar, Angshuman Modak, Ankur Raina, Anoop Nayak, Harikiran, and Purvi Agrawal. Their company was not only fruitful for many aspects of daily life, but was also helpful to make my stay more comfortable on campus. I would like to offer my special thanks to Neeraj Kumar Sharma for his treasured support which was greatly influential in shaping my experiment methods and critiquing my results. His intellectual advice and avid insights greatly helped to improve the quality of my research work. I will always cherish my friendship with Sai Gunaranjan who was also my neighbour in the hostel. His unwavering attitude always helped to keep my motivation high for research. Keeping the company of Neeraj, Sai, and Arthi was like going into the sunlight which spreads all around almost instantly. I would like to thank them all for their timely advice, comments and the encouragement during my study.

Life was incomplete without getting involved in sports activities at IISc. I was happy to be part of Sunday Cricket League (SCL), Table Tennis, and Volleyball teams. The fun and joyful moments with all of you will always remain special and memorable for me. I will also cherish vivid memories of my interactions with Venket Reddy, Sreekhar, Tajendra Singh from the Department of Physics, and Sanjeev Kumar Mittal from Center for Nano Science and Engineering, for all those endless philosophical discussions about various aspects of life. Their treasured company was not only useful to uplift my moral and emotional state, but also helped me to find some meaning in our discussions during the last phase of my stay in the campus:).

I would also like to thank the EE office for all the academic help I received from the staff in the office. Further, I recognise the efforts of my advisor's secretary Manjula who has provided immense help for a smooth conduct of administrative work. She is the one who is always working at the background. I would like to convey my sincere thanks to her for making my stay comfortable in the lab. Even outside the lab, her help can not be underestimated, she is a good human being. My appreciation also goes out to my family and friends (those I missed above due to my limited memory) for their encouragement and support all through my studies.

Contents

List of Figures

List of Tables

Abstract

Speech signals possess a rich time-varying spectral content, which makes their analysis a challenging signal processing problem. Developing methods for accurate speech analysis has a direct impact on applications such as speech synthesis, speaker recognition, speech recognition, voice morphing, etc. A widely used tool to visualize the time-varying spectral content is the spectrogram, which represents the spectral content of the signal in the joint time-frequency plane. A spectrogram can be viewed as a collection of several localized spectrotemporal patches. By analyzing the structure of two-dimensional (2-D) patterns in the spectrogram, we propose modeling it using 2-D amplitude-modulated and frequency-modulated (AM-FM) sinusoids. The justification for the 2-D AM-FM model for speech can be provided based on the physical process behind its generation. From a speech production perspective, the AM and FM components correspond to the vocal-tract smooth envelope and excitation signal, respectively. We demonstrate that analyzing speech jointly in time and frequency reveals several important characteristics, which are otherwise not evident either in purely time-domain or frequency-domain analysis.

The central problem in this dissertation is 2-D demodulation of a speech spectrogram, which yields 2-D AM and FM components. We advocate the use of the Riesz transform, which is a 2-D extension of the Hilbert transform, to demodulate narrowband and pitch adaptive spectrograms. Interestingly, the 2-D AM and FM components obtained as a result of demodulation have potential benefits for speech

analysis. We demonstrate the impact of the proposed modeling technique for vocal tract filter estimation, voiced/unvoiced component separation, pitch tracking, speech synthesis, and periodic/aperiodic decomposition of speech signals. The accuracy of the estimated speech parameters is validated considering the task of speech reconstruction.

The first part of the thesis is focused on theoretical developments related to 2-D modeling. We consider prototypical 2-D cosine signals, analyze their Fourier transform properties, solve the problem of demodulation of a 2-D AM-FM cosine signal and extend the model to spectrotemporal patches. Following this, we examine the taxonomy of time-frequency patterns in the FM component, highlighting the salient attributes of different types of phonation in speech. We show that 2-D patterns specific to different speech sounds (voiced/unvoiced) can be captured by computing two novel time-frequency maps from the 2-D FM component: the coherencegram and orientationgram. The usefulness of the maps is demonstrated for the problem of periodic and aperiodic decomposition of speech signals.

In the second part, we use the FM component for estimating the source parameters. We show that the FM component is a rich representation of the source signal in 2-D and use it to estimate the speaker's fundamental frequency (or pitch), speech aperiodicity, and voiced/unvoiced segmentation of the speech signal. We propose novel spectrotemporal features for voiced and unvoiced segmentation of speech. In contrast to time-domain features such as short-time energy, zero crossings, and autocorrelation coefficients, the proposed features are relatively insensitive to local variations of the speech waveform. The FM component is obtained by demodulating the narrowband speech spectrogram, which exhibits high frequency resolution. Consequently, the FM component encodes the speaker's pitch. Hence, we propose methods for estimating the pitch from the FM component. Another critical component of a speech signal is its aperiodicity. Voiced sounds are quasi-periodic and have a noise component of strength relatively weaker than unvoiced sounds.

Utilizing the time-frequency properties of the FM component, we propose methods for the estimation of speech aperiodicity.

While the FM component is used to estimate the source parameters, the 2-D AM component models the slowly varying vocal-tract filter. However, estimation of the vocal-tract filter is challenging due to its interaction with the quasi-periodic excitation. Two issues arise in this context: the first one is related to the length of the analysis window used for computing the spectrogram. We argue that a fixed-length analysis window is not ideal for vocal tract estimation. We show that the best results can be obtained by adapting the window length to the speaker's pitch while computing the spectrogram. Such a spectrogram is referred to as the *pitch-adaptive spectrogram*. The second issue is related to the processing involved in demodulation, which has the undesirable effect of broadening the formant bandwidths. Hence, we propose a method to compensate for the formant broadening. It is crucial to estimate the optimum formant bandwidths as they determine the shape of the vocal tract filter and govern speech intelligibility during synthesis.

The effectiveness of the estimated source and filter parameters is shown by incorporating them in a spectral synthesis model and a neural vocoder for speech reconstruction. For neural vocoder, we use WaveNet, which is a deep generative model for audio generation. By conditioning the model on acoustic features, one can guide WaveNet to produce realistic speech waveforms. We use the Riesz transform-based acoustic features as conditional features in WaveNet vocoder. The quality of generated speech waveforms is evaluated by using objective and subjective measures.

Chapter 1

Introduction

Speech communication is one of the most effective ways of communication in our everyday life. Humans produce speech by expelling air from the lungs through the layrnx and controlling the vocal-fold vibrations and vocal-tract shape. In the process, the acoustic energy gets modulated and comes out in the form of speech. The modulations carry a distinct acoustic signature and are important to analyze in numerous applications such as speech synthesis, automatic speech recognition, speech enhancement, design of psychoacoustic stimuli, speech coding, etc. Speech is a nonstationary signal, which also means that the modulations vary with time in a seamless fashion. The fundamental problem of studying the modulations constitutes the *speech analysis* task and the nonstationarity makes the analysis challenging. One typically resorts to the assumption of *quasi-stationarity*, i.e., the signal is assumed to be relatively stationary over short durations (typically, 20 to 30 ms long). Processing of speech signals over short time windows is referred to as *short-time analysis*. Consecutive windowed segments are overlapped in the analysis to ensure temporal continuity. Computing Fourier transforms of windowed segments results in the *short-time Fourier transform (STFT)*.

The STFT of a signal $s(t)$ is given by

$$S(t,\omega) = \int_{-\infty}^{\infty} s(\tau) w(\tau - t) e^{\mathrm{j}\omega\tau} \mathrm{d}\tau, \tag{1.1}$$

where $w(t)$ denotes the analysis window, t denotes the time instant at which the

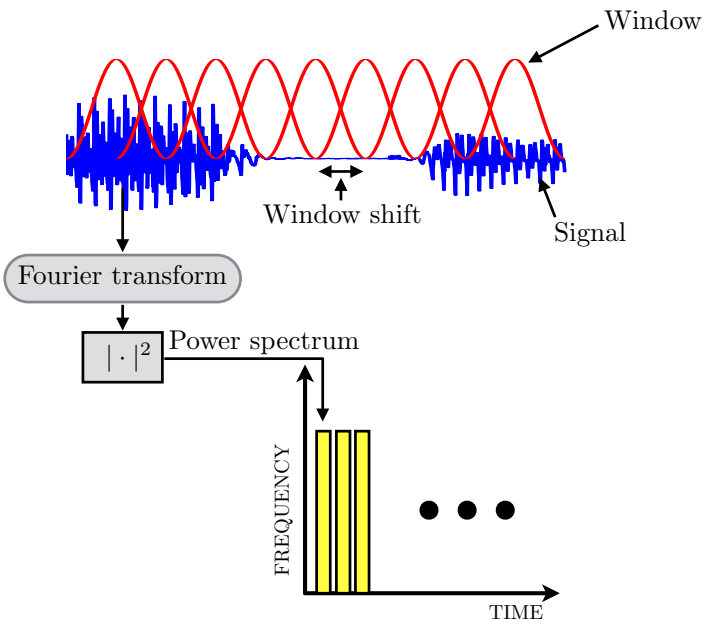
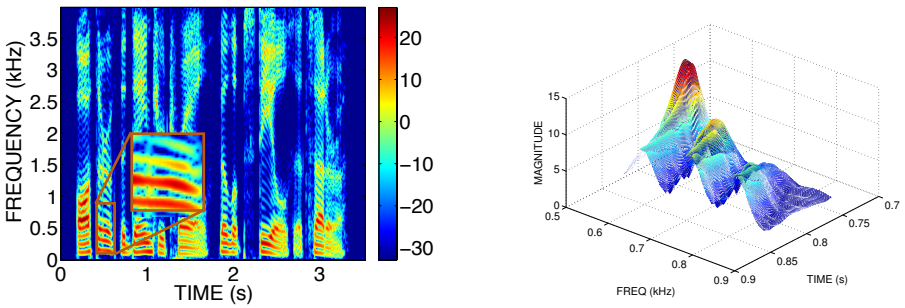


Figure 1.1: Illustration of the short-time processing of a speech signal.

window is centered, and ω denotes the digital frequency. The magnitude-square of STFT, i.e., $|S(t,\omega)|^2$ is the power spectrum of the windowed signal at time t. The STFT can be computed for a fixed hop duration ΔT at time instants $t = k\Delta T, k = 1, 2, 3, \ldots$ and the corresponding power spectra stacked over time yields a t-f representation known as the *spectrogram*. Figure 1.1 illustrates the spectrogram computation. Depending on the length of analysis window, the spectrogram is usually available in two flavors *wideband* or *narrowband*. A short window gives a wideband spectrogram and a long one results in a narrowband spectrogram. A short window gives better temporal resolution and a long window results in better spectral resolution. One could also vary the window length as a function of time. Features derived from the spectrogram have been deployed for various applications such as speech activity detection [3,4], language identification [5,6], speaker identification [7], and speech recognition [8]. In this thesis, the narrowband spectrogram is of particular interest and is the default meaning of the term spectrogram. Figure 1.2(a) shows a spectrogram. A zoomed-in voiced spectrotemporal patch is shown in Figure 1.2(b).



(a) (b)

Figure 1.2: (a) A narrowband spectrogram along with a zoomed in voiced spectrotemporal patch; and (b) 3-D view of the voiced spectrotemporal patch. The patch contains amplitude and frequency modulations of speech.

The spectrogram has been widely used mainly for visual representation and understanding of the speech signal. Speech analysis is largely about picking the right kind of modulations for the task at hand. One popular approach to analyze modulations in the spectrogram is referred to as *modulation filtering*, which consists of filtering the temporal trajectories of the short-time spectrum of speech [9]. Slow temporal modulations ($<$ 10 Hz) are associated with the syllable rate in speech, while fast and intermediate modulations ($>$ 10 Hz) capture the segmental transitions such as onsets and offsets [10]. Modulation filtering aims at removing the spectral content (usually due to noise) that changes slower or faster than the speech spectrum. Humans are most sensitive to modulation frequencies in the range 4-16 Hz, which also coincides with the rate of phoneme production. The modulation spectrum has been shown to be important in human speech recognition [11,12]. In the literature, various approaches have been developed for the estimation of modulation spectra [13,14], which benefited applications such as speaker separation, audio fingerprinting, and content identification [15–17]. In contrast, the modulations that we shall consider in this thesis are of a different variety – we consider spectrotemporal modulations embedded in the spectrogram and propose robust techniques to estimate them from the spectrogram. We do not confine ourselves to short-time processing. On the

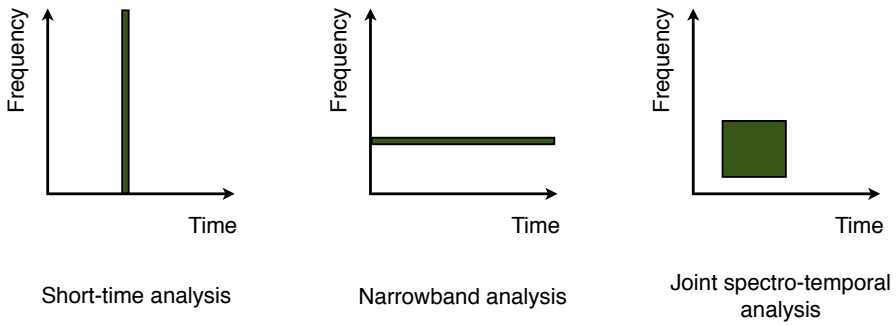


Figure 1.3: Illustration of different approaches for speech analysis.

contrary, we analyze the speech signal over large time-frequency regions, which has received relatively little interest compared with the short-time approaches. A fixed approach does not give access to all the modulations. Depending on the target modulations, speech processing algorithms can be broadly classified as short-time analysis approaches, narrowband approaches, or spectrotemporal approaches. Figure 1.3 compares the three approaches. Short-time approaches focus on short segments of the signal for determining temporal properties or for estimating spectra. Examples include linear prediction analysis [18] and cepstral analysis [19], which have been used successfully for speech coding, speaker/speech recognition, etc. Narrowband analysis algorithms have higher frequency resolution than time resolution. Spectrotemporal analysis operates on larger time-frequency patches and analyzes 2-D amplitude and frequency modulations (AM and FM, respectively).

In this thesis, our objective is to develop novel methods for spectrotemporal modeling and analysis of speech signals. The motivation for this stems from recent findings in auditory neuroscience. Neurophysiological studies on animal models have shown that certain neurons in the primary auditory cortex (A1) are tuned to specific spectrotemporal patterns of sounds [20–23]. The tuning is quantified by studying the response of a cortical neuron to varied time-frequency patterns [24–27]. The resulting response is referred to as the spectrotemporal receptive field (STRF) of the neuron [28–32]. STRFs are hypothesized to detect and extract time-

frequency patterns of interest from the auditory spectrograms. Inspired by STRFs, neurophysiology based algorithms and representations have been developed, which have found applications in speech/non-speech separation [33], speech denoising [34], assessment of speech intelligibility [35], and speech recognition [36], etc. The evidence of time-frequency tuning in auditory cortical neurons forms the motivation for this thesis with the key objective of developing spectrotemporal speech processing algorithms.

1.1 Notations

Functions in 1-D and 2-D are denoted by lowercase and uppercase letters, respectively. Vector quantities are denoted by boldface letters. The 1-D Fourier transform of a signal $s(t)$ is denoted by $\hat{s}(\omega)$, $t, \omega \in \mathbb{R}$. The short-time Fourier transform is denoted by $S(t, \omega)$. The time and frequency variables are represented jointly as $\boldsymbol{\omega} = (t, \omega) \in \mathbb{R}^2$. The 2-D Fourier transform of a spectrotemporal patch is given by

$$\hat{S}(\Omega_t, \Omega_\omega) = \iint S(t, \omega)\, W(t, \omega)\, e^{-\mathrm{j}t\,\Omega_t - \mathrm{j}\omega\,\Omega_\omega}\, \mathrm{d}t\, \mathrm{d}\omega,$$

where Ω_t and Ω_w denote the Fourier variables corresponding to t and ω, respectively, and W denotes the t-f window. The 2-D Fourier variable is represented succinctly as $\boldsymbol{\Omega} = (\Omega_t, \Omega_\omega) \in \mathbb{R}^2$. The quantity will be referred to as the *grating compression transform* (GCT).

The symbols $\mathbb{R}_{\geq 0}$ denotes the set of non-negative real numbers. The symbol * denotes convolution and $\delta(\cdot)$ denotes the Dirac delta. The operation $\lfloor x \rfloor$, $x \in \mathbb{R}$ gives the greatest integer less than or equal to x. The symbol ∇ and $\|\cdot\|$ denote the gradient operator and the ℓ_2 norm, respectively.

1.2 AM-FM Demodulation

Central to AM-FM modeling is the problem of *demodulation*, which involves estimating the amplitude and frequency modulations from a given signal. The demodulation in 1-D is typically achieved using the *Hilbert transform* by constructing the analytic signal [37]. Consider the signal

$$x(t) = a(t)\cos\phi(t), \tag{1.2}$$

where $a(t)$ and $\phi(t)$ represent the AM and instantaneous phase/phase modulation (PM), respectively. Typically, the modulations are slowly varying than the signal. The demodulation problem is to determine $a(t)$ and $\phi(t)$ given $x(t)$. Unlike the case of a pure tone/sinusoid, the Hilbert transform of an AM-FM signal in general does not yield the quadrature. This is where the Bedrosian theorem [38] becomes useful. If the bandwidth of $a(t)$ is smaller than the carrier frequency, then the Bedrosian theorem guarantees that the Hilbert transform generates exact quadratures:

$$x_q(t) = \mathcal{H}\{x(t)\} = a(t)\sin\phi(t).$$

Otherwise, the Hilbert transform generates approximate quadratures. The analytic signal is constructed as follows:

$$x_a(t) = x(t) + \mathrm{j}x_q(t) = a(t)(\cos\phi(t) + \mathrm{j}\sin\phi(t)) = a(t)e^{\mathrm{j}\phi(t)}. \tag{1.3}$$

The modulus and angle of the analytic signal give the AM and PM respectively. The instantaneous frequency or FM can be obtained as the derivative of the instantaneous phase $\omega_i(t) = \dfrac{\mathrm{d}\phi(t)}{\mathrm{d}t}$.

While the Hilbert transform readily solves the 1-D demodulation problem, the 2-D counterpart is not so straightforward. We need an appropriate extension of the Hilbert transform and this is precisely where the *Complex Riesz Transform (CRT)* becomes relevant [39–41].

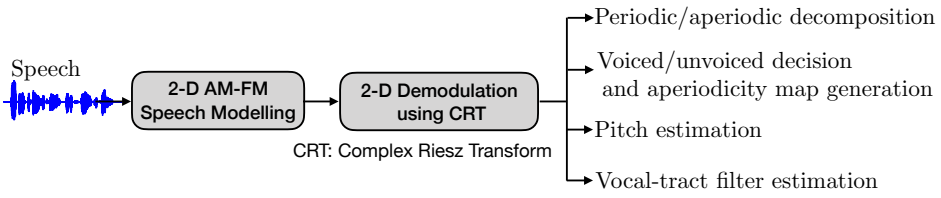


Figure 1.4: Overview of the thesis.

1.3 Overview of the Thesis

The key idea is to solve the problem of demodulation in 2-D and estimate the 2-D AM and FM components from a spectrogram patch. Most speech processing approaches have remained largely 1-D or at best "stacked 1-D" giving the impression of a 2-D approach. On the other hand, we develop a bona fide 2-D approach based on a sound mathematical foundation. The 2-D modeling takes into account the mechanism of speech production and the t-f structure of various speech sounds. The demodulation is based on the complex Riesz transform, which has recently proven to be successful in various imaging applications. We shall show that estimation of speech parameters such as fundamental frequency of the speaker, voiced/unvoiced demarcation, quantification of aperiodicity, extraction of source and vocal tract filter parameters can all be addressed by working in the spectrotemporal domain. These parameters are useful for several tasks such as speech synthesis, voice conversion, prosodic modifications, etc. The accuracy of the estimated parameters is assessed by carrying out signal reconstruction. Figure 1.4 gives an overview of the thesis.

1.4 Source-Filter Model for Speech Production

Figure 1.5 illustrates the speech production system and Figure 1.6 depicts the source-filter model of speech. The term *source* refers to the phonation that occurs at the vocal folds (or glottis) during speech production and *filter* refers to the region from the vocal folds to the lips, which defines the vocal tract (see Figure 1.5). Various

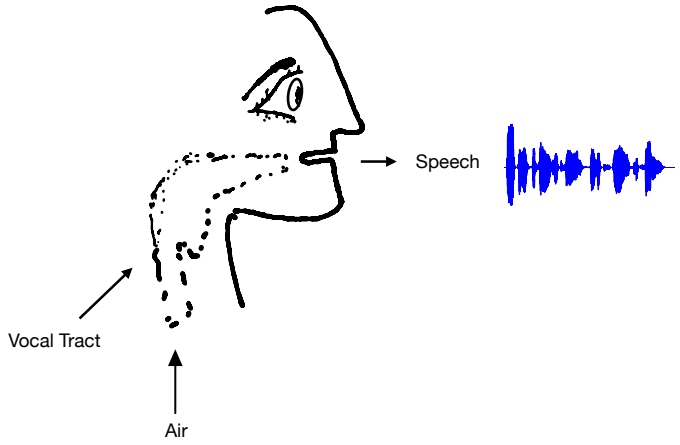


Figure 1.5: Diagram for the speech production system.

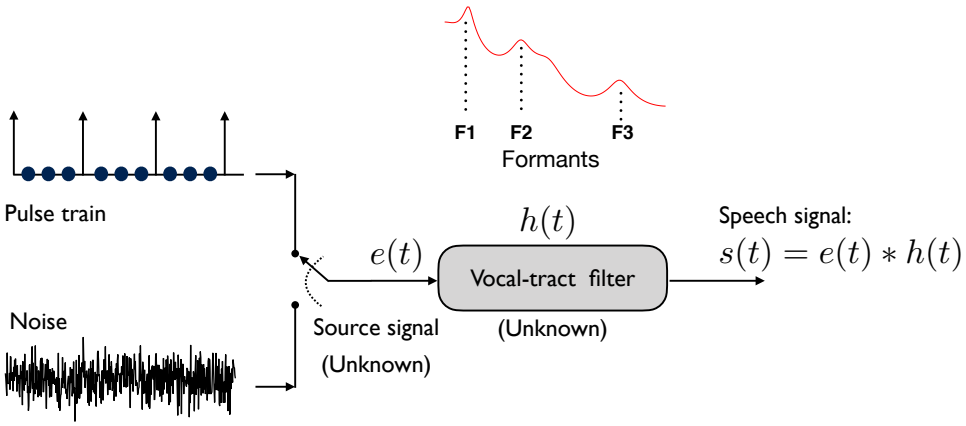


Figure 1.6: The source-filter model for speech production system.

speech sounds are produced by controlling the type of source and the shape of the vocal tract in a dynamic fashion. The glottal excitation may be quasi-periodic, aperiodic or a mixed one depending on the type of sound. The excitation is quasi-periodic for voiced sounds such as /a/, /e/,/i/,/o/,/u/,' which is caused by nearly periodic opening and closing of vocal folds. It is aperiodic/noise-like for unvoiced sounds (when the vibrations at the glottis are absent) such as the /f/, and /s/ in the word 'fuss'. The vocal-tract filter acts like a tube with varying cross-sectional area and length. The resonances of the vocal tract are called *formants*, which are

characteristic to the speaker. The glottis and vocal tract are mutually interactive and nonlinearly coupled to each other. However, the source-filter model [42] makes a simplifying linear system assumption with independent source/excitation and filter.

The speech signal is given by a convolution of the excitation $e(t)$ and the impulse response $h(t)$ of the vocal-tract filter:

$$s(t) = h(t) * e(t). \tag{1.4}$$

In Fourier-domain, the convolution manifests as

$$\hat{s}(\omega) = \hat{h}(\omega)\hat{e}(\omega), \tag{1.5}$$

where $\hat{h}$ is slowly varying and $\hat{e}$ is fast-varying. Therefore, the vocal-tract "modulates" the excitation. Given a speech signal, the problem of estimating the source and filter components is ill-posed. Only by taking into account the speech production mechanism can suitable priors be constructed that allow for "deconvolution" or "demodulation" of the excitation and filter components.

1.5 Speech Reconstruction

While speech analysis aims at estimating the source and filter parameters, the task of speech reconstruction involves combining the estimated parameters to synthesize an intelligible and natural-sounding speech without relying on the STFT phase. A system that performs the analysis with the objective of reconstruction is known as the *vocoder* [43]. A vocoder is an integral part of the pipeline in text-to-speech technology. A well known vocoder is STRAIGHT [44], which stands for *Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum*. Its successor is the WORLD vocoder [45]. Both these vocoders follow the *analysis-by-synthesis* approach. In the analysis stage, the source and filter parameters are estimated which include the fundamental frequency of the speaker, aperiodicity parameters, voiced/unvoiced decisions, and the frequency response of the vocal-tract filter. The

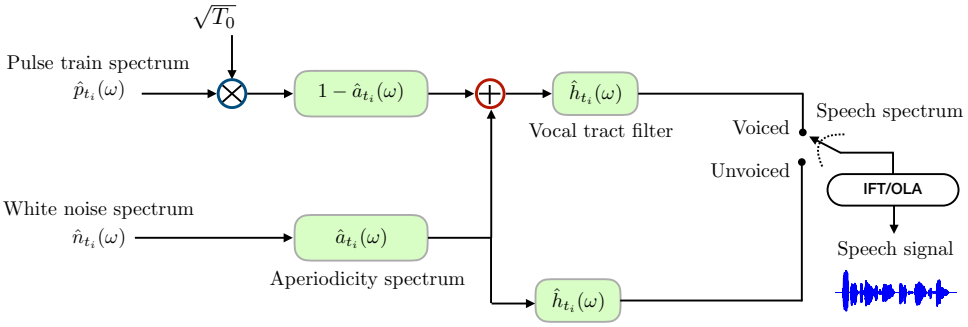


Figure 1.7: The block diagram of spectral synthesis model used for speech reconstruction. The phase spectrum for vocal tract and aperiodicity spectrum is modeled using the minimum-phase approximation.

synthesis stage reconstructs the speech waveform by using the parameters obtained in the analysis stage. A drawback of STRAIGHT and WORLD is that they require a lot of manual tuning to obtain high quality synthesis, which has been done over several years. The STRAIGHT and WORLD vocoder versions have evolved over the years and achieved a high degree of perfection. Our objective is to develop if a totally new approach relying on spectrotemporal modulations that would be competitive with these vocoders.

1.5.1 *Spectral Synthesis Model*

This is a frequency-domain approach where voiced and unvoiced speech segments are reconstructed using the V/UV decisions. Figure 1.7 shows the block diagram of spectral synthesis model. The spectrum of a voiced speech frame is modeled as follows:

$$\hat{s}_{v,t_i}(\omega) = \hat{h}_{t_i}(\omega)\hat{e}_{t_i}(\omega), \tag{1.6}$$

where $\hat{e}_{t_i}(\omega)$ denotes the excitation spectrum comprising two parts:

$$\hat{e}_{t_i}(\omega) = \underbrace{\sqrt{T_0}(1-\hat{a}_{t_i}(\omega))\hat{p}_{t_i}(\omega)}_{\text{periodic part}} + \underbrace{\hat{a}_{t_i}(\omega)\hat{n}_{t_i}(\omega)}_{\text{aperiodic part}}, \tag{1.7}$$

Table 1.1: Recently developed vocoders based on deep learning.

Neural vocoder type	Vocoder	Remark
Autoregressive	WaveNet [46]	Fully autoregressive, uses dilated convolutions to model the long-term dependencies
	WaveRNN [47]	Uses a stack of Recurrent neural networks
	FFTNet [48]	Uses a simplified architecture based on FFT butterfly
	LPCNet [49]	A variant of WaveRNN that combines linear prediction with RNN
Non-autoregressive	WaveGlow [50]	Flow-based generative model
	Parallel WaveGAN [51]	Uses GAN architecture
	Neural source-filter [52]	Inspired from conventional source-filter model

where in turn $\hat{p}_{t_i}(\omega)$ is the spectrum of an impulse at synthesis time-instant t_i, $\hat{a}_{t_i}(\omega)$ is the filter response derived from the aperiodicity parameters, T_0 denotes the pitch period, and $\hat{n}_{t_i}(\omega)$ is complex white Gaussian noise with zero mean and unit variance. The first term models the periodic part while the second one accounts for aperiodicity in the excitation. The frequency response of the vocal-tract filter in Equation (1.6) is given by

$$\hat{h}_{t_i}(\omega) = \hat{v}_{t_i}(\omega)e^{\mathrm{j}\theta_{\min,t_i}(\omega)},$$

where $\theta_{\min,t_i}(\omega)$ denotes the phase spectrum derived from the magnitude spectrum $\hat{v}_{t_i}(\omega)$ using the minimum-phase approximation (Appendix A). Similarly, the aperiodicity spectrum $\hat{a}(t_i,\omega)$ is also derived using the minimum-phase approximation. At every synthesis time instant, the spectrum $\hat{s}_{v,t_i}(\omega)$ is subjected to inverse Fourier transform (IFT) and the resulting speech segments are overlap-added (OLA) pitch-synchronously.

The spectrum of unvoiced segments is modeled as follows:

$$\hat{s}_{uv,t_i}(\omega) = \hat{h}_{t_i}(\omega)\hat{a}_{t_i}(\omega)\hat{n}_{t_i}(\omega). \tag{1.8}$$

The unvoiced segments are synthesized by inverting the spectrum $\hat{s}_{uv,t_i}(\omega)$ followed by OLA.

1.5.2 *Deep Learning Models*

In the past five years, there has been a surge of deep learning based vocoders, which have surpassed traditional vocoders such as STRAIGHT and WORLD in terms of performance. The deep learning vocoders use generative modeling techniques for waveform generation. They are trained to learn the underlying probability distribution of the data conditioned on the acoustic features. The acoustic features could correspond to source, filter or a combination of both. The prominent deep learning based vocoder is *WaveNet* [2, 46], which uses an autoregressive model for waveform generation. A drawback of WaveNet is that the sample-by-sample generation mechanism makes it considerably slow. In order to speed up waveform generation, non-autoregressive vocoders have also been proposed — knowledge distillation networks [53], flow-based generation [54], and generative adversarial networks (GAN) [55]. Table 1.1 lists recently developed deep learning vocoders.

1.6 Organization of the Thesis

The chapter-wise organization of the thesis is given below.

Chapter 2: AM-FM Modeling of the Speech Spectrogram and Demodulation in 2-D

We provide an overview of 2-D cosine signals, their AM-FM counterparts and discuss the salient properties of their Fourier transforms. Using the source-filter theory of speech production, we show that a voiced spectrogram-patch can be modeled by using weighted sum of 2-D AM-FM cosines, referred to as the multicomponent AM-FM model. One of the key findings is that the optimum model order is proportional to the variations in the F0 of the speaker. In the context of the demodulation problem, we introduce the complex Riesz transform. The estimated AM and FM are used for estimating weights of the 2-D cosine carriers in the multicomponent model. Objective evaluation on a speech database confirms that the multicomponent model

has superior model accuracy than its moncomponent counterpart.

Chapter 3: Periodic and Aperiodic Decomposition of Speech Signals

Here, we focus on the 2-D FM (the *carrier spectrogram*), which carries rich information of the glottal excitation. Visual inspection of the carrier spectrogram shows that it exhibits two distinct spectrotemporal signatures – one corresponding to periodic sounds and the other to aperiodic sounds. We also derive t-f maps that capture the local *coherence* and orientation. The effectiveness of these maps is demonstrated for solving the problem of periodic/aperiodic decomposition of speech, which is formulated as a binary classification problem. Since there are no ground truth labels available, the problem is solved in an unsupervised manner.

Chapter 4: Voiced/Unvoiced Segmentation and Quantification of Speech Aperiodicity

The problem of voiced/unvoiced segmentation is viewed as a binary class classification problem, for which we derive novel features from the coherencegram. Experimental evaluation shows that the new features are robust to variabilities in the waveform. We then propose a numerical measure for the estimation of aperiodicity of FM sinusoids. The idea is extended for estimating the aperiodicity content of voiced speech frames. The carrier spectrogram, once again, turns out to be a useful representation for the estimation of speech aperiodicity. We derive band-wise aperiodicity parameters suitable for modeling the stochastic component of speech signals in traditional vocoders. The effectiveness of V/UV segmentation and band-wise aperiodicity parameters is shown by incorporating them in the WORLD vocoder.

Chapter 5: Pitch Estimation From the Carrier Spectrogram

Estimation of the pitch of the speaker is an important problem in speech analysis/synthesis, for which we employ the carrier spectrogram. Objective evaluation with state-of-the-art pitch estimation algorithms on two speech databases shows the superiority of the proposed methods.

Chapter 6: Vocal-tract Filter Estimation and Speech Reconstruction

In this chapter, we address the problem of estimating the vocal-tract filter, formants and bandwidths, based on the 2-D AM. The formant bandwidths play a crucial role for high-quality speech synthesis. A novel formant bandwidth correction method is proposed to mitigate estimation errors. Effectiveness of the proposed source and filter parameters is shown by incorporating them in the spectral synthesis model and in the WaveNet vocoder.

1.7 Databases Used for Evaluation

The following databases have been used for evaluation.

CMU-ARCTIC [56]: There are approximately 1200 utterances per speaker (sampling rate of 32 kHz) with parallel Electroglottogram (EGG) recordings. The database was designed specifically for speech synthesis research at Carnegie Mellon University (CMU). It is available for free download at: `http://festvox.org/cmu_arctic/`.

CSTR-FDA: This database includes 50 speech utterances (20 kHz, 16bit) each from one and one female speaker with parallel EGG recordings. This database was mainly developing by the Center for Speech Technology Research (CSTR), University of Edinburgh for F0 Determination Algorithm Evaluation (FDA). It is available for commercial use at: `https://www.cstr.ed.ac.uk/research/projects/fda/`.

Starkey [57]: The speech material is recorded from 16 American speakers out of which 8 are male and 8 are female. Each speaker reads the standard rainbow passage [58]. The recordings are at sampling rate of 44.1 kHz with 16 bit quantization. The database is freely available at: `https://starkeypro.com/research/research-resources/open-access-stimuli.html`.

TIMIT [59]: This database contains a total of 6300 utterances spoken by 630 speakers, with each speaker contributing 10 utterances. The prompts for the 6300 utterances consist of 2 dialect "shibboleth" sentences (SA), 450 phonetically-compact sentences (SX), and 1890 phonetically-diverse sentences (SI). TIMIT is divided into

Table 1.2: Absolute Category Rating Scale for MOS test.

Score	Description
1	**Bad:** Very annoying artifacts and/or very bad resynthesis quality
2	**Poor:** Annoying artifacts and/or bad resynthesis quality
3	**Fair:** Some artifacts and/or good synthesis quality but not identical
4	**Good:** Very few artifacts and/or very good resynthesis quality but not identical
5	**Excellent:** No artifacts and/or identical resynthesis quality

a training set and a test set. The training set contains 4620 utterances, and the test set contains 1344 utterances of which 192 form a core test set.

VTR [60]: The Vocal Tract Resonance (VTR) database developed by Microsoft is composed of 538 utterances (only SX and SI sentences) selected as a subset of TIMIT corpus. The database provides trajectories of the first three formants corrected through extensive manual labeling. The selected subset of utterances in VTR contains 192 and 346 utterances from the TIMIT test set and training set, respectively. The 192-utterance test subset contains a total of 24 speakers with 5 SX and 3 SI sentences for each speaker, and 173 speakers in the 346-utterance training with 1 SX and 1 SI sentences for each speaker. Thus, the selected 538 utterances represent a balanced selection of dialect, speaker and gender while consisting of rich phonetic contexts. The formant trajectories are first estimated by employing the VTR tracking algorithm described in [61], followed by extensive manual correction. The database is available for download at: `http://www.seas.ucla.edu/spapl/VTRFormants.html`.

1.8 Performance Measures

The performance measures used are Mean Opinion Score (MOS) and Perceptual Evaluation of Speech Quality (PESQ). MOS is used for subjective evaluation of speech quality as recommended by International Telecommunication Union-Telecommunication Sector (ITU-T) [62,63]. It is based on the opinion of a number of listeners about the speech quality and the score assignment is as per the descrip-

tion given in Table 1.2. Although a bit time-consuming, MOS accurately reflects the perceived quality of speech. The PESQ metric recommended by the ITU-T P.862 [64] standard is used for the objective evaluation of speech quality. PESQ is computed between a reference (typically the clean speech) and a processed signal (synthesized signal in our case). It lies in the range from -0.5 to 4.5 with a higher value indicating a quality closer to the reference. PESQ was developed primarily for automatic assessment of end-to-end speech quality in telecommunications. Metrics such as signal-to-noise ratio do not accurately quantify the user experience of the speech quality. PESQ solves this problem by incorporating a perceptual model that can distinguish between audible and inaudible artifacts.

Chapter 2

AM-FM Modeling of the Speech Spectrogram and Demodulation in 2-D

In this chapter, we describe two-dimensional (2-D) AM-FM modeling of a narrowband speech spectrogram and address the problem of separating AM and FM components from the spectrogram. In Section 2.1, we consider the 2-D AM-FM cosine signal model and its Fourier transform. We also discuss the taxonomy of the Fourier transform of a voiced spectrotemporal patch taken from a spectrogram and show that 2-D AM-FM cosine signals can be employed to model a voiced patch. In view of this, a summary of state-of-the-art 2-D AM-FM models for a speech spectrgram is given in Section 2.2. In Section 2.3, we derive the 2-D AM-FM model of a speech signal by using the basic principles of human speech production system. In particular, we show that a voiced spectrotemporal patch of a narrowband speech spectrogram can be modeled using a weighted sum of 2-D AM-FM cosine signals, which gives rise to a multicomponent AM-FM model. Next, we solve the problem of demodulation in 2-D, which gives access to the amplitude and frequency modulations. The problem is typically accomplished with the help of the quadrature. While the quadrature component of a sinusoid in 1-D is obtained by means of the *Hilbert transform*, the 2-D counterpart requires a consistent generalization of the Hilbert transform. The complex Riesz transform (CRT) meets this specification [65, 66]. Hence, we discuss the complex Riesz transform and its properties in Section 2.4. The demodulation of a 2-D AM-FM cosine signal using CRT is described in Section 2.4. In Section 2.5, we employ CRT-based demodulation for the estimation of AM and FM components

from speech spectrogram. In Section 2.6, the performance of the multicomponent AM-FM model is evaluated with respect to its monocomponent counterpart. We conclude in Section 2.7 with the chapter summary.

2.1 Review of Two-dimensional Cosines, Fourier Transform, and AM-FM Cosines

The 2-D AM-FM modeling of a speech spectrogram and the demodulation algorithm require a clear understanding of a 2-D bandpass signal and its AM-FM representation. Hence, we begin by providing several examples of 2-D cosines, Fourier transform and 2-D AM-FM cosine signals.

2.1.1 *Two-dimensional Cosines*

A 2-D cosine with constant amplitude and frequency is expressed in cartesian coordinates as $\cos(t\Omega_t + \omega\Omega_\omega)$, and in polar coordinates as

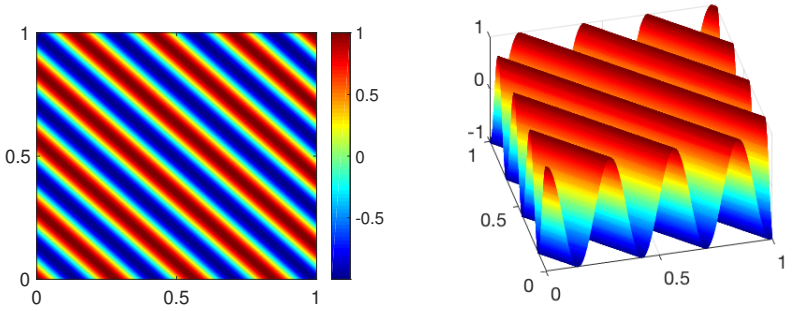
$$F(t,\omega) = \cos\left(\Omega_0(t\cos\beta_0 + \omega\sin\beta_0)\right), \tag{2.1}$$

where Ω_0 denotes the spatial frequency of the sinusoid, β_0 denotes the orientation of the sinusoid, measured with respect to the t-axis. The phase of the cosine is a linear function of continuous variables t and ω. The discretized version of a 2-D cosine in Equation (2.1) is expressed as

$$F[k_1,k_2] = \cos\left(\Omega_0(k_1T_t\cos\beta_0 + k_2T_\omega\sin\beta_0)\right), \tag{2.2}$$

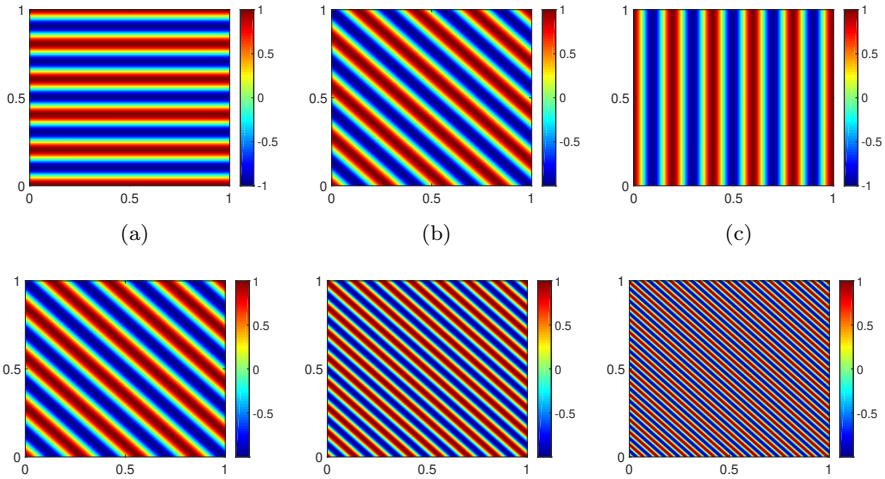
where T_t and T_ω denote the sampling steps along t-axis and ω-axis, respectively.

Figure 2.1 displays a 2-D cosine and its 3-D view. Figure 2.2 displays 2-D cosines obtained by changing either frequency or orientation in Equation (2.1).



(a) (b)

Figure 2.1: (a) A 2-D cosine with $\beta_0 = \pi/4$, $\Omega_0 = 10\pi$, $T_t = T_\omega = 0.002$, and (b) its 3-D view.



(a) (b) (c)

(d) (e) (f)

Figure 2.2: Illustration of 2-D cosines. In the first row, the frequency is constant and the orientation is varied. In the second row, the spatial frequency is varied and the orientation is kept constant.

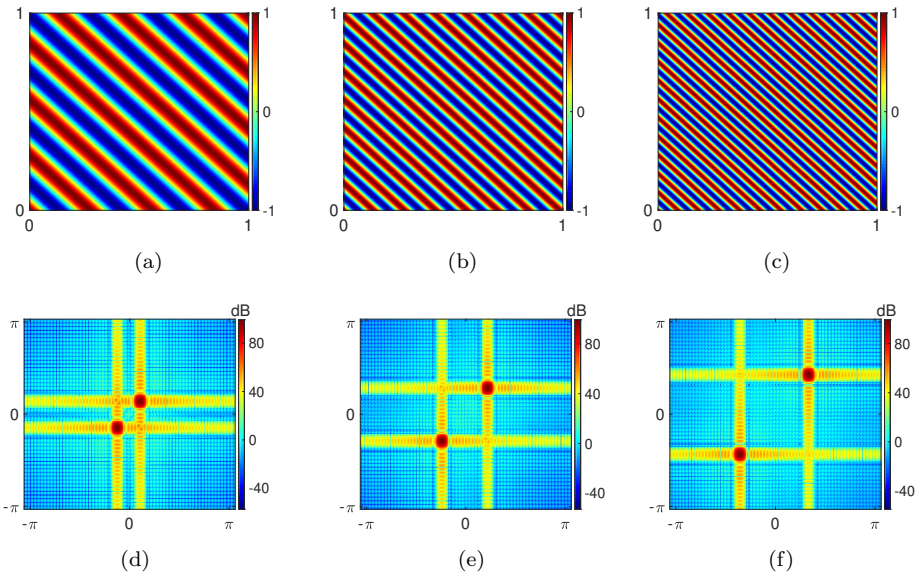


Figure 2.3: Illustration of 2-D cosines and corresponding Fourier spectra. The first row displays 2-D cosines with fixed orientation but varying spatial frequency. The higher the spatial frequency of the sinusoid, the greater the distance of the impulses from the origin in the Fourier domain. The Fourier transform of sinusoidal patterns (which resemble gratings) is also referred to as the *grating compression transform* (GCT) [1]

.

2.1.2 *Two-dimensional Fourier Transform*

The Fourier transform $\hat{F}(\Omega_t, \Omega_\omega) : \mathbb{R}^2 \to \mathbb{C}$ of a real-valued function $F(t, \omega) : \mathbb{R}^2 \to \mathbb{R}$ is defined as follows:

$$\hat{F}(\Omega_t, \Omega_\omega) \triangleq \int_\omega \int_t F(t, \omega) e^{-j(t\Omega_t + \omega\Omega_\omega)} \, \mathrm{d}t\mathrm{d}\omega, \tag{2.3}$$

where Ω_t and Ω_ω denote the frequency variables in the Fourier domain corresponding to the variables t and ω, respectively.

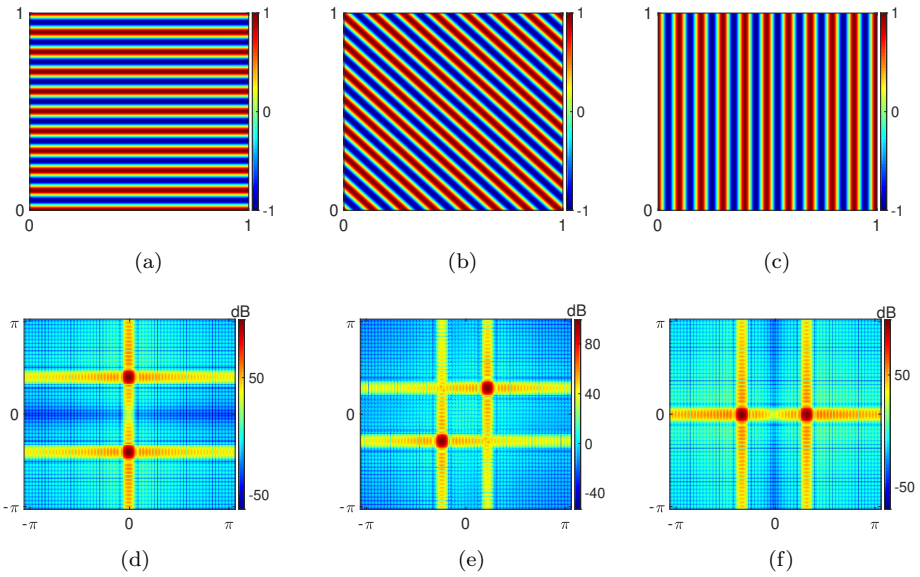


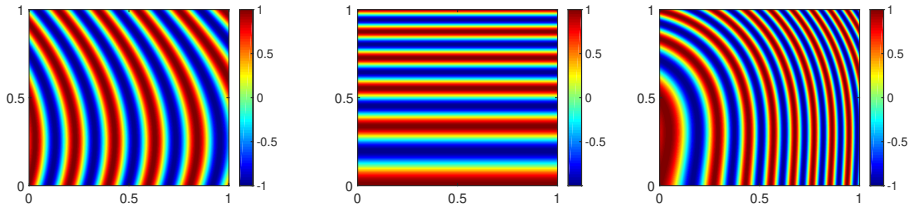
Figure 2.4: Illustration of 2-D cosines and the corresponding Fourier spectra. The first row displays 2-D cosines with fixed frequency but varying orientation. The line joining the impulse pair is orthogonal to the orientation of the sinusoid.

2.1.3 *Fourier Transform of a 2-D Cosine*

The 2-D Fourier transform of the 2-D cosine $F(t,\omega) = \cos\Omega_0(t\cos\beta_0 + \omega\sin\beta_0)$ is given by

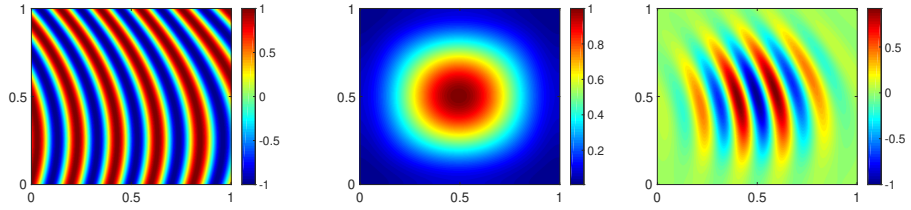
$$\hat{F}(\Omega_t,\Omega_\omega) = \pi\delta(\Omega_t - \Omega_0\cos\beta_0, \Omega_\omega - \Omega_0\sin\beta_0) + \pi\delta(\Omega_t + \Omega_0\cos\beta_0, \Omega_\omega + \Omega_0\sin\beta_0), \tag{2.4}$$

which is a pair of 2-D Dirac deltas. Figure 2.3 and Figure 2.4 display examples of 2-D cosines and their Fourier magnitude spectra. The locations of the impulses are governed by the spatial frequency and orientation of the sinusoid.



(a) (b) (c)

Figure 2.5: Illustration of how frequency modulations are introduced in a 2-D cosine by changing (a) only the orientation, (b) only the frequency, and (c) both orientation and frequency.



(a) (b) (c)

Figure 2.6: Illustration of (a) a frequency-modulated 2-D cosine, (b) an amplitude modulating function, and (c) a 2-D AM-FM signal.

2.1.4 2-*D AM-FM Cosines*

An amplitude and frequency modulated 2-D cosine $F(\boldsymbol{\omega}) : \mathbb{R}^2 \rightarrow \mathbb{R}$ is written as follows:

$$F(\boldsymbol{\omega}) = A(\boldsymbol{\omega}) \cos\Big(\Omega_0(\boldsymbol{\omega})\Phi(\boldsymbol{\omega})\Big), \tag{2.5}$$

where $\Omega_0(\boldsymbol{\omega}) = \Omega_0 + \Delta\Omega(\boldsymbol{\omega})$ denotes the spatial frequency, $\Delta\Omega(\boldsymbol{\omega})$ represents the frequency modulation around the center frequency Ω_0, $A(\boldsymbol{\omega})$ denotes the amplitude modulation, and $\Phi(\boldsymbol{\omega})$ denotes the phase expressed as:

$$\Phi(\boldsymbol{\omega}) = t \cos\beta_0(\boldsymbol{\omega}) + \omega \sin\beta_0(\boldsymbol{\omega}), \tag{2.6}$$

where $\beta_0(\boldsymbol{\omega})$ is the local orientation of the 2-D cosine. In contrast to a 1-D cosine, a 2-D cosine has the orientation $\beta_0(\boldsymbol{\omega})$ as an additional degree of freedom. Consequently,

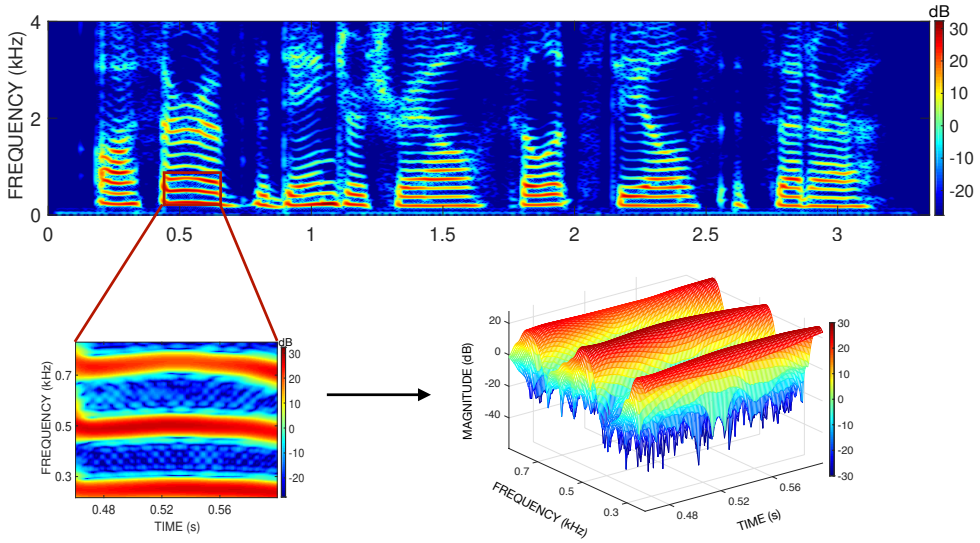
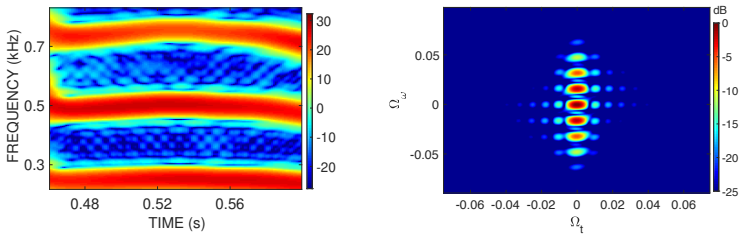


Figure 2.7: A voiced spectrogram patch.



(a) (b)

Figure 2.8: (a) A voiced spectrogram patch, and (b) its Fourier transform magnitude.

frequency modulations in a 2-D cosine can be introduced either by changing the orientation and frequency independently or jointly. Figure 2.5 illustrates this effect. Figure 2.6 shows a 2-D AM-FM signal obtained by using Equation (2.5).

2.1.5 *A Voiced Speech Patch and its Fourier Transform*

Thus far, we have seen examples of stylized 2-D AM-FM signals. Next, we consider a real voiced signal t-f patch taken from a speech spectrogram and its Fourier transform. Figure 2.7 shows a narrowband speech spectrogram and a patch taken

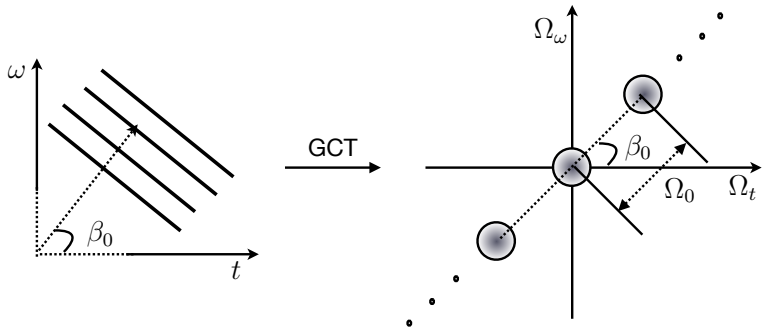


Figure 2.9: Schematic of Grating Compression Transform (GCT); Ω_0 and β_0 denote the 2-D frequency and orientation of the 2-D sinusoid, respectively.

from a voiced t-f region. Its Fourier transform is displayed in Figure 2.8(b). The Fourier transform of a real voiced patch exhibits peaks at the DC, the dominant frequency, and its harmonics. The key observation is that each pair of peaks can be modeled by using a single 2-D AM-FM cosine.

The 2-D Fourier transform of sinusoidal patterns (which resemble gratings) is also known as *grating compression transform* (GCT) [1]. The GCT of a 2-D AM-FM cosine is a useful tool that shows its bandpass nature in the Fourier domain. Some of the salient properties of GCT of a 2-D cosine can be described with the help of the schematic in Figure 2.9. The parallel lines in the figure illustrates a 2-D sinusoid and the circular lobes in the GCT domain represent the locations where most of the energy is concentrated. The energy at $(\Omega_t, \Omega_\omega) = (0, 0)$ in the GCT domain is due to the DC component. The radius of a circle reflects the spread of energy around the peak in the GCT domain.

2.2 State-of-the-art Spectrogram Patch Models

A seminal contribution for 2-D speech modeling was made by Wang and Quatieri [1], who extended the idea of sinusoidal modeling of speech [42] and proposed a 2-D sinusoidal series-based modulation model for small regions of narrowband speech spectrograms. They also extended the model to wideband spectrograms [67]. A

windowed patch is represented as follows:

$$S_W(\boldsymbol{\omega}) \approx V(\boldsymbol{\omega})\Bigg(\alpha_0 + \sum_{k=1}^{\infty} \alpha_k \cos\left(k\Omega_0(t\cos\beta + \omega\sin\beta)\right)\Bigg), \tag{2.7}$$

where $V(\boldsymbol{\omega})$ is the amplitude modulation that represents time-varying vocal tract envelope and $\alpha_0 \in \mathbb{R}_{\geq 0}$ is the DC value of the patch, which makes it a non-negative quantity. The pitch harmonics/carrier are modeled as stationary cosines with constant spatial frequency Ω_0 and orientation β. Ezzat *et al.* [68] proposed production-based speech spectrogram patch models and investigated how different acoustic events (e.g. voicing/unvoicing, plosives, onset/offset, etc.) manifest in the 2-D Fourier domain. They expressed the localized AM components using Gabor atoms [69]. They showed that 2-D AM and FM encode the phonetic and speaker's attributes, respectively.

Aragonda and Seelamantula [41] proposed a 2-D AM-FM model where the stationary FM assumption proposed in [1] was generalized to spatially varying FM. They proposed an accurate demodulation strategy using the complex Riesz transform (CRT). They showed that the generalized model and CRT-based demodulation algorithm resulted in a superior performance over the sinusoidal-series based model proposed in [1].

Motivated from the success of the 2-D AM-FM model proposed in [41] and to keep the exposition self-contained, we describe this model and its multicomponent counterpart.

2.3 Multicomponent 2-D AM-FM Signal Model

To begin with, we describe modeling of the speech spectrum in 1-D that relies on the source-filter theory of speech production [42]. The analysis carried out in 1-D for modeling the magnitude spectrum of a voiced speech provides a direct link to 2-D modeling of t-f localized voiced spectrotemporal patches.

2.3.1 *Modeling the 1-D Magnitude Spectrum*

The source-filter model considers voice production as involving two almost separate processes: excitation generation and vocal tract filtering. The excitation is typically modeled by using an impulse train for voiced sounds while the unvoiced sounds are considered to be noise-like. A windowed voiced speech $s_w(t) : \mathbb{R} \to \mathbb{R}$ is modeled as follows:

$$s_w(t) = w(t)\bigg(v(t) * \sum_{k=-\infty}^{\infty} \delta(t - kT_0)\bigg), \tag{2.8}$$

where $w(t)$, $v(t)$, and T_0 represent the 1-D analysis window, the vocal tract impulse response, and the pitch period, respectively. The equivalent representation in Fourier domain is given by

$$\hat{s}_w(\omega) = \hat{w}(\omega) * \bigg(\hat{v}(\omega) \sum_{k=-\infty}^{\infty} \delta\bigg(\omega - \frac{2\pi k}{T_0}\bigg)\bigg), \tag{2.9}$$

where $\hat{s}_w(\omega) : \mathbb{R} \to \mathbb{C}$. The magnitude of the function $\hat{s}_w(\omega)$ can be approximated as follows [42]:

$$|\hat{s}_w(\omega)| \approx |\hat{v}(\omega)| \underbrace{\bigg| \sum_{k=-\infty}^{\infty} \hat{w}\bigg(\omega - \frac{2\pi k}{T_0}\bigg)\bigg|}_{\hat{p}(\omega)}, \tag{2.10}$$

where $\hat{p}(\omega)$ represents the magnitude spectrum of the sound source signal and it is periodic in ω with period $\frac{2\pi}{T_0}$. Hence, a Fourier-series expansion of $\hat{p}(\omega)$ is given by

$$\hat{p}(\omega) = \alpha_0 + \sum_{k=1}^{\infty} \alpha_k \cos(kT_0\omega + \psi_k). \tag{2.11}$$

Substituting Equation (2.11) in Equation (2.10), we get

$$\begin{aligned} |\hat{s}_w(\omega)| &\approx |\hat{v}(\omega)|\bigg(\alpha_0 + \sum_{k=1}^{\infty} \alpha_k \cos(kT_0\omega + \psi_k)\bigg) \\ &= \alpha_0|\hat{v}(\omega)| + \alpha_1|\hat{v}(\omega)|\cos(T_0\omega + \psi_1) \\ &\quad + \alpha_2|\hat{v}(\omega)|\cos(2T_0\omega + \psi_2) + \cdots, \end{aligned} \tag{2.12}$$

which represents a decomposition of the power spectrum of a speech signal in terms of the amplitude modulations given by $|\hat{v}(\omega)|$ and harmonically related 1-D cosine carriers, each term in the summation is referred to as a *component*. The coefficient α_k determines the strength of the amplitude modulation for the k^{th} component. The 1-D AM-FM model given in Equation (2.12) can be directly extended in 2-D to model a spectrotemporal patch.

2.3.2 *Multicomponent AM-FM Model for a Spectrogram Patch*

The 2-D counterpart of the 1-D magnitude spectrum given in Equation (2.12) can be expressed as follows:

$$\begin{aligned} S_W(\boldsymbol{\omega}) &\approx V(\boldsymbol{\omega})\Big(\alpha_0 + \sum_{k=1}^{K} \alpha_k \cos k\Phi(\boldsymbol{\omega})\Big) \\ &= \underbrace{\alpha_0 V(\boldsymbol{\omega})}_{\text{low-pass component}} + \underbrace{\alpha_1 V(\boldsymbol{\omega}) \cos \Phi(\boldsymbol{\omega})}_{\text{fundamental band-pass component}} \\ &+ \underbrace{\alpha_2 V(\boldsymbol{\omega}) \cos 2\Phi(\boldsymbol{\omega}) + \ldots}_{\text{higher-order band-pass components}}, \end{aligned}$$

where $S_W(\boldsymbol{\omega}) : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ denotes a windowed voiced patch. The model order is denoted by K, and $\{\alpha_k\}_{k=1}^{K} \in \mathbb{R}$ act as weights on the carrier and its harmonics. The amplitude modulation and the phase component are represented by $V(\boldsymbol{\omega}) : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ and $\Phi(\boldsymbol{\omega})$, respectively. The phase component of a planar 2-D cosine with spatial frequency $\Omega_0(\boldsymbol{\omega}) : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ and local orientation $\beta(\boldsymbol{\omega})$ is written as follows

$$\Phi(\boldsymbol{\omega}) = \Omega_0(\boldsymbol{\omega})(t \cos \beta(\boldsymbol{\omega}) + \omega \sin \beta(\boldsymbol{\omega})), \tag{2.13}$$

where $\Omega_0(\boldsymbol{\omega}) = \Omega_0 + \Delta\Omega_0(\boldsymbol{\omega})$ with $\Delta\Omega_0(\boldsymbol{\omega})$ representing frequency modulation and the local orientation $\beta_0(\boldsymbol{\omega}) = \beta_0 + \Delta\beta_0(\boldsymbol{\omega}) \in (-\pi, \pi)$ with $\Delta\beta_0(\boldsymbol{\omega})$ representing the small variations around β_0. In this model, a voiced source signal is represented as a sum of harmonically related and weighted 2-D sinusoidal carriers where each carrier is modulated by a slowly varying 2-D envelope $V(\boldsymbol{\omega})$ that represents the local

spectrotemporal shaping, e.g., due to dynamic formant/glottal flow structure. The 2-D AM-FM spectrogram-patch model given by Equation (2.13) is referred to as the *multicomponent 2-D AM-FM model*. It is a *monocomponent model* for $K = 1$. Unlike a sinusoidal-series based model given in Equation (2.7), which assumes a stationary carrier, the multicomponent 2-D AM-FM model given in Equation (2.13) makes no such assumption on the 2-D carrier. This argument is justified by the fact that a real voiced speech patch exhibits frequency modulations due to time-varying fundamental frequency in natural speech. We have seen in Section 2.1.4 that such frequency modulations are coupled to the frequency and the orientation of the 2-D sinusoid.

The unknown parameters of the model in Equation (2.13) are the AM $V(\boldsymbol{\omega})$, phase $\Phi(\boldsymbol{\omega})$, model coefficients $\alpha_0, \alpha_1, \cdots, \alpha_K$, and the model order K. In Section 2.5, we describe the details of estimation of the unknown parameters of the model. The model parameters are estimated in two steps: (1) estimate AM and FM components, and (2) estimate model coefficients using the estimated AM and FM in Step (1). The AM and FM components are obtained by 2-D demodulation for which we use the complex Riesz transform. Before proceeding further, we explain the Riesz transform, its key properties, and its action on a 2-D AM-FM cosine signal.

2.4 The Complex Riesz Transform (CRT)

The analytic representation of 1-D signals using the Hilbert transform is central to many applications in signal processing such as AM-FM demodulation, spectral analysis, interferometry [70–73], and single sideband modulation [74]. Motivated from these successes, there were many attempts made to extend the Hilbert transform to two dimensions, constructed using the product of 1-D functions, which resulted in the half-plane and quadrature-plane Hilbert transforms [75–77]. A drawback of these extensions is that they are not isotropic.

The complex Riesz transform provides an elegant isotropic extension of the

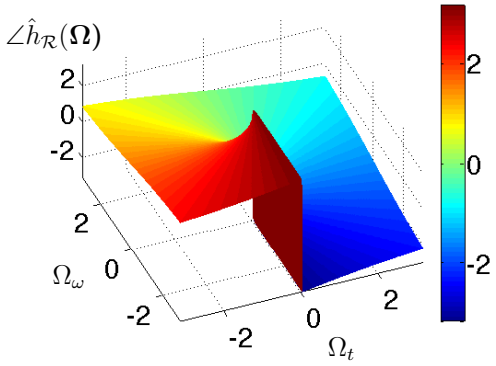


Figure 2.10: Phase response of the complex Riesz transform over the domain $[-\pi, \pi] \times [-\pi, \pi]$. The units of all axes are radians.

Hilbert transform. The Riesz transform has found applicability in fringe pattern analysis [40,66,78] and amplitude and phase decomposition [79].

The complex Riesz transform $f_{\mathcal{R}}(\boldsymbol{\omega}) : \mathbb{R}^2 \to \mathbb{C}$ of a scalar function $f(\boldsymbol{\omega}) : \mathbb{R}^2 \to \mathbb{R}$ is defined as follows:

$$f_{\mathcal{R}}(\boldsymbol{\omega}) \triangleq (f_t + \mathrm{j} f_\omega)(\boldsymbol{\omega}) = (h_t * f + \mathrm{j} h_\omega * f)(\boldsymbol{\omega}), \tag{2.14}$$

where $h_t(\boldsymbol{\omega})$ and $h_\omega(\boldsymbol{\omega})$ denote the Riesz kernels along time and frequency axes, respectively. Analogous to the 1-D Hilbert transform, the frequency responses of the Riesz kernels along Ω_t-axis and Ω_ω-axis are expressed as

$$\hat{h}_t(\boldsymbol{\Omega}) \triangleq -\mathrm{j} \frac{\Omega_t}{\|\boldsymbol{\Omega}\|}, \tag{2.15}$$

and

$$\hat{h}_\omega(\boldsymbol{\Omega}) \triangleq -\mathrm{j} \frac{\Omega_\omega}{\|\boldsymbol{\Omega}\|}, \tag{2.16}$$

respectively. Applying the 2-D Fourier transform on both sides of Equation (2.14) gives

$$\hat{f}_{\mathcal{R}}(\boldsymbol{\Omega}) = \underbrace{(\hat{h}_t(\boldsymbol{\Omega}) + \mathrm{j}\hat{h}_\omega(\boldsymbol{\Omega}))}_{\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega})} \hat{f}(\boldsymbol{\Omega}), \tag{2.17}$$

where $\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega})$ denotes the frequency response of the CRT kernel:

$$\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega}) \triangleq \hat{h}_t(\boldsymbol{\Omega}) + \mathrm{j}\hat{h}_{\omega}(\boldsymbol{\Omega}) = \frac{-\mathrm{j}\Omega_t + \Omega_{\omega}}{\|\boldsymbol{\Omega}\|}, \tag{2.18}$$

where $\|\boldsymbol{\Omega}\| = \sqrt{\Omega_t^2 + \Omega_{\omega}^2}$. An equivalent polar form is written as follows:

$$\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega}) = e^{\mathrm{j}\tan^{-1}\left(-\frac{\Omega_t}{\Omega_{\omega}}\right)}. \tag{2.19}$$

Equation (2.19) shows that the function $\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega})$ is a phase-only function with unity magnitude and a phase response that resembles a spiral as shown in Figure 2.10. The action of CRT on a function $f(\boldsymbol{\omega})$ is denoted by the Riesz operator $\mathcal{R}$, and its spectral behaviour is described by the following relation:

$$\mathcal{R}f(\boldsymbol{\omega}) \triangleq \frac{-\mathrm{j}\Omega_t + \Omega_{\omega}}{\|\boldsymbol{\Omega}\|}\hat{f}(\boldsymbol{\Omega}). \tag{2.20}$$

The CRT kernel $\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega})$ in Equation (2.18) satisfies the following properties:

(1) It is unitary, which follows directly from the property that $|\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega})|^2 = 1$.
(2) It is anti-symmetric: $\hat{h}_{\mathcal{R}}(-\boldsymbol{\Omega}) = -\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega})$.
(3) Analogous to the 1-D Hilbert transform, the CRT kernel also possesses a singularity at the origin.

In this section, we derive the complex Riesz transform of a 2-D cosine and a 2-D AM-FM cosine signal, by considering the quasi-eigenfunction property.

2.4.1 *Quasi-eigenfunction Property*

Complex exponentials are eigenfunctions of linear, shift-invariant (LSI) systems. In 1-D, the output of an LSI system with frequency response $\hat{h}(\omega)$ corresponding to the input $e^{\mathrm{j}\omega_0 t}$ is given by $\hat{h}(\omega_0)e^{\mathrm{j}\omega_0 t}$. Considering the 2-D counterpart of this property, the output of an LSI system to the input $e^{\mathrm{j}\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle}$ is $\hat{h}(\boldsymbol{\Omega}_0)e^{\mathrm{j}\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle}$ where $\hat{h}(\boldsymbol{\Omega})$ denotes the frequency response of the LSI system. Naturally, an extension of this property can be sought for amplitude and frequency modulated eigenfunctions. This property does not hold for an AM-FM signal. In this case, one can resort to an

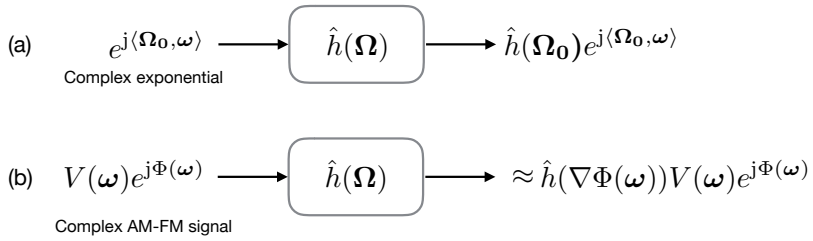


Figure 2.11: Illustration of the (a) eigenfunction property, and (b) quasi-eigenfunction approximation for LSI systems.

approximation which is referred to as the *quasi-eigenfunction* property.

Consider a 2-D AM-FM complex exponential function $V(\boldsymbol{\omega})e^{j\Phi(\boldsymbol{\omega})}$, where $V(\boldsymbol{\omega})$ and $\Phi(\boldsymbol{\omega})$ denote the 2-D AM and phase function, respectively. The quasi-eigenfunction property is based on the following assumptions on the AM and the phase function:

(1) The phase function is of the form $\Phi(\boldsymbol{\omega}) = \langle \mathbf{\Omega}_0, \boldsymbol{\omega} \rangle + \phi(\boldsymbol{\omega})$, where $\phi(\boldsymbol{\omega})$ denotes the nonlinear phase variation about the carrier frequency with $\|\nabla\phi(\boldsymbol{\omega})\| \ll \|\mathbf{\Omega}_0\|$.
(2) The function $V(\boldsymbol{\omega})$ is smooth and its rate of variation is much smaller than that of the carrier frequency $\|\mathbf{\Omega}_0\|$.

Under the above assumptions, $V(\boldsymbol{\omega})e^{j\Phi(\boldsymbol{\omega})}$ can be locally approximated using a 2-D sinusoid and the quasi-eigenfunction approximation can be invoked. The output of an LSI system can be approximated as $\hat{h}(\nabla\Phi(\boldsymbol{\omega}))V(\boldsymbol{\omega})e^{j\Phi(\boldsymbol{\omega})}$. Figure 2.11 summarizes the quasi-eigenfunction property.

We employ the eigenfunction and quasi-eigenfunction properties to determine the Riesz transform a 2-D cosine and 2-D AM-FM cosine, respectively

2.4.2 *Riesz Transform of a 2-D Cosine*

The complex Riesz transform of a 2-D cosine $\cos(t\Omega_0 \cos\beta_0 + \omega\Omega_0 \sin\beta_0)$ is given by

$$\mathcal{R}\{\cos\Omega_0(t\cos\beta_0 + \omega\sin\beta_0)\} = e^{j\beta_0}\sin\Omega_0(t\cos\beta_0 + \omega\sin\beta_0). \tag{2.21}$$

Proof: For notational brevity, we express $t\Omega_0 \cos\beta_0 + \omega\Omega_0 \sin\beta_0 = \langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle$, where $\boldsymbol{\Omega}_0 = [\Omega_0 \cos\beta_0 \ \Omega_0 \sin\beta_0]^{\mathrm{T}}$ and $\boldsymbol{\omega} = [t \ \omega]^T$. Using Euler's formula, we write

$$\cos(\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle) = \frac{1}{2}\big(e^{\mathrm{j}\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle} + e^{-\mathrm{j}\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle}\big). \tag{2.22}$$

Consider the CRT of the complex exponential:

$$\mathcal{R}\left\{e^{\mathrm{j}\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle}\right\} = \hat{h}_{\mathcal{R}}(\nabla\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle) e^{\mathrm{j}\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle} = \hat{h}_{\mathcal{R}}(\boldsymbol{\Omega}_0) e^{\mathrm{j}\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle}, \tag{2.23}$$

where using Equation (2.18), we get

$$\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega}_0) = -\mathrm{j}\cos\beta_0 + \sin\beta_0 = -\mathrm{j}e^{\mathrm{j}\beta_0} \tag{2.24}$$

$$\implies \mathcal{R}\left\{e^{\mathrm{j}\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle}\right\} = -\mathrm{j}e^{\mathrm{j}\beta_0} e^{\mathrm{j}\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle}. \tag{2.25}$$

Similarly, it can be shown that

$$\mathcal{R}\{e^{-\mathrm{j}\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle}\} = \mathrm{j}e^{\mathrm{j}\beta_0} e^{-\mathrm{j}\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle}. \tag{2.26}$$

Combining Equation (2.25) and Equation (2.26) gives the Riesz transform of a 2-D cosine:

$$\mathcal{R}\{\cos(\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle)\} = -\frac{1}{2}\mathrm{j}e^{\mathrm{j}\beta_0}(e^{\mathrm{j}\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle} - e^{-\mathrm{j}\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle}) = e^{\mathrm{j}\beta_0} \sin(\langle \boldsymbol{\Omega}_0, \boldsymbol{\omega} \rangle). \tag{2.27}$$

2.4.3 *Riesz Transform of a 2-D AM-FM Cosine*

Using the quasi-eigenfunction property, the complex Riesz transform of a 2-D AM-FM cosine $V(\boldsymbol{\omega}) \cos\Phi(\boldsymbol{\omega})$ can be approximated as

$$\mathcal{R}\big\{V(\boldsymbol{\omega}) \cos\Phi(\boldsymbol{\omega})\big\} \approx e^{\mathrm{j}\beta(\boldsymbol{\omega})} V(\boldsymbol{\omega}) \sin\Phi(\boldsymbol{\omega}), \tag{2.28}$$

where $\beta(\boldsymbol{\omega}) = \tan^{-1}\left(\frac{\Phi_\omega(\boldsymbol{\omega})}{\Phi_t(\boldsymbol{\omega})}\right)$ with $\Phi_t(\boldsymbol{\omega})$ and $\Phi_\omega(\boldsymbol{\omega})$ denoting the partial derivatives of $\Phi(\boldsymbol{\omega})$ along t-axis and ω-axis, respectively.

Proof: Using the quasi-eigenfunction property, the Riesz transform of an AM-FM

complex exponential function $V(\boldsymbol{\omega})e^{\mathrm{j}\Phi(\boldsymbol{\omega})}$ is written as follows:

$$\begin{aligned}\mathcal{R}\{V(\boldsymbol{\omega})e^{\mathrm{j}\Phi(\boldsymbol{\omega})}\} &\approx \hat{h}_{\mathcal{R}}(\nabla\Phi(\boldsymbol{\omega}))V(\boldsymbol{\omega})e^{\mathrm{j}\Phi(\boldsymbol{\omega})} \\ &= \frac{-\mathrm{j}\Phi_t(\boldsymbol{\omega})+\Phi_\omega(\boldsymbol{\omega})}{\sqrt{\Phi_t^2(\boldsymbol{\omega})+\Phi_\omega^2(\boldsymbol{\omega})}}V(\boldsymbol{\omega})e^{\mathrm{j}\Phi(\boldsymbol{\omega})} \\ &= -\mathrm{j}\frac{\Phi_t(\boldsymbol{\omega})+\mathrm{j}\Phi_\omega(\boldsymbol{\omega})}{\sqrt{\Phi_t^2(\boldsymbol{\omega})+\Phi_\omega^2(\boldsymbol{\omega})}}V(\boldsymbol{\omega})e^{\mathrm{j}\Phi(\boldsymbol{\omega})} \\ &= -\mathrm{j}e^{\mathrm{j}\beta(\boldsymbol{\omega})}V(\boldsymbol{\omega})e^{\mathrm{j}\Phi(\boldsymbol{\omega})}, \end{aligned} \tag{2.29}$$

where $\beta(\boldsymbol{\omega}) = \tan^{-1}\left(\frac{\Phi_\omega(\boldsymbol{\omega})}{\Phi_t(\boldsymbol{\omega})}\right)$ denotes the local orientation. Similarly, we have

$$\mathcal{R}\{V(\boldsymbol{\omega})e^{-\mathrm{j}\Phi(\boldsymbol{\omega})}\} = \mathrm{j}e^{\mathrm{j}\beta(\boldsymbol{\omega})}V(\boldsymbol{\omega})e^{-\mathrm{j}\Phi(\boldsymbol{\omega})}. \tag{2.30}$$

Combining Equation (2.29) and Equation (2.30), results in Equation (2.28).

From Equation (2.28), one can observe that the CRT of a 2-D AM-FM cosine is the product of three terms: the original AM, quadrature component of the carrier sinusoid, and a complex exponential that involves the local orientation. In order to obtain the quadrature component of original AM-FM sinusoid in 2-D, the effect of local orientation must be removed, which is done by multiplying Equation (2.28) by $e^{-\mathrm{j}\beta(\boldsymbol{\omega})}$ on both sides:

$$\underbrace{e^{-\mathrm{j}\beta(\boldsymbol{\omega})}\mathcal{R}}_{\mathcal{V}}\{V(\boldsymbol{\omega})\cos\Phi(\boldsymbol{\omega})\} \approx V(\boldsymbol{\omega})\sin\Phi(\boldsymbol{\omega}), \tag{2.31}$$

where $\mathcal{V}$ is referred to as the *vortex operator* [40] and is defined as the operation of taking Riesz transform followed by the orientation compensation step. In practice, the local orientation is an unknown quantity and must be estimated. We develop the procedure to estimate the orientation.

2.4.4 *Estimation of Local Orientation*

The problem of computing the local orientation is formulated as an optimization problem, which relies on the directional Hilbert transform and its relation to CRT.

Definition 2.4.1 (Directional Hilbert Transform)**.** *The directional Hilbert transform of a function* $f(\boldsymbol{\omega}) : \mathbb{R}^2 \to \mathbb{R}$ *along angle* β *is given by*

$$\mathcal{H}_\beta f(\boldsymbol{\omega}) \triangleq \cos\beta f_t(\boldsymbol{\omega}) + \sin\beta f_\omega(\boldsymbol{\omega}), \tag{2.32}$$

where $f_t(\boldsymbol{\omega}) = (h_t * f)(\boldsymbol{\omega})$ *and* $f_\omega(\boldsymbol{\omega}) = (h_\omega * f_\omega)(\boldsymbol{\omega})$.

The directional Hilbert Transform and the CRT are closely related. Consider a function $f(\boldsymbol{\omega}) : \mathbb{R}^2 \to \mathbb{R}$, then

$$\text{Real}\big\{e^{-\mathrm{j}\beta}\mathcal{R}f(\boldsymbol{\omega})\big\} = \cos\beta f_t(\boldsymbol{\omega}) + \sin\beta f_\omega(\boldsymbol{\omega}) \triangleq \mathcal{H}_\beta f(\boldsymbol{\omega}). \tag{2.33}$$

This property can be proved by expanding the left-hand side:

$$\begin{aligned}\text{Real}\big\{(e^{-\mathrm{j}\beta}\mathcal{R}f)(\boldsymbol{\omega})\big\} &= \text{Real}\big\{(e^{-\mathrm{j}\beta}((\underbrace{h_t * f}_{f_t} + \mathrm{j}\underbrace{h_\omega * f}_{f_\omega}))(\boldsymbol{\omega}))\big\}\\ &= (f_t\cos\beta + f_\omega\sin\beta)(\boldsymbol{\omega})\\ &= \mathcal{H}_\beta f(\boldsymbol{\omega})\end{aligned}$$

Let $\hat{\beta}(\boldsymbol{\omega})$ be an estimate of $\beta(\boldsymbol{\omega})$. Rewriting Equation (2.28) as

$$\mathcal{R}\big\{f(\boldsymbol{\omega})\big\} = \mathcal{R}\big\{V(\boldsymbol{\omega})\cos\Phi(\boldsymbol{\omega})\big\} \approx e^{\mathrm{j}\beta(\boldsymbol{\omega})}V(\boldsymbol{\omega})\sin\Phi(\boldsymbol{\omega}). \tag{2.34}$$

Multiplying both sides of Equation (2.34) with $e^{-\mathrm{j}\hat{\beta}(\boldsymbol{\omega})}$, we have

$$e^{-\mathrm{j}\hat{\beta}(\boldsymbol{\omega})}\mathcal{R}\big\{f(\boldsymbol{\omega})\big\} = e^{\mathrm{j}(\beta(\boldsymbol{\omega})-\hat{\beta}(\boldsymbol{\omega}))}\sin\Phi(\boldsymbol{\omega}). \tag{2.35}$$

Expressing $\mathcal{R}\big\{f(\boldsymbol{\omega})\big\} = f_t(\boldsymbol{\omega}) + \mathrm{j}f_\omega(\boldsymbol{\omega})$ (using (2.14)) and taking the real part gives

$$\underbrace{\text{Real}\Big\{e^{-\mathrm{j}\hat{\beta}(\boldsymbol{\omega})}\mathcal{R}f(\boldsymbol{\omega})\Big\}}_{\mathcal{H}_{\hat{\beta}(\boldsymbol{\omega})}f(\boldsymbol{\omega})} = \cos(\beta(\boldsymbol{\omega}) - \hat{\beta}(\boldsymbol{\omega}))\sin\Phi(\boldsymbol{\omega}). \tag{2.36}$$

Since the directional Hilbert transform in Equation (2.36) is maximum when $\hat{\beta}(\boldsymbol{\omega}) = \beta(\boldsymbol{\omega})$, the orientation of a function $f(\boldsymbol{\omega})$ at location $\boldsymbol{\omega}_0 = (t_0, \omega_0) \in \mathbb{R}^2$ is estimated

by solving the following optimization problem:

$$\hat{\beta}(\boldsymbol{\omega}_0) = \underset{\beta\in[-\pi,\pi]}{\arg\max}(\psi * |\mathcal{H}_\beta f|^2)(\boldsymbol{\omega}_0), \tag{2.37}$$

where ψ is a positive, radially symmetric localizing function. Smoothing with ψ is important to obtain an accurate estimate of the local orientation. Typically, a symmetric Gaussian smoothing is used, and the degree of smoothing can be varied by adjusting the variance of the Gaussian kernel.

From Equation (2.33), we can write

$$(\mathcal{H}_\beta f)(\boldsymbol{\omega}_0) = (f_t \cos\beta + f_\omega \sin\beta)(\boldsymbol{\omega}_0) = (\boldsymbol{f}^{\mathrm{T}}\boldsymbol{u})(\boldsymbol{\omega}_0), \tag{2.38}$$

where $\boldsymbol{u} = [\cos\beta \;\; \sin\beta]^{\mathrm{T}}$ denotes a unit vector, and $\boldsymbol{f} = [f_t(\boldsymbol{\omega}) \;\; f_\omega(\boldsymbol{\omega})]^{\mathrm{T}}$. Using Equation (2.38), we express

$$\begin{aligned}(\psi * |\mathcal{H}_\beta f|^2)(\boldsymbol{\omega}_0) &= (\psi * (\boldsymbol{f}^{\mathrm{T}}\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{f}^{\mathrm{T}}\boldsymbol{u}))(\boldsymbol{\omega}_0)\\ &= (\boldsymbol{u}^{\mathrm{T}}(\psi * \boldsymbol{f}\boldsymbol{f}^{\mathrm{T}})\boldsymbol{u})(\boldsymbol{\omega}_0)\\ &= \boldsymbol{u}^{\mathrm{T}}\boldsymbol{J}(\boldsymbol{\omega}_0)\boldsymbol{u},\end{aligned} \tag{2.39}$$

where the matrix $\boldsymbol{J}(\boldsymbol{\omega}_0)$ is referred to as *structure tensor* [80] and is given by

$$\boldsymbol{J}(\boldsymbol{\omega}_0) = \begin{bmatrix} (\psi * f_t^2)(\boldsymbol{\omega}_0) & (\psi * f_\omega f_t)(\boldsymbol{\omega}_0) \\ (\psi * f_\omega f_t)(\boldsymbol{\omega}_0) & (\psi * f_\omega^2)(\boldsymbol{\omega}_0) \end{bmatrix}. \tag{2.40}$$

Using Equation (2.39), the optimization problem in Equation (2.37) can be expressed equivalently as follows:

$$\hat{\beta}(\boldsymbol{\omega}_0) = \underset{\beta\in[-\pi,\pi]}{\arg\max}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{J}(\boldsymbol{\omega}_0)\boldsymbol{u}). \tag{2.41}$$

The solution of Equation (2.41) is the eigenvector of matrix $\boldsymbol{J}(\boldsymbol{\omega})$ corresponding to its maximum eigenvalue [81]. The closed-form solution of Equation (2.41) is given by

$$\hat{\beta}(\boldsymbol{\omega}_0) = \frac{1}{2}\tan^{-1}\left(\frac{2\times(\psi * f_t f_\omega)(\boldsymbol{\omega}_0)}{(\psi * f_\omega^2)(\boldsymbol{\omega}_0) - (\psi * f_t^2)(\boldsymbol{\omega}_0)}\right). \tag{2.42}$$

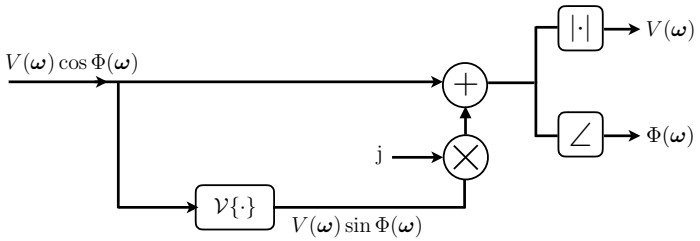


Figure 2.12: Block diagram illustrating CRT-based demodulation of a 2-D AM-FM cosine.

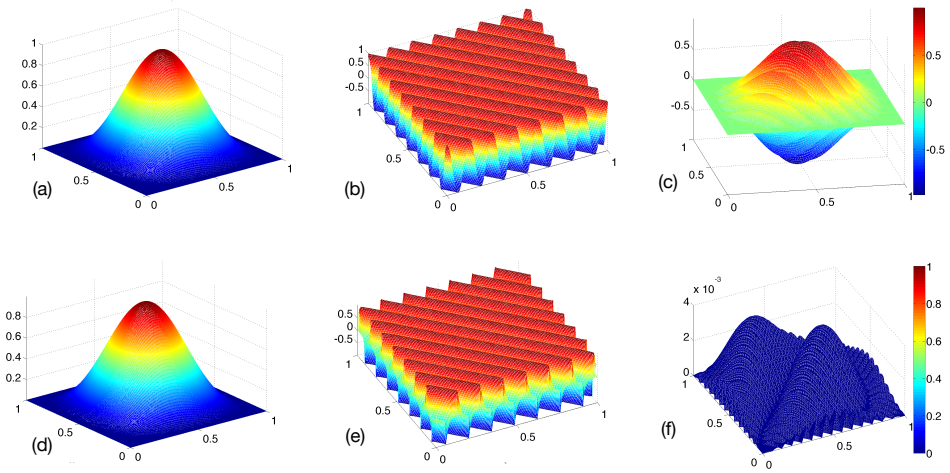


Figure 2.13: (Color online) Demodulation of an amplitude modulated 2-D cosine using CRT: (a) Amplitude modulation obtained as the outer product of a 1-D Hamming window function, (b) original carrier, (c) amplitude modulated carrier, (d) estimated amplitude modulation, (e) estimated carrier signal, and (f) the error in amplitude modulation estimation. The estimation error in the carrier was also found to be of the same order as the error in amplitude estimation.

2.4.5 *Demodulation of 2-D AM-FM Cosine Using CRT*

The key result obtained in Equation (2.31) is used to demodulate a 2-D AM-FM cosine. The quadrature component given by Equation (2.31) is combined with the original AM-FM cosine in complex number format that gives the so-called *monogenic*

signal in higher dimensions [82, 83]:

$$S_a(\boldsymbol{\omega}) = V(\boldsymbol{\omega}) \cos \Phi(\boldsymbol{\omega}) + \mathrm{j} V(\boldsymbol{\omega}) \sin \Phi(\boldsymbol{\omega}) = V(\boldsymbol{\omega}) e^{\mathrm{j}\Phi(\boldsymbol{\omega})}. \tag{2.43}$$

The quantities $V(\boldsymbol{\omega})$ and $\Phi(\boldsymbol{\omega})$ are obtained from $S_a(\boldsymbol{\omega})$ as follows:

$$V(\boldsymbol{\omega}) = |S_a(\boldsymbol{\omega})|, \text{ and} \tag{2.44}$$

$$\Phi(\boldsymbol{\omega}) = \angle S_a(\boldsymbol{\omega}). \tag{2.45}$$

Figure 2.12 shows the block diagram for the demodulation of a 2-D AM-FM cosine using CRT. Figure 2.13 illustrates an example of 2-D AM-FM signal and its AM and FM components estimated by employing the CRT-based demodulation.

2.5 Estimation of Multicomponent 2-D AM-FM Model Parameters

We describe the estimation of the unknown parameters of the multicomponent-patch model described in Section 2.3.2. We follow a two-step procedure for estimating the model parameters. From Equation (2.13), one can observe that the AM and FM components appear in product form in the bandpass term. Hence, in the first step, we estimate the AM and FM components from the fundamental bandpass component in Equation (2.13) by employing CRT-based demodulation. In the next step, the model coefficients $\{\alpha_k\}_{k=1}^{K}$ are estimated using least-squares regression relying on the estimates of AM and FM from the previous step. Section 2.5.3 addresses the choice of the optimum value of the model order K.

2.5.1 *Demodulation of Speech Spectrogram Using CRT*

We divide a speech spectrogram into localized spectrotemporal patches of size 100 ms $\times$ 600 Hz and multiply each patch by a 2-D Hamming window, the patch size along frequency axis is chosen such that at least two pitch harmonics are included even

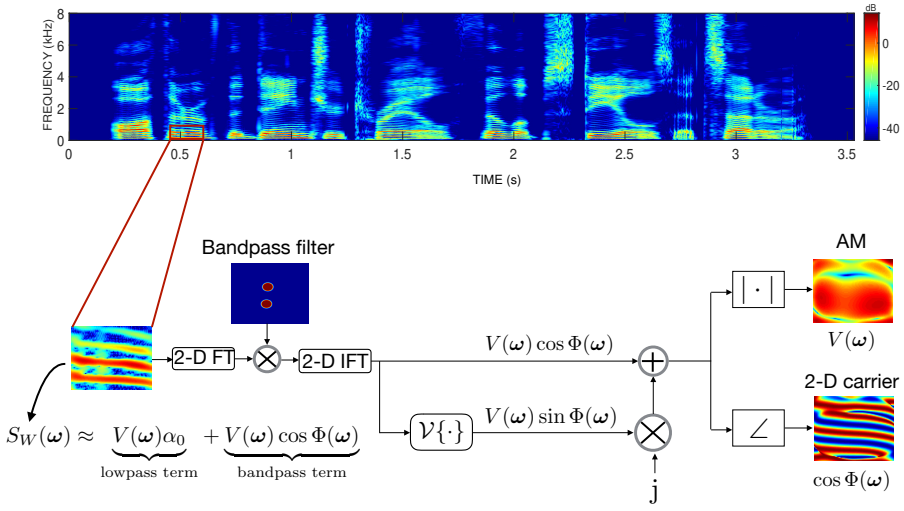
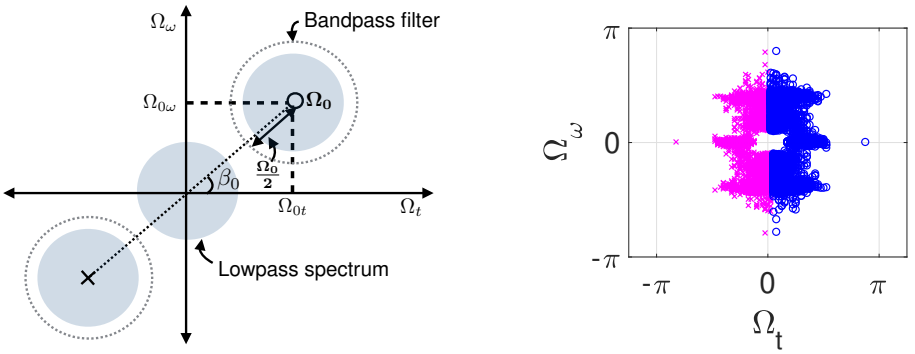


Figure 2.14: Demodulation of a spectrogram patch using CRT.



(a) (b)

Figure 2.15: (a) (Color online) Illustration of the 2-D bandpass filter placement in the GCT domain and its bandwidth for $0 < \alpha < 1$. The filter is always placed at the dominant peak $(\Omega_{0t}, \Omega_{0\omega}) = (\Omega_0 \cos\beta_0, \Omega_0 \sin\beta_0)$, where $\Omega_0 = \sqrt{\Omega_{0t}^2 + \Omega_{0\omega}^2}$ from the origin. A similar argument holds for the filter placement in second and third quadrants and (b) the distribution of 2-D BPF center locations in GCT plane corresponding to the spectrogram patches of a female speech utterance, "*Author of The Danger Trail, Philip Steels, etc.*" A pair of peaks for a patch is marked by a combination of $\times$ (red) and $\circ$ (blue).

for high pitch sounds such as female voices. Consecutive patches have an overlap of 75% and 35% along time and frequency axes, respectively. Since the spectrogram is a non-negative quantity, a patch has DC component, which must be removed before performing demodulation. This is done by passing the patch through a 2-D bandpass filter. The filtered patch is then subjected to CRT-based demodulation. Figure 2.14 illustrates demodulation of a speech spectrogram patch using CRT. Next, we describe the design of the 2-D bandpass filter.

2.5.1.1 2-*D Bandpass Filter*

The GCT of a voiced spectrogram patch shows multiple dominant peaks, which include peaks at the zero frequency, the spatial frequency Ω_0, and its harmonics (see Figure 2.8). We use a 10^{th}-order 2-D Butterworth filter designed with its center frequency located at the second dominant peak in the GCT domain, which occurs at the spatial frequency $\mathbf{\Omega}_0$ as shown in Figure 2.15(a). Centered at $\mathbf{\Omega}_0 = (\Omega_{0t}, \Omega_{0\omega}) \in \mathbb{R}^2$, the bandwidth of the bandpass filter (BPF) is chosen to be $\alpha\Omega_0$ where $0 < \alpha < 1$ and $\Omega_0 = \sqrt{\Omega_{0t}^2 + \Omega_{0\omega}^2}$. The transfer function of an n^{th}-order circular Butterworth 2-D BPF filter centered at $\mathbf{\Omega}_0$ is given by

$$H(\Omega_t, \Omega_\omega) = \frac{1}{1 + \left(\frac{R(\Omega_t, \Omega_\omega)}{\Omega_c}\right)^{2n}}, \tag{2.46}$$

where $R(\Omega_t, \Omega_\omega) = \sqrt{(\Omega_t - \Omega_{0t})^2 + (\Omega_\omega - \Omega_{0\omega})^2}$, and $\Omega_c \in \mathbb{R}_{>0}$ denotes the cut-off frequency of the filter. The optimum value of α is obtained empirically by evaluating the model accuracy on a speech database for which the details are provided in Section 2.6.3. For unvoiced spectrogram patches, the location of the dominant peak occurs at random locations. Figure 2.15(b) shows the distribution of the center locations of BPF in GCT plane for the spectrogram patches of a continuous speech utterance spoken by a female speaker. The peak-pairs along Ω_t-axis mostly correspond to unvoiced patches, whereas the peak-pairs corresponding to voiced patches occur in either the first and third quadrants or second and fourth quadrants

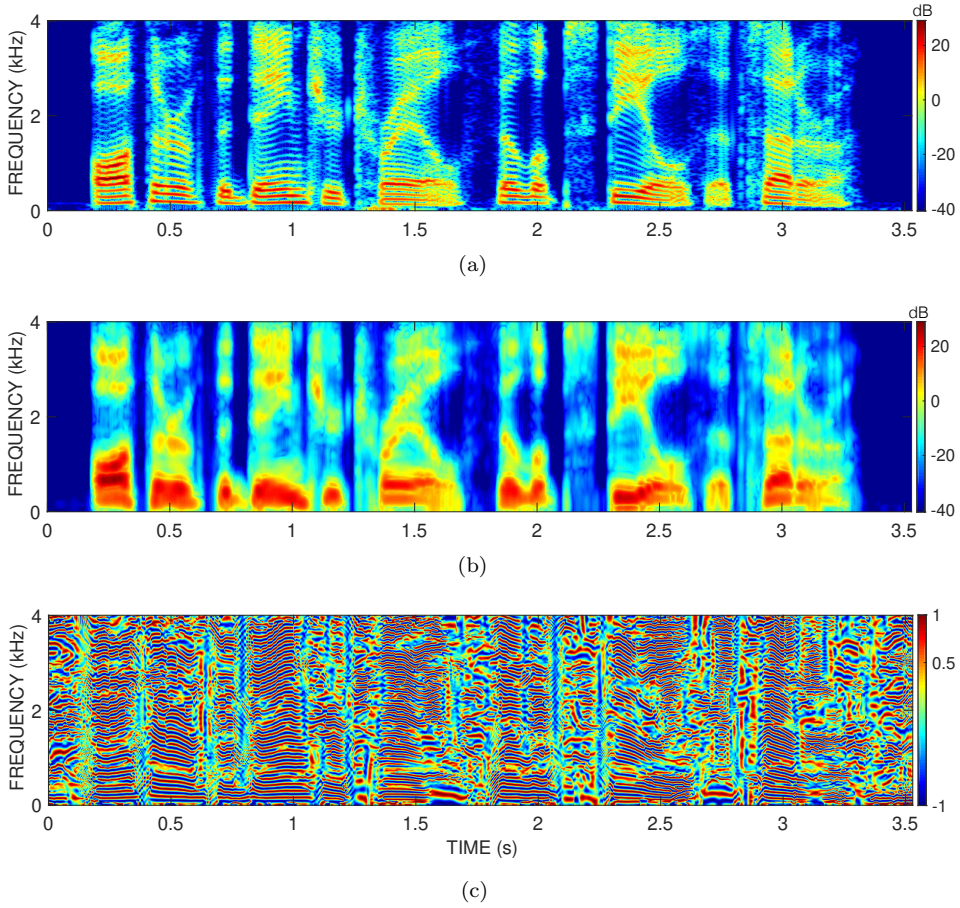


Figure 2.16: The t-f maps of (a) the spectrogram, (b) AM, and (c) the corresponding 2-D carrier for a speech utterance, "*Author of the danger trail, Philip Steels, etc.*," spoken by a male speaker.

depending on whether the pitch is decreasing or increasing, respectively. The occurrence of peak-pair along Ω_ω-axis indicates a flat or nearly constant pitch.

The bandpass filter effectively retains the fundamental bandpass term in Equation (2.13) and the estimates of $V(\boldsymbol{\omega})$ and $\cos\Phi(\boldsymbol{\omega})$ for each patch are obtained from this term using Equations (2.43)-(2.45). The spectrotemporal characteristics of a speech signal are more prominent in the full t-f maps of AM and 2-D carrier $\cos\Phi(\boldsymbol{\omega})$. The t-f map of the AM is obtained by combining the AM components of

all the patches using 2-D overlap-add in least-squares sense (2-D OLA-LSE) (Appendix B). The same procedure is followed to obtain the t-f map of the 2-D carrier. Figure 2.16 illustrates the t-f maps of the AM and the 2-D carrier $\cos\Phi(\boldsymbol{\omega})$. We observe that the AM captures slowly varying magnitude response of the vocal-tract filter/formant-structure and the 2-D carrier predominantly exhibits the excitation characteristics such as pitch and the evolution of its harmonics in the t-f plane.

2.5.2 *Estimation of the Model Coefficients*

After obtaining estimates of $V(\boldsymbol{\omega})$ and $\Phi(\boldsymbol{\omega})$ for a patch, the parameter set $\boldsymbol{\theta}$ is estimated using least-squares regression. Let $\mathbf{m} = (\mathbf{l}, \mathbf{k})$ denote the discrete counterpart of $\boldsymbol{\omega} = (t, \omega)$. A spectrogram patch $S_W(\mathbf{m})$ is vectorized to a column vector $\boldsymbol{s}$. Similarly, the column vectors corresponding to $V(\mathbf{m})$ and $\Phi(\mathbf{m})$ are denoted by $\boldsymbol{v}$ and $\boldsymbol{\phi}$, respectively. The coefficients $\boldsymbol{\theta} = [\alpha_0, \alpha_1, \cdots, \alpha_K]^{\mathrm{T}}$ are obtained by solving the following problem:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \left\| \boldsymbol{s} - \boldsymbol{v} \odot \left(\alpha_0 + \sum_{j=1}^{K} \alpha_j \cos j\boldsymbol{\phi} \right) \right\|^2, \tag{2.47}$$

where $\odot$ denotes element-wise product. Taking derivative of the cost function with respect to $\boldsymbol{\theta}$ in Equation (2.47) and equating it to zero gives a set of $K+1$ linear equations:

$$\underbrace{\begin{bmatrix} \|\boldsymbol{v}_0\|^2 & \boldsymbol{v}_0^{\mathrm{T}}\boldsymbol{v}_1 & \cdots & \boldsymbol{v}_0^{\mathrm{T}}\boldsymbol{v}_K \\ \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_0 & \|\boldsymbol{v}_1\|^2 & \cdots & \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_K \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{v}_K^{\mathrm{T}}\boldsymbol{v}_0 & \boldsymbol{v}_K^{\mathrm{T}}\boldsymbol{v}_1 & \cdots & \|\boldsymbol{v}_K\|^2 \end{bmatrix}}_{A_{(K+1)\times(K+1)}} \underbrace{\begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix}}_{\boldsymbol{\theta}_{(K+1)\times 1}} = \underbrace{\begin{bmatrix} \boldsymbol{v}_0^{\mathrm{T}}\boldsymbol{s} \\ \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{s} \\ \vdots \\ \boldsymbol{v}_K^{\mathrm{T}}\boldsymbol{s} \end{bmatrix}}_{\boldsymbol{b}_{(K+1)\times 1}}, \tag{2.48}$$

where $\boldsymbol{v}_j = \boldsymbol{v} \odot \cos j\boldsymbol{\phi}$ for $j = 0, 1, 2, \ldots K$. The closed-form least-squares solution to Equation (2.48) is given by $\boldsymbol{\theta}^* = A^{\dagger}\boldsymbol{b}$, where $A^{\dagger}$ denotes the pseudo-inverse of A. An illustration of the obtained AM, FM, and $\boldsymbol{\theta}^*$ for a voiced spectrogram patch is shown in Figure 2.17.

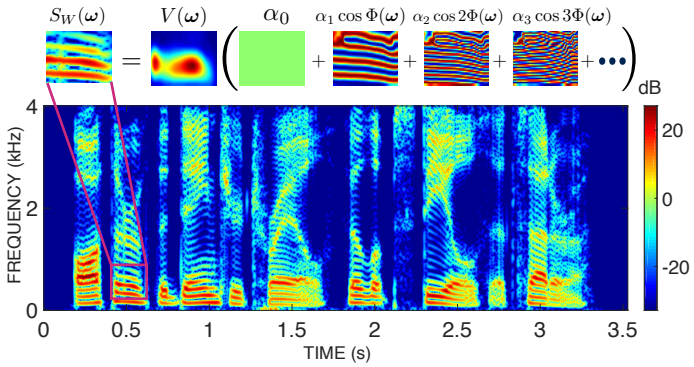


Figure 2.17: (Color online) A narrowband spectrogram of a male speech utterance. The decomposition of a voiced patch into its AM-FM components using the multicomponent AM-FM model. The estimated model coefficients for the patch were $\alpha_0 = 0.83, \alpha_1 = 1, \alpha_2 = 0.26$, and $\alpha_3 = 0.01$.

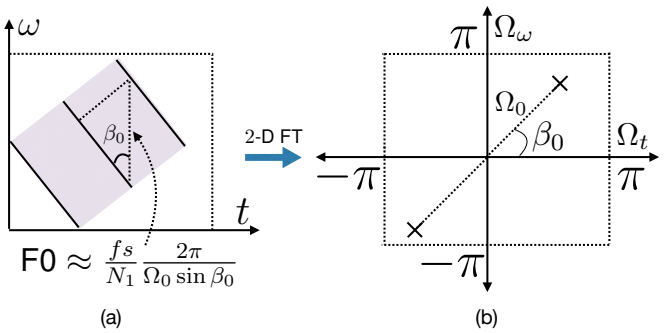


Figure 2.18: (a) Schematic for a voiced spectrogram patch, and (b) its GCT with dominant peak illustrated by a symbol "×".

2.5.3 *Choice of the Model Order*

We show that the optimal choice of the model order is proportional to the instantaneous fundamental frequency (pitch) of the speaker, and hence the model order must be adapted to the pitch. Figure 2.18 depicts the fanning structure of the harmonic lines corresponding to a voiced spectrogram patch and the corresponding pair of peaks in its GCT (marked with ×). If the patch size is small enough such that the harmonic lines can be assumed to be approximately parallel to the local variations of the speaker's fundamental frequency F0, then with reference to the schematic

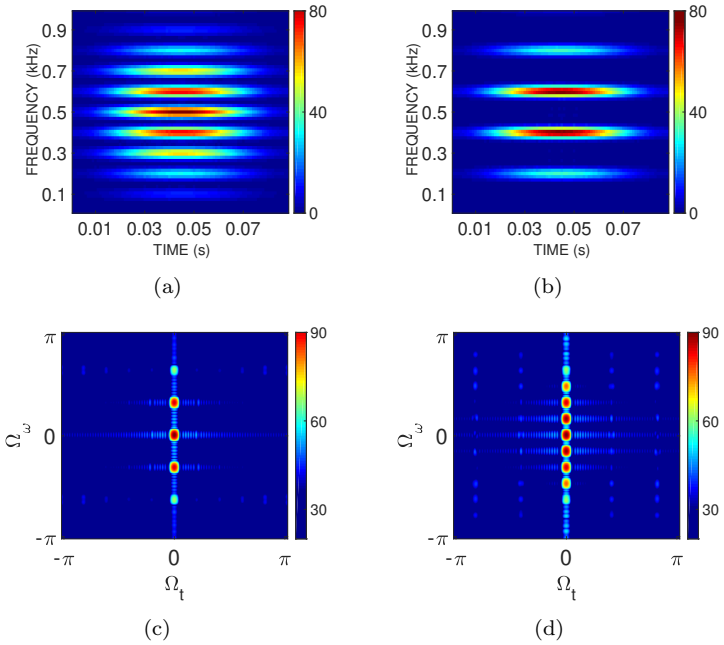


Figure 2.19: (Color online) Spectrogram patches (of size 1 kHz $\times$ 80 ms) of a signal consisting of a sum of harmonically related sinusoids with fundamental frequency (a) $f_0 = 100$ Hz, (b) $f_0 = 200$ Hz and their corresponding GCTs in (c) and (d). The spectrogram was computed with a 40 ms Hamming window with a frameshift of 1 ms and 512 FFT points. The signal sampling frequency is 8 kHz. The figure shows that, for a given spectrogram patch size, a signal with higher fundamental frequency has more number of peaks in GCT plane than a signal with lower fundamental frequency.

shown in Figure 2.18, we have

$$\text{F0} \approx \frac{f_s}{N_1} \frac{2\pi}{\Omega_0 \sin \beta_0}, \quad \beta_0 \in (0, \pi), \tag{2.49}$$

where f_s and N_1 denote the sampling frequency of the speech signal and the number of FFT points used for computing the STFT, respectively. For a given spectrogram patch, the model order K is the count of the number of peaks that occur at fundamental frequency and its harmonics in the GCT domain within the Nyquist

frequency bounds. Hence, the model order K is given by

$$K = \min\left\{\left\lfloor\frac{\pi}{\Omega_0 \sin\beta_0}\right\rfloor, \left\lfloor\frac{\pi}{\Omega_0|\cos\beta_0|}\right\rfloor\right\}. \tag{2.50}$$

Using Equation (2.49), one can express the model order K in terms of F0 as follows

$$K = \min\left\{\left\lfloor\frac{N_1}{2f_s}\mathrm{F0}\right\rfloor, \left\lfloor\frac{N_1|\tan\beta_0|}{2f_s}\mathrm{F0}\right\rfloor\right\}, \tag{2.51}$$

which shows that, for given f_s, N_1 and β_0, the model order is directly proportional to F0. A higher value of F0 results in a higher value of K. For illustration, consider the signal

$$s(t) = \sum_{j=1}^{J} \sin(2\pi j f_0 t), \tag{2.52}$$

which is a sum of J harmonically related sinusoids with fundamental frequency f_0. We consider a spectrogram patch of dimension 1 kHz $\times$ 80 ms. Figure 2.19 shows the patches and their GCTs for two cases: (1) $f_0 = 100$ Hz, and (2) $f_0 = 200$ Hz with $J = 10$. From the figure, one can observe that, for a fixed patch size, the number of peaks in the GCT domain becomes double when the frequency f_0 is increased from 100 Hz to 200 Hz. Hence, for higher f_0 values, the model order K can take on high values in accordance with Equation (2.51). Therefore, the model order in Equation (2.13) should be chosen in a pitch-adaptive manner based on the relation given in Equation (2.51). For instance, female speakers have a higher F0 and correspondingly a higher K compared to male speakers.

2.5.3.1 *Upper Bound on the Model Order*

Irrespective of the values taken by $\beta_0 \in (0, \pi)$, the upper bound on the model order K using Equation (2.51) is given by

$$K \leq \left\lfloor\frac{N_1}{2f_s}\mathrm{F0}_{\max}\right\rfloor, \tag{2.53}$$

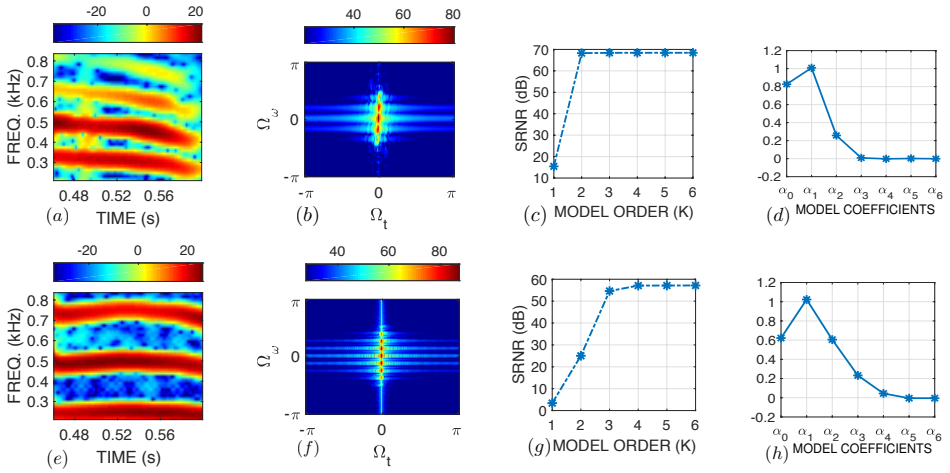


Figure 2.20: (Color online) Illustration of a spectrogram patch, its 2-D Fourier transform, SRNR values with respect to K and the values of model coefficients when the patch is subjected to multicomponent modeling with a fixed model order $K = 6$. The first row corresponds to male speakers and the second one to female speakers.

where $\text{F0}_{\max}$ is the maximum fundamental frequency of the speaker. Let $K_{\max} = \left\lfloor \frac{N_1}{2f_s}\text{F0}_{\max} \right\rfloor$. Consider an example with $N_1 = 512, f_s = 8000$ Hz and a female speaker with $\text{F0}_{\max}$= 300 Hz, then using Equation (2.53), we get $K_{\max} = 8$. On the other hand, the 2-D Fourier transform of an unvoiced patch does not show harmonically separated peaks. Assuming Ω_0 corresponds to the highest peak location in an unvoiced patch, we use the same criterion for selecting the model order as given in Equation (2.53).

2.5.4 *Model Order versus Model Accuracy*

We conduct a preliminary experiment to compute the accuracy of the proposed model versus model order for a voiced spectrogram patch. Two patches having same dimensions are considered from the narrowband spectrograms corresponding to speech utterances spoken by male and female speakers. The spectrogram patch modeling accuracy is measured using the objective measure, *Signal-to-Reconstruction*

Noise-Ratio (SRNR), which is defined as follows:

$$\text{SRNR} = 20\log_{10}\left(\frac{\|S(\boldsymbol{\omega})\|}{\|S(\boldsymbol{\omega}) - \hat{S}(\boldsymbol{\omega})\|}\right)\text{dB}, \tag{2.54}$$

where $S(\boldsymbol{\omega})$ and $\hat{S}(\boldsymbol{\omega})$ denote the original and reconstructed spectrogram patches, respectively. Figure 2.20 shows two spectrogram patches corresponding to a male speaker and a female speaker and their corresponding GCTs. The figure also shows the SRNR values versus model order. The highest model order was set using Equation (2.53) with F0_{max} set to the average pitch of the female speaker, which was obtained using the standard software *Praat* [84]. The figure shows that the SRNR improves by increasing the model order for both the speakers. For male speakers, the SRNR improves by about 55 dB when the model order is increased from $K = 1$ to $K = 2$. For female speakers, the SRNR improves by about 50 dB when the model order increases from $K = 1$ to $K = 4$. Also, we see a saturation in SRNR beyond a certain value of K. Because female speakers have a higher F0, the saturation occurs at a higher value of K (cf. Equation (2.53)) than for a male speaker. This also implies that more the of 2-D cosines (equivalently higher model order) are required for modeling a high-pitched sound. The figure also shows the variations of the model coefficients when the highest model order was set to $K = 6$. Observe that the model coefficients of the male and female speakers patches under consideration are close to zero for $K > 2$ and $K > 4$, respectively. In conclusion, this experiment shows that the model order must be chosen depending on the significant harmonic peaks present within the 2-D Fourier transform of a patch.

Evaluation of the multicomponent AM-FM model on a speech database is reported in Section 2.6.

2.5.5 *Multicomponent AM-FM Decomposition of a Spectrogram Patch*

We now show that the model error decreases as the model order increases.

Algorithm 2.1 Multicomponent decomposition of a speech spectrogram

Step 1: Input the spectrogram patch $S_W(\boldsymbol{\omega})$ to a bandpass filter designed to pick the fundamental bandpass component $V(\boldsymbol{\omega})\cos\Phi(\boldsymbol{\omega})$.
Step 2: Using Riesz transform based demodulation, estimate AM $V(\boldsymbol{\omega})$ and FM $\Phi(\boldsymbol{\omega})$.
Step 3: Determine the optimum model order K using Equation (2.50).
Step 4: Find the optimum model coefficients $\boldsymbol{\theta}^* = \{\alpha_k^*\}_{k=0}^K$ by solving the following optimization cost:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \left\| S_W(\boldsymbol{\omega}) - V(\boldsymbol{\omega})\left(\alpha_0 + \sum_{j=1}^{K} \alpha_j \cos j\phi(\boldsymbol{\omega})\right)\right\|_{\mathrm{F}}^2.$$

Outputs: $\boldsymbol{\theta}^*$, AM component $V(\boldsymbol{\omega})$, and 2-D carriers $\{\cos\Phi(\boldsymbol{\omega}), \cos 2\Phi(\boldsymbol{\omega}), \ldots, \cos K\Phi(\boldsymbol{\omega})\}$.

Theorem 2.5.1. *The least-squares error for a K^{th}-order model is given by*

$$E_K(\alpha_0, \ldots, \alpha_K) = \iint \left(S_W(\boldsymbol{\omega}) - V(\boldsymbol{\omega}) \sum_{k=0}^{K} \alpha_k \cos k\Phi(\boldsymbol{\omega})\right)^2 \mathrm{d}\boldsymbol{\omega}. \tag{2.55}$$

The model coefficients $\{\alpha_k\}_{k=0}^K$ are obtained by solving the following optimization problem:

$$\arg\min_{\alpha_0,\ldots,\alpha_K} \quad E_K(\alpha_0, \ldots, \alpha_K). \tag{2.56}$$

Let $(\tilde{\alpha}_0^*, \ldots, \tilde{\alpha}_K^*)$ denote the optimum solution in Equation (2.56). Also, let $(\alpha_0^*, \ldots, \alpha_{K+1}^*)$ be the optimum solution set for model order $(K+1)$.
The claim is that

$$E_{K+1}(\alpha_0^*, \ldots, \alpha_{K+1}^*) \leq E_K(\tilde{\alpha}_0^*, \ldots, \tilde{\alpha}_K^*); \quad K = 1, 2, 3, \ldots,$$

i.e., the least-squares error with a $(K+1)^{th}$-order model is lower than that of a K^{th}-order model.

Proof: Since $(\alpha_0^*, \ldots, \alpha_{K+1}^*)$ is the optimal solution to the $(K+1)^{th}$-order model, we have

$$E_{K+1}(\alpha_0^*, \ldots, \alpha_{K+1}^*) \leq E_{K+1}(\alpha_0, \ldots, \alpha_{K+1}), \quad \forall \alpha_0, \ldots, \alpha_{K+1}. \tag{2.57}$$

Further, with $\alpha_{k+1} = 0$, we have

$$E_{K+1}(\alpha_0, \ldots, \alpha_K, 0) = E_K(\alpha_0, \cdots, \alpha_K), \quad \forall \alpha_0, \ldots, \alpha_K. \tag{2.58}$$

Setting $(\alpha_0, \ldots, \alpha_K, \alpha_{K+1}) = (\tilde{\alpha}_0^*, \ldots, \tilde{\alpha}_K^*, 0)$ in Equation (2.57) gives

$$E_{K+1}(\alpha_0^*, \ldots, \alpha_{K+1}^*) \leq E_{K+1}(\tilde{\alpha}_0^*, \ldots, \tilde{\alpha}_K^*, 0). \tag{2.59}$$

Plugging Equation (2.58) in Equation (2.59), we get

$$E_{K+1}(\alpha_0^*, \ldots, \alpha_{K+1}^*) \leq E_K(\tilde{\alpha}_0^*, \ldots, \tilde{\alpha}_K^*), \tag{2.60}$$

which is the desired result.

2.6 Performance Evaluation on Speech Data

2.6.1 *Objective Measures*

We use four objective measures to quantify the accuracy of the proposed model. The first three are computed between the input speech signal and the reconstructed signal. Higher values of these scores reflect a better model accuracy. The fourth one quantifies the error in demodulation — the lower it is, the better is the modeling accuracy.

(1) Global signal-to-noise ratio (GSNR): GSNR quantifies the reconstruction error in the time domain and is given by

$$\text{GSNR} = 20 \log_{10} \left(\frac{\|x(t)\|}{\|x(t) - \hat{x}(t)\|} \right) \text{dB},$$

where $x(t)$ and $\hat{x}(t)$ denote the original speech signal and reconstructed speech signal, respectively.

(2) Average frame-wise (or segmental) signal-to-noise ratio (SSNR): SSNR is obtained by averaging the frame-wise SNR over speech frames of duration 20 ms. Prior to computing SSNR, the silence region are removed using a short-time

energy based detector.

(3) Perceptual evaluation of speech quality (PESQ).

(4) Patch error: Patch error measures the error between the original and reconstructed spectrogram patch relative to the energy in the original patch $S_W(\mathbf{m})$ and is given by

$$\zeta_p = \frac{\sum_{\mathbf{m}} |S_W(\mathbf{m}) - \tilde{S}_W(\mathbf{m})|^2}{\sum_{\mathbf{m}} |S_W(\mathbf{m})|^2}, \tag{2.61}$$

where $\tilde{S}_W(\mathbf{m})$ is the reconstructed spectrogram patch obtained using the estimated AM and weighted carriers.

2.6.2 *Database and Experimental Settings*

We use the Starkey database [57], which has 8 male and 8 female American speakers, reading the standard *rainbow passage* [58]. We randomly pick 5 speech utterances (each about 4 s long) for each of the speakers, thus giving rise to a total of 40 male and 40 female utterances. Additionally, the database has speakers with different voice quality, speaking style, and pitch. The speech signals in the database are downsampled to 8 kHz. A narrowband spectrogram is computed using a Hamming window with frame update interval of 1 ms. We next address the optimum length of the analysis window to compute STFT and the bandwidth factor for 2-D bandpass filter as alluded to in Section 2.5.1.1.

2.6.3 *Optimum Duration of the Analysis Window and Bandwidth of the 2-D Bandpass Filter*

We vary the duration of 1-D analysis window and the 2-D bandpass filter bandwidth and analyze the impact of the parameters on the average model accuracy evaluated over speech waveforms taken from the Starkey database. The spectrogram is subjected to the proposed multicomponent modeling and the model parameters (AM, FM, and $\boldsymbol{\theta}$) are estimated for each spectrogram patch. An approximation of

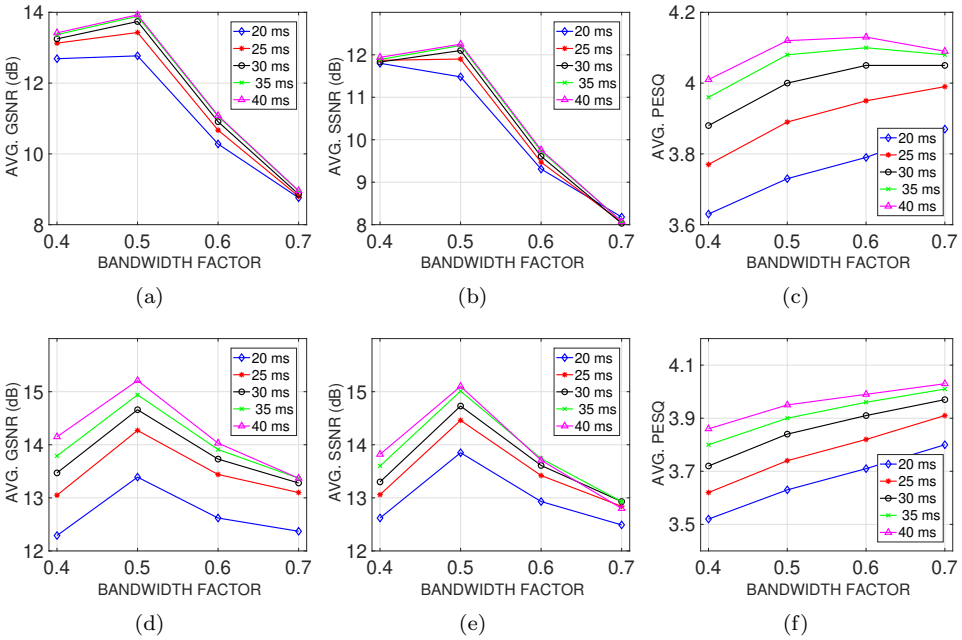


Figure 2.21: (Color online) Average values of objective scores on Starkey database for varying duration of analysis window with respect to bandwidth factor α. The first row corresponds to the male speakers and the second one to the female speakers.

the spectrogram patch is obtained using Equation (2.13). The patches are combined using OLA-LSE in 2-D to obtain the full spectrogram matrix. The reconstructed spectrogram using 2-D OLA-LSE is combined with the original STFT phase. The time-domain speech signal is reconstructed using inverse short-time Fourier transform and the overlap-add of the speech frames.

We compute the average objective measures (see Section. 2.6.1) between the original and reconstructed speech signals for varying window duration and filter bandwidth factor α. Taking into consideration the typical duration of the window used for short-time analysis of speech signals, we vary it from 20 ms to 40 ms in steps of 5 ms. Figure 2.21 shows the average values of the objective measures as a function of the bandwidth factor α that varies from 0.4 to 0.7 in steps of 0.1 for different window durations. From the figure, an improvement in all the three objective measures is

observed when α varies from 0.4 to 0.5 irrespective of the window duration and the speaker's gender. For both male and females speakers, there is significant degradation in the model performance for $\alpha > 0.5$ and the highest model accuracy is achieved for $\alpha = 0.5$ and window duration of 40 ms. The PESQ scores show an increasing trend with increasing bandwidth factor, however, the variations are small. One can observe that for a given value of α, GSNR and SSNR scores are better for window duration 40 ms than any other choice of the window duration with no significant performance gain when the window duration is changed from 30 ms to 40 ms. Based on these observations, a reasonable choice of bandwidth factor and window duration can be made for a relatively high model accuracy irrespective of the gender. We conclude that the optimum values of bandwidth factor and window duration can be set to be 0.5 and 40 ms, respectively.

2.6.4 *Performance Comparison: Monocomponent Versus Multicomponent Model*

2.6.4.1 *Highest achievable model accuracy without demodulation*

We use continuous speech utterances from Starkey database and evaluate the performance of spectrogram analysis and synthesis steps without demodulation. In the analysis step, the spectrogram is divided into overlapping patches, each patch is multiplied by a 2-D window, which is followed by synthesis step where the patches are stitched back using 2-D OLA-LSE (Appendix B). The performance of 2-D OLA-LSE is evaluated using the objective measures GSNR, SSNR, and PESQ. This gives the upper limits of the objective measures without subjecting a spectrogram to demodulation.

Table 2.1 shows the objective scores. High values of GSNR and SSNR indicate a high reconstruction accuracy of the speech waveforms, and a high value of PESQ indicates that the reconstructed speech is of high perceptual quality.

Table 2.1: Average values of objective scores for spectrogram reconstruction after splitting and 2-D OLA-LSE for Starkey database.

	GSNR (dB)	SSNR (dB)	PESQ
Female	56.23 ± 1.77	72.07 ± 1.38	4.45 ± 0.03
Male	55.64 ± 1.73	73.08 ± 1.45	4.46 ± 0.02

Table 2.2: Performance comparison between monocomponent and multicomponent models on Starkey database.

	Male speakers		Female speakers	
	Monocomponent	Multicomponent	Monocomponent	Multicomponent
GSNR (dB)	12.23 ± 1.66	**14.00 ± 1.60**	11.96 ± 1.53	**15.37 ± 2.01**
SSNR (dB)	10.50 ± 1.20	**12.30 ± 1.24**	11.23 ± 1.02	**15.14 ± 1.69**
PESQ	3.89 ± 0.13	**4.13 ± 0.10**	3.33 ± 0.20	**3.94 ± 0.14**

2.6.4.2 *Model accuracy on a continuous speech database*

In this section, the evaluation is done by subjecting the spectrogram to demodulation. We apply the proposed model for analysis/synthesis to the continuous speech utterances from the database. In particular, we carry out a performance comparison between a monocomponent model ($K = 1$) and its multicomponent counterpart, where the model order is selected in a pitch-adaptive fashion. The objective scores for both the models are given in Table 2.2. The table shows that the inclusion of higher-order ($K > 1$) weighted cosine carriers indeed improves the model accuracy over a monocomponent model. Also, a multicomponent model gives about 2 to 3 dB improvement over a monocomponent model for male and female speakers, respectively.

2.6.4.3 *The cumulative average normalized count of the model order K across patches*

We mentioned in Section 2.5.3 that the multicomponent model benefits from choosing the model order for each patch in a pitch-adaptive fashion. To elucidate further,

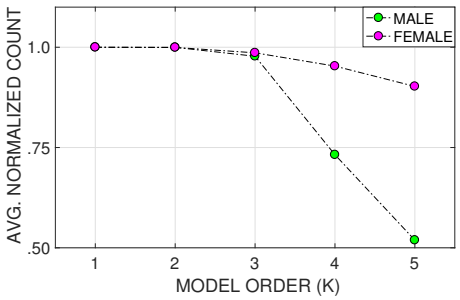


Figure 2.22: Cumulative average normalized count.

we analyze the model order required by different patches obtained while analyzing the waveforms from the database. We propose a measure namely, *the cumulative average normalized count*, given by

$$C_k = \frac{\sum_{j=k}^{K_{\max}} n_j}{\sum_{j=1}^{K_{\max}} n_j}, \tag{2.62}$$

where $k = 1, 2, 3, 4, \cdots, K_{max}$, with K_{max} denoting the maximum model order, n_k denotes the total number of patches with model order k. The cumulative average normalized count C_k is a measure of the occurrence of patches having model order k or higher and satisfies the following relation:

$$C_k > C_{k+1} \quad \text{with} \quad k = 1, 2, 3, \cdots, K_{max} - 1. \tag{2.63}$$

The average normalized count of the model order after pooling all patches is shown in Figure 2.22 with $K_{max} = 5$. The spectrogram patches corresponding to female speakers feature a higher occurrence of model orders 3, 4, and 5, relative to the patches from male speakers. This is attributed to the higher F0 of female speakers. This result shows that it is advantageous to adapt the model order.

2.6.5 *Performance Comparison for All-voiced Speech Utterances*

The performance comparison between monocomponent and multicomponent model is carried out for all-voiced speech utterances taken from TIMIT database [85]. Speech files corresponding to the sentences "S1: Where were you while we we were

Table 2.3: Comparison of global and average frame-wise SNRs of speech reconstructed using a monocomponent and multicomponent model. The higher SNR is indicated in boldface. The speech files were taken from TIMIT database.

Filename	F0 range (Hz)	Global SNR		Avg. frame-wise SNRs	
		Monocomponent	Multicomponent	Monocomponent	Multicomponent
mS1	(120, 188)	19.38	**21.17**	19.1	**20.45**
fS1	(190, 270)	12.17	**22.52**	13.76	**24.48**
mS2	(104, 156)	21.00	**22.38**	20.00	**21.31**
fS2	(199, 310)	10.89	**21.48**	13.07	**23.73**

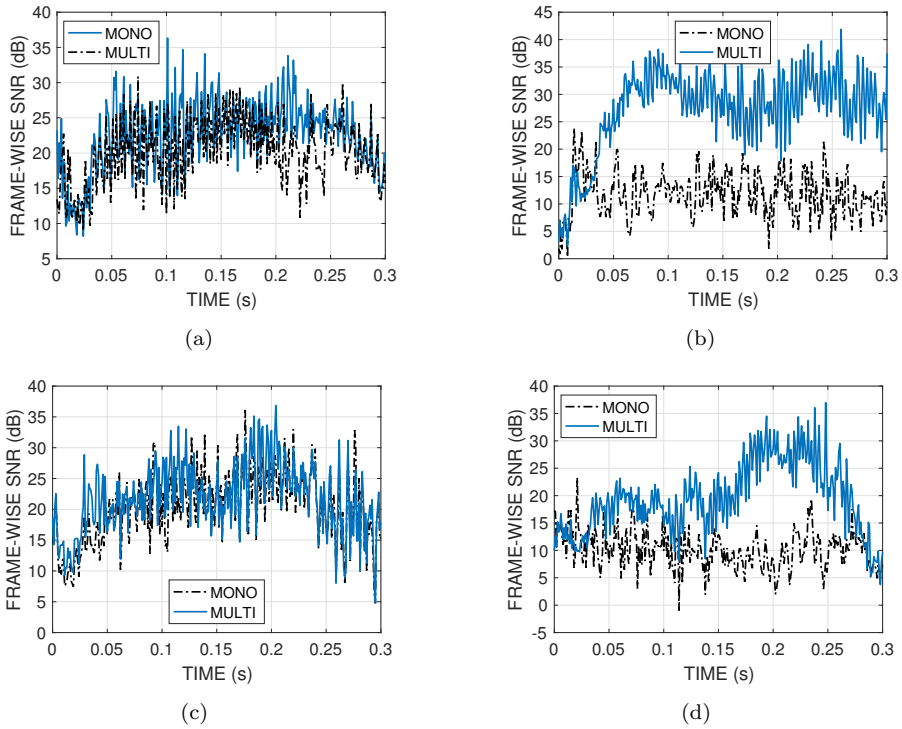


Figure 2.23: (Color online) Frame-wise SNRs of voiced speech files reconstructed using demodulated AM and carriers. The dashed black and thick blue lines correspond to monocomponent and multicomponent model, respectively. (a) mS1, (b) fS1, (c) mS2, and (d) fS2.

away?" and "S2: He will allow a rare lie." were chosen. The identity of male and female speakers is indicated by prefixing 'm' and 'f' to the the sentence label. 'mS1', 'fS1', 'mS2', and 'fS2' correspond to the speakers with TIMIT database IDs 'DAC2',

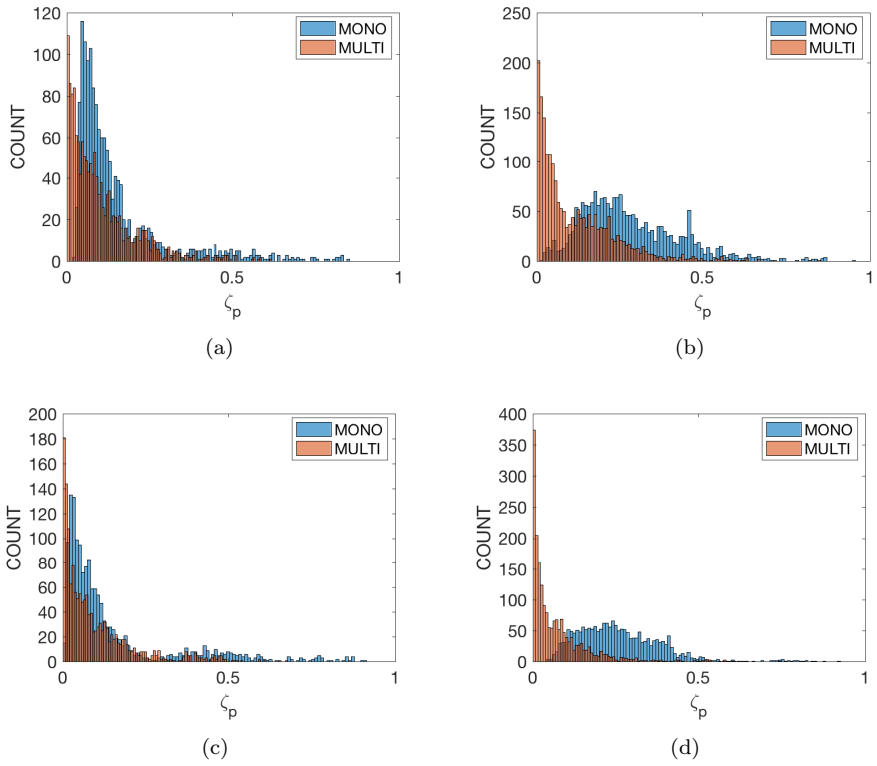


Figure 2.24: (Color online) Histograms of patch error measure ζ_p corresponding to different voiced speech files: (a) mS1, (b) fS1, (c) mS2, (d) fS2.

'GJD0', 'GJF0', and 'JMG0,' respectively. All the speech files were downsampled to 8 kHz and silence regions were manually removed. The speech files were normalized to have a peak time-domain magnitude of unity. From the conclusion drawn in Section 2.6.3, we choose a 40 ms Hamming window for spectrogram computation and bandwidth factor $\alpha = 0.5$.

Table 2.3 shows global and frame-wise SNR values for reconstructed speech files using monocomponent and multicomponent models. One can observe that for the male speakers, the objective scores are about 1 to 2 dB higher for multicomponent model over the monocomponent model. On the other hand, a significant improvement of about 11 dB is observed for female speakers. This is because the average

fundamental frequency of female speakers is higher than that of the male speakers. Hence, choosing model order adapted to pitch variations of the speaker is more effective for high-pitched speech sounds. Figure 2.23 displays the variations of frame-wise SNRs over a time duration of 0 to 0.3 seconds for the voiced speech files. From the figure, we observe that frame-wise SNR is higher for multicomponent model with respect to monocomponent model across the speech frames over the specified time duration. The advantage of the multicomponent model over the monocomponent one is more prominent in the case of female speakers than males speakers.

Figure 2.24 shows histograms of ζ_p for different all voiced speech utterances. We observe that more spectrogram patches take on low values of ζ_p for the multicomponent model than a monocomponent model. This also shows that a multicomponent model gives superior modeling accuracy than a monocomponent model.

2.7 Chapter Summary

We provided a review of 2-D cosines, Fourier transform and 2-D AM-FM cosines. A voiced patch of a speech spectogram was modeled as a 2-D AM-FM signal. We investigated a multicomponent 2-D AM-FM model for the spectrotemporal patches and developed a novel scheme for optimal choice of the model order. We showed that the model order for a given patch is proportional to the variations in the instantaneous fundamental frequency of the speaker. The model order was varied across spectrogram patches depending on the speaker's pitch. We estimated the model parameters in two steps: (1) the AM and FM components were estimated by solving the 2-D demodulation problem, and (2) the weights of the 2-D cosine carriers were estimated by using the least-squares method. For solving the demodulation problem, we used the Riesz transform technique, which gives accurate estimates of the AM and FM components. The proposed model was applied for the decomposition of a narrowband spectrogram and compared with its monocomponent counterpart. We used four types of objective measures to quantify

the model accuracy: GSNR, SSNR, patch reconstruction error, and PESQ. First, we evaluated the model accuracy for all voiced speech utterances and compared multicomponent model with a monocomponent model. Second, we hypothesized that, in addition to voiced speech sounds, the unvoiced sounds are also better represented by incorporating more number of cosine carriers in the model. This argument was supported by running an experiment on a speech database containing continuous speech utterances from a variety of speakers. We showed that the multicomponent model gives superior model accuracy over its monocomponent counterpart in terms of all the four measures.

Chapter 3

Periodic and Aperiodic Decomposition of Speech Signals

The speech signal broadly comprises voiced and unvoiced segments. The voiced segments can be considered the output of a time-varying vocal-tract filter with the excitation being a combination of a quasiperiodic stream of pulses and noise. In the case of unvoiced segments, the excitation is noise-like. Therefore, the speech signal is a mixture of a periodic (deterministic) and an aperiodic (stochastic) component. Decomposing a speech signal into its periodic and aperiodic constituents is an important task and finds applications in speech synthesis [86], denoising [87], voice analysis [88], etc. For instance, such a decomposition is useful for controlling the characteristics of excitation source signal for speech synthesis application [89] where the aperiodic component characterizes the voice attributes such as *breathiness* and *roughness*. Breathiness is caused due to glottal air leakage or turbulence during phonation. Roughness is defined by the presence of low-frequency noise component [42]. Unvoiced speech sounds exhibit a higher degree of the aperiodic component. In contrast, although the voiced speech sounds have relatively lesser energy of the aperiodic component, they are never completely devoid of it [90].

Previous studies [91] have shown that there are mainly two sources of aperiodicity in voiced speech sounds: 1) additive random noise, and 2) modulation aperiodicity.
• *Additive random noise*: This source of aperiodicity represents frication or aspiration noise. It is present in segments of voiced fricatives or breathy vowels [92–94]. This is modeled as additive because the noise is superimposed onto the voice source. The

location of the source of additive noise in the vocal apparatus indicates the nature of the noise. For example, the noise is aspiration noise if it is generated at the glottis (especially when the glottal closure is incomplete), or frication noise if it is generated at a constriction in the vocal tract (e.g. voiced fricatives). The frication noise and aspiration noise also differ in their spectral properties — the frication noise is highpass, and the aspiration noise is broadband.

• *Modulation aperiodicity*: Random perturbations in the duration of glottal cycles and their peak amplitudes cause modulation aperiodicity in the speech signal. Aperiodicity may also be introduced by voluntary changes in the source characteristics as in prosody, and in formant transitions.

We study the characteristics of the demodulated 2-D carrier for different types of speech sounds. In particular, we observe that the 2-D carrier exhibits mainly two types of spectrotemporal regions: (1) coherent regions, which contain a strong structure; and (2) incoherent regions where there is no prominent structure. These spectrotemporal properties are characterized by maps computed from the 2-D carrier using the complex Riesz transform: (1) the coherencegram, and (2) the orientationgram. We employ these t-f maps for periodic-aperiodic decomposition (PAPD).

In Section 3.1, we first elaborate on the spectrotemporal properties of the 2-D carrier in connection to the periodic/aperiodic components of speech. We show that different speech sounds exhibit distinguishable spectrotemporal patterns in the 2-D carrier. Such patterns can be effectively captured by computing the coherence and orientation of the 2-D carrier. In Section 3.2, we propose a new t-f map referred to as the *tracegram*. Unlike coherencegram and the orientationgram, which are computed from the 2-D carrier, the tracegram is computed directly from the speech spectrogram. Analogous to short-time energy in 1-D, the tracegram gives spectrotemporal distribution of energy. In Section 3.3, we construct a joint feature vector in $\mathbb{R}^3$ by taking values from the coherencegram, orientation, and the

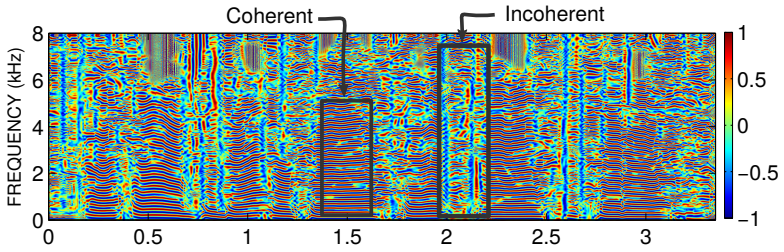


Figure 3.1: Illustration of the coherent and the incoherent time-frequency regions in a carrier spectrogram.

tracegram at a t-f bin. The problem of periodic/aperiodic decomposition is viewed as an unsupervised binary classification problem where each t-f bin is classified as either periodic or aperiodic. The classifier output results in a binary mask in the t-f domain, which could be used to decompose the speech signal into its periodic and aperiodic components.

3.1 The Carrier Spectrogram and its Time-Frequency Properties

We have also seen that the carrier in a 2-D AM-FM multicomponent model can be modeled as consisting of a 2-D sinusoid oscillating at the fundamental frequency and the sinusoids oscillating at its harmonics. We focus on the 2-D carrier component that oscillates at the fundamental frequency and refer its t-f map to as the *carrier spectrogram*.

We have seen in Chapter 2 that the Riesz transform approach enables demodulation of the spectrogram into AM and FM components. The FM component or the *carrier spectrogram* carries information not only about the evolution of the fundamental frequency of the speaker but also the perturbations in pitch partials due to the aperiodicity in the speech signal. An example carrier spectrogram corresponding to a continuous speech utterance is shown in Figure 3.1. Some observations are in order: (1) the interference due to the formants of the vocal-tract filter has been removed by the demodulation technique; (2) the temporal evolution of the fundamental

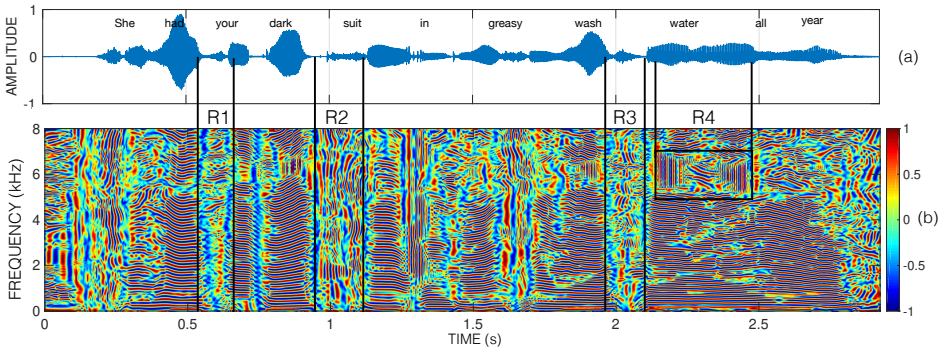


Figure 3.2: [Color online] (a) A speech waveform, and (b) its carrier spectrogram. The speech utterance is "*She had your dark suit in greasy wash water all year*," spoken by a female speaker. The labels R_1, R_2, R_3 and R_4 indicate the correspondence of the specific sounds to the spectrotemporal signature in the carrier spectrogram.

frequency and its harmonics for voiced sounds manifests prominently; and (3) the t-f signatures for voiced and unvoiced sounds are distinct. Certain regions are more *coherent* in the sense that they are structured and have a specific orientation, whereas others are *incoherent* because they lack directionality and structure. The coherent regions are predominantly voiced whereas the incoherent ones are unvoiced. The 2-D carrier depends on the type of spectrogram used for demodulation. For instance, a narrowband spectrogram has a finer frequency resolution and the carrier has horizontal striations in voiced regions. On the other hand, a wideband spectrogram would have exactly the opposite effect — the carrier would contain vertical striations in the voiced regions. In this study, we focus on the narrowband spectrogram. An example of a carrier spectrogram and the corresponding speech waveform is shown in Figure 3.2. We observe that voiced sounds are characterized by a carrier that is predominantly oriented parallel to the time axis, whereas fricatives do not have a preferred orientation (R1, R2, R3). Certain voiced regions may have spectrotemporal regions (such as R4) that are coherent but their orientation is different from those in the other regions. Carrier orientation alone does not suffice and one must take coherence also into account. In Sections 3.1.1 and 3.1.2, we describe how to compute the coherence map and orientation map, respectively.

3.1.1 *The Coherencegram*

We compute the coherence using the structure tensor approach discussed in Section 2.4.4 of Chapter 2, but with a difference. The structure tensor now operates on the carrier spectrogram and not the narrowband speech spectrogram. A 2×2 structure tensor is determined at every t-f location in the carrier spectrogram. The relative discrepancy between the eigenvalues of the structure tensor reflects the degree of uniformity of the underlying 2-D pattern and is quantified by the following *coherence measure*:

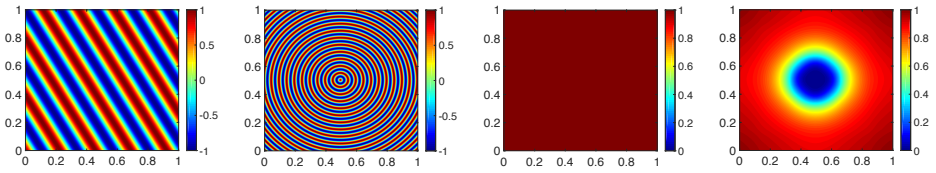
$$C(\boldsymbol{\omega}_0) \triangleq \begin{cases} \left(\frac{\lambda_1-\lambda_2}{\lambda_1+\lambda_2}\right)^2, \ \lambda_1 \neq 0, \lambda_2 \neq 0, \\ 0, \qquad \text{otherwise}, \end{cases} \tag{3.1}$$

where λ_1 and λ_2 are the eigenvalues of the structure tensor. By definition, coherence takes a value between zero and unity.

The eigenvalues and eigenvectors of the structure tensor have found applications in edge detection [95, 96], image segmentation [97, 98] and estimation of the local orientation [80, 99].

If the eigenvalues are small, it indicates that the energy is low and that locally, there is no preferred orientation. The coherence takes a small value in this case. Spectrotemporal regions representing unvoiced sounds have this property. If both eigenvalues are large and comparable, it corresponds to a high-energy region. In image processing, this would correspond to a corner. In spectrotemporal representations of speech, such regions correspond to voiced/unvoiced transitions. If the maximum eigenvalue significantly dominates the minimum eigenvalue, then it indicates a clear directional preference and high coherence (closer to unity) — this kind of a structure is possessed by voiced regions.

To illustrate further, consider a planar cosine (Figure 3.3(a)) and a radial cosine (Figure 3.3(b)). A planar cosine has a directional preference and is highly coherent everywhere (Figure 3.3(c)). A radial cosine has directional preference only away from the center (Figure 3.3(d)).



(a) (b) (c) (d)

Figure 3.3: (a) A planar cosine, (b) a radial cosine, (c) coherence of the planar cosine, and (d) coherence of the radial cosine. The images are of size 900×900 pixels. The smoothing window $\psi(\boldsymbol{\omega})$ used in the computation of the structure tensor is a 2-D Gaussian of size 90×90 pixels. The coherence is 1 for a planar cosine. For a radial cosine, it is closer to one away from the center. This is because, away from the center, the ripples of a radial cosine are approximately planar.

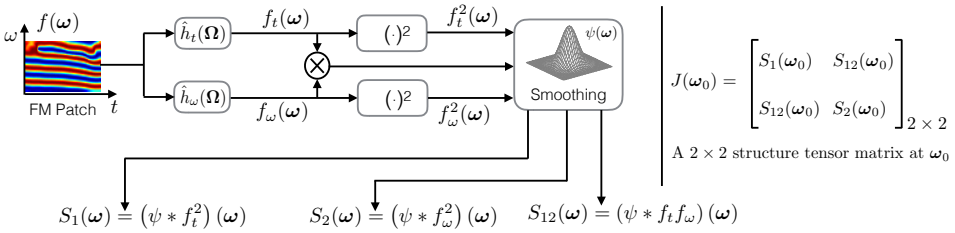


Figure 3.4: The computation of a 2×2 structure tensor matrix at location $\boldsymbol{\omega}_0$ for a given FM patch $f(\boldsymbol{\omega})$. The dimensions of $S_1(\boldsymbol{\omega})$, $S_2(\boldsymbol{\omega})$ and $S_{12}(\boldsymbol{\omega})$ are same and equal to the dimensions of the FM patch. $\hat{h}_t(\boldsymbol{\Omega})$ and $\hat{h}_\omega(\boldsymbol{\Omega})$ represent the complex Riesz kernels along t-axis and ω-axis, respectively. An example of 2×2 structure tensor matrix $J(\boldsymbol{\omega}_0)$ is shown on the right.

In the case of speech, we perform the computations patch-wise, which is more suitable for parallel processing. The patches are of size 600 Hz $\times$ 100 ms with an overlap of 75% along time and frequency axes. Figure 3.4 illustrates the computation of the structure tensor. Finally, overlap-add synthesis of the spectrotemporal coherence patches gives rise to the *coherencegram* that has the same dimensions as the carrier spectrogram. Figure 3.5(b) displays the coherencegram along with the corresponding carrier spectrogram.

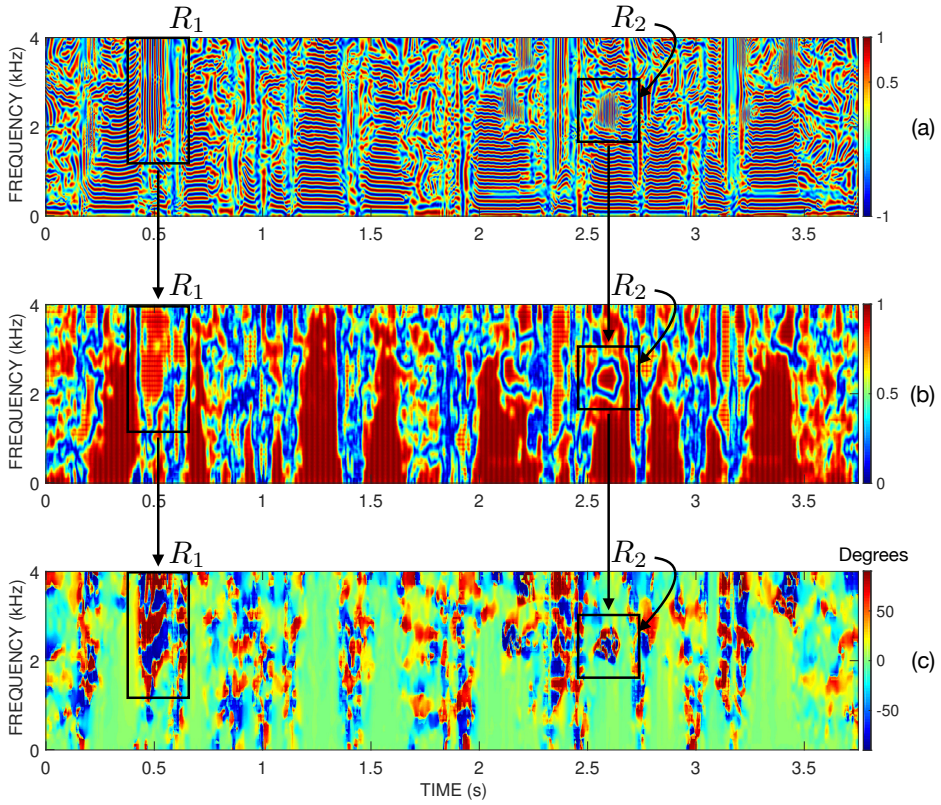
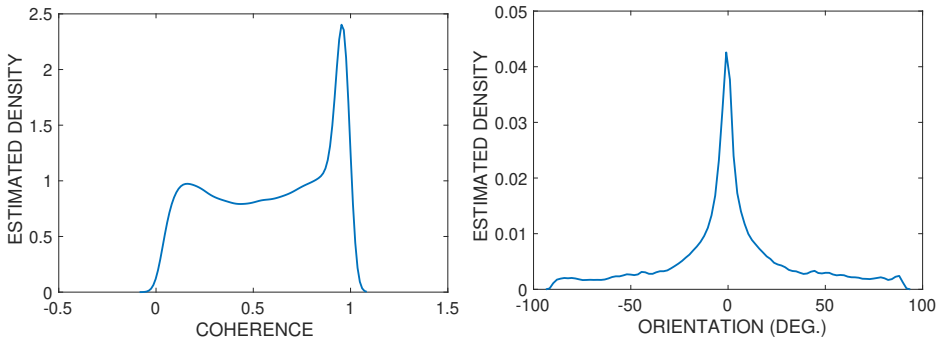


Figure 3.5: Illustration of (a) the carrier spectrogram, (b) coherencegram, and (c) the orientationgram. The t-f regions enclosed by the boxes highlight complementary information captured by coherence and orientation.

3.1.2 *The Orientationgram*

The orientationgram is a that contains the estimate of the local orientation in the carrier spectrogram at each t-f point. We use the optimization formulation given in Equation (2.37) (Section 2.4.4 of Chapter 2) to determine the preferred orientation at a t-f point. The orientation is computed with respect to the time axis. Similar to the coherencegram, the orientationgram is also computed patch-wise followed by overlap-add synthesis. Figure 3.5(c) displays an orientationgram. The orientation is close to 0° for the t-f regions corresponding to voiced sounds – this property is exhibited by pitch partials in the carrier spectrogram and correspond to voiced



(a) (b)

Figure 3.6: (a) Estimated density of the coherencegram values corresponding to a female speech utterance. The coherence is strictly between 0 and 1. The spill-over beyond this interval is due to smoothing caused by the kernel-density estimator, and (b) estimated density of the orientation values corresponding to a female speech utterance, "*Not at this particular case, Tom apologized Whittemore.*"

sounds. On the other hand, unvoiced sounds have an orientation that significantly deviates from 0°.

Figure 3.6(a) shows the kernel density estimate of coherence values obtained from continuous speech utterance spoken by a female speaker. The strong mode closer to unity indicates that the percentage of voiced sounds is higher. The mode around 0 is weaker and spread out, which corresponds to unvoiced regions. A similar behavior was found for a male speaker. Figure 3.6(b) illustrates the kernel density estimate of orientation values for a continuous speech utterance spoken by a female speaker. The presence of a strong mode around 0° is indicative of largely flat pitch. The rising and falling pitch correspond to 90° and −90°, respectively.

3.2 The Tracegram

In addition to the coherencegram and orientationgram computed from the carrier spectrogram, one could also determine the local energy computed as the trace of the structure tensor. However, in this case, the structure tensor is computed

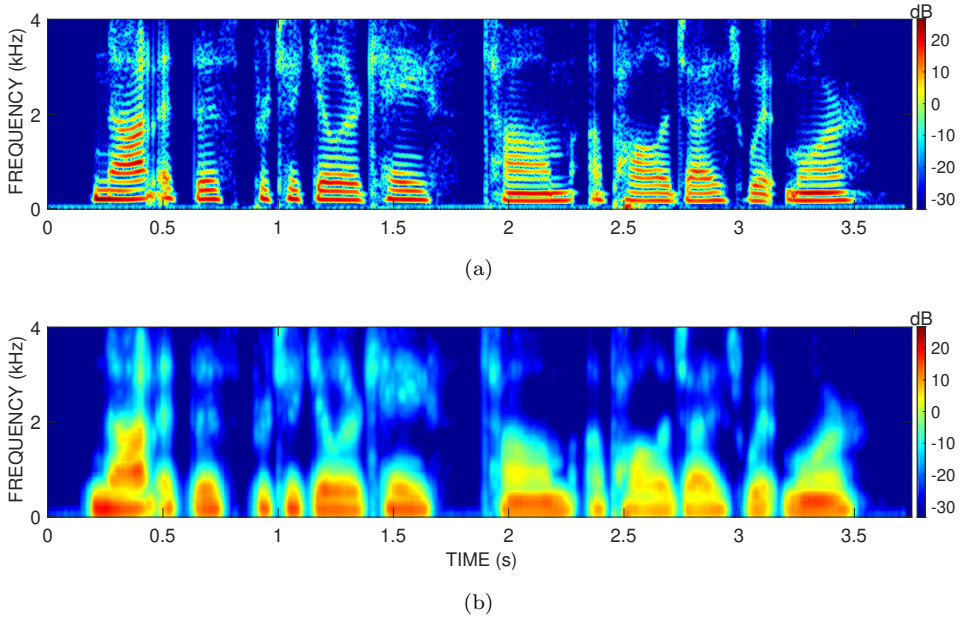


Figure 3.7: (a) Narrowband spectrogram; and (b) its corresponding tracegram for a speech utterance "*Not at this particular case, Tom apologized Whittemore" spoken by a female speaker.*

from the speech spectrogram and not the carrier spectrogram as the latter does not have any amplitude information. Periodic regions have high energy and hence higher trace of the structure tensor, whereas aperiodic regions have relatively lower trace values. The t-f map of the trace of the structure tensor is referred to as the tracegram. Computation of the tracegram also proceeds in a patch-by-patch fashion with overlap-add synthesis. Figure 3.7 displays a tracegram and the corresponding spectrogram. Observe that the values are high in harmonic regions and relatively low in inharmonic regions.

3.2.1 *Juxtaposing the Coherencegram and Orientationgram*

Figure 3.5 shows the coherencegram and the orientationgram capture complementary information for various speech sounds. Consider the t-f region labeled $R-2$ around (2.6 s, 2 kHz) of the carrier spectrogram. The corresponding temporal segment is

overall voiced since the low-frequency structure shows strong voicing. However, R_2 is an *island of aperiodicity* within a voiced segment. Interestingly, the boundary of the island is delineated by a low coherence. However, the island of aperiodicity is not distinguished by coherence, but is reflected in the orientationgram since the striations are vertical as opposed to the pitch harmonics, which are nearly horizontal. A similar phenomenon can be observed in the region R_1, although this is not a voiced region. These observations indicate that both coherence and orientation must be used to determine whether a region is periodic or not.

3.3 Application to Periodic and Aperiodic Decomposition (PAPD) of the Speech Signal

In this section, we describe how the coherence, orientation, and trace can be used to perform PAPD of speech. The problem of PAPD is viewed as a binary classification problem, wherein each t-f bin must be classified as either periodic or aperiodic. We solve this problem in an unsupervised manner.

3.3.1 *Data Standardization*

Let $C(\boldsymbol{\omega})$, $O(\boldsymbol{\omega})$ and $T(\boldsymbol{\omega})$ denote the coherencegram, orientationgram, and the tracegram, respectively. We use the absolute of the orientationgram, denoted by $\tilde{O}(\boldsymbol{\omega}) = |O(\boldsymbol{\omega})|$. The tracegram is mapped to a logarithmic scale: $\tilde{T}(\boldsymbol{\omega}) = 10 \log T(\boldsymbol{\omega})$. Let the dimensions of the spectrogram be $M \times N$. Also, let $\boldsymbol{c} = [c_1, c_2, \ldots, c_d]^{\mathrm{T}}$, $\boldsymbol{o} = [o_1, o_2, \ldots, o_d]^{\mathrm{T}}$ and $\boldsymbol{t} = [t_1, t_2, \ldots, t_d]^{\mathrm{T}}$ denote the vectorized form of $C(\boldsymbol{\omega})$, $\tilde{O}(\boldsymbol{\omega})$ and $\tilde{T}(\boldsymbol{\omega})$, respectively, and $d = MN$. The feature matrix is

constructed as

$$\mathbf{X} = \begin{bmatrix} c_1 & o_1 & t_1 \\ c_2 & o_2 & t_2 \\ \vdots & \vdots & \vdots \\ c_d & o_d & t_d \end{bmatrix}_{d \times 3} \tag{3.2}$$

where each row is a feature vector. We perform feature normalization of $\mathbf{X}$ as follows

$$\tilde{\mathbf{X}} = (\mathbf{X} - \underline{\mathbf{1}}\boldsymbol{\mu}^{\mathrm{T}}) \oslash \underline{\mathbf{1}}\boldsymbol{\sigma}^{\mathrm{T}} \tag{3.3}$$

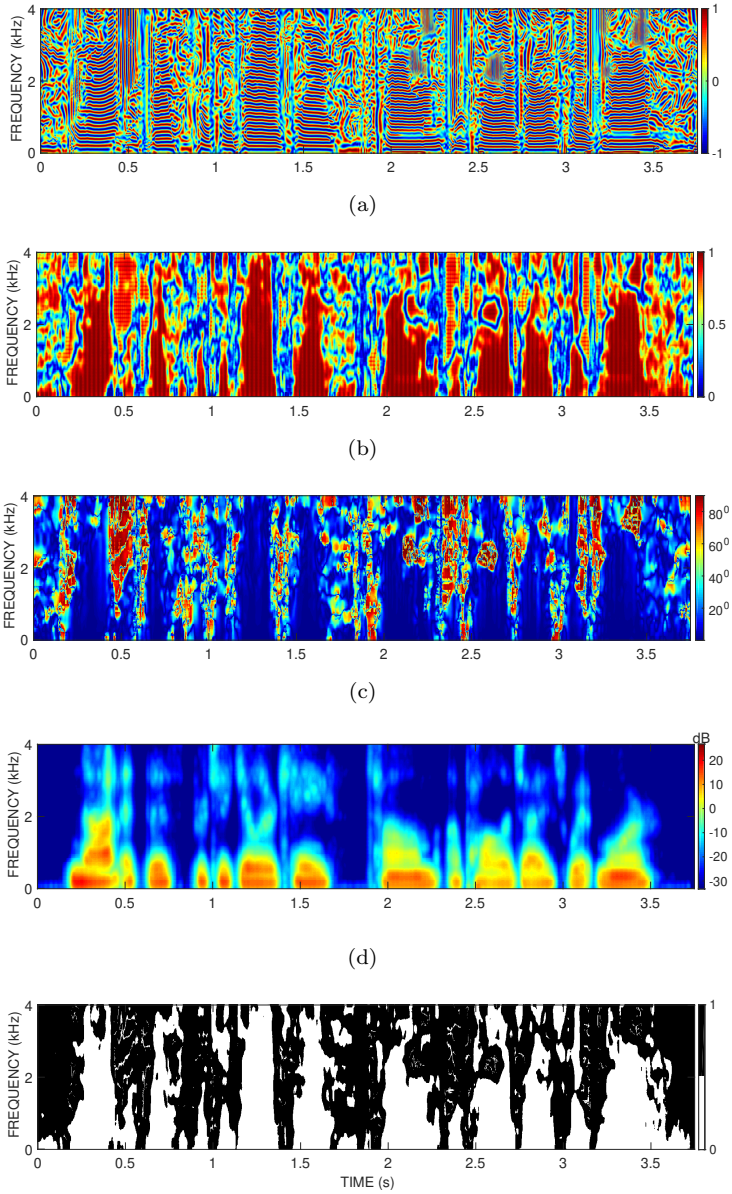
where $\underline{\mathbf{1}}$ is a $d \times 1$ vector of all 1s, $\oslash$ denotes element-wise division, $\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3]^{\mathrm{T}}$ and $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \sigma_3]^{\mathrm{T}}$ with

$$\mu_j = \frac{1}{d}\sum_{i=1}^{d} X_{ij}, \quad \text{and} \quad \sigma_j = \frac{1}{d}\sum_{i=1}^{d}(X_{ij} - \mu_j)^2.$$

For a signal of duration 3 s, $M \times N = 512 \times 3000$, and the number of feature vectors is $15,36,000$.

3.3.2 *Unsupervised Binary Mask Estimation for PAPD*

We employed the K-means algorithm [100] to identify two clusters, one corresponding to periodic and the other corresponding to aperiodic. Twenty iterations of the K-means algorithm were found to suffice. The K-means algorithm is unsupervised and provides two clusters. Which cluster corresponds to periodic regions and which one to aperiodic must be determined. Toward this, we rely on high-frequency regions which can be considered predominantly aperiodic for speech sounds. Consequently, one would expect a majority of the high-frequency t-f bins to be aperiodic. The cutoff is empirically selected as $f_s/4$. The majority cluster corresponding to the t-f bins above $f_s/4$ is labelled as aperiodic and the minority as periodic. To explain further, if the two clusters are C_1 and C_2 and if C_1 has a majority of points above $f_s/4$, then the cluster C_1 is labelled as aperiodic and C_2 as periodic, and vice versa. Figure 3.8(e) displays the predicted binary mask along with coherencegram and



(a)

(b)

(c)

(d)

(e)

Figure 3.8: (a) The carrier spectrogram, (b) coherencegram, (c) absolute orientationgram, (d) tracegram, and the predicted (e) binary mask by the K-means algorithm for the speech utterance, "*Not at this particular case, Tom apologized Whittemore,*" *spoken by a female speaker.*

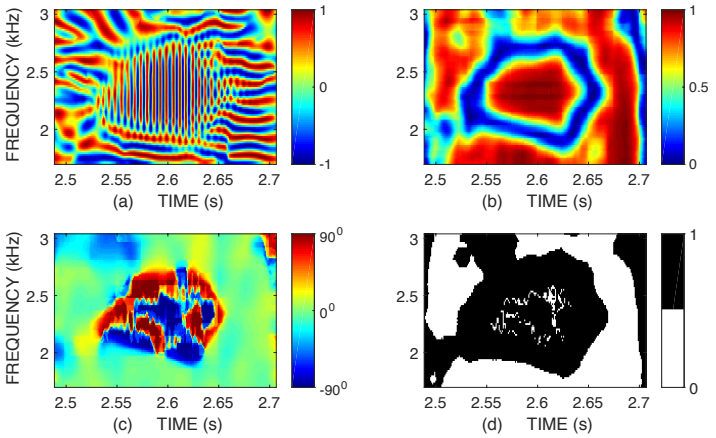


Figure 3.9: (a) The carrier spectrogram patch, (b) coherencegram patch, (c) orientationgram patch, and the predicted (d) binary mask obtained by K-means algorithm for the speech utterance, "*Not at this particular case, Tom apologized Whittemore" spoken by a female speaker.*

the orientationgram. The black regions in the binary mask are aperiodic and the white regions are periodic. We observe that the binary mask preserves the fuzzy boundaries indicating soft decisions for periodicity and aperiodicity in the t-f domain. A zoomed-in portion is shown in Figure 3.9. The corresponding coherence shown in Figure 3.9(b) is largely high indicating a periodic region, but the orientation map shown in Figure 3.9(c) suggests otherwise and indicates an *island of aperiodicity*. The final decision taking into account both orientation and coherence shown in Figure 3.9(d) is more accurate.

3.3.3 *PAPD of the Speech Signal*

The estimated binary decisions could be used to decompose a speech signal into periodic/aperiodic components. The decomposition depends on several parameters such as choice of window, window length, and DFT length. We considered a sampling rate of 8 kHz, DFT size of 512 points, Hamming window of duration 25 ms and a frame-shift of 1 ms. Once the binary decisions are obtained, the original STFT is multiplied point-wise with the estimated binary mask, which retains only the

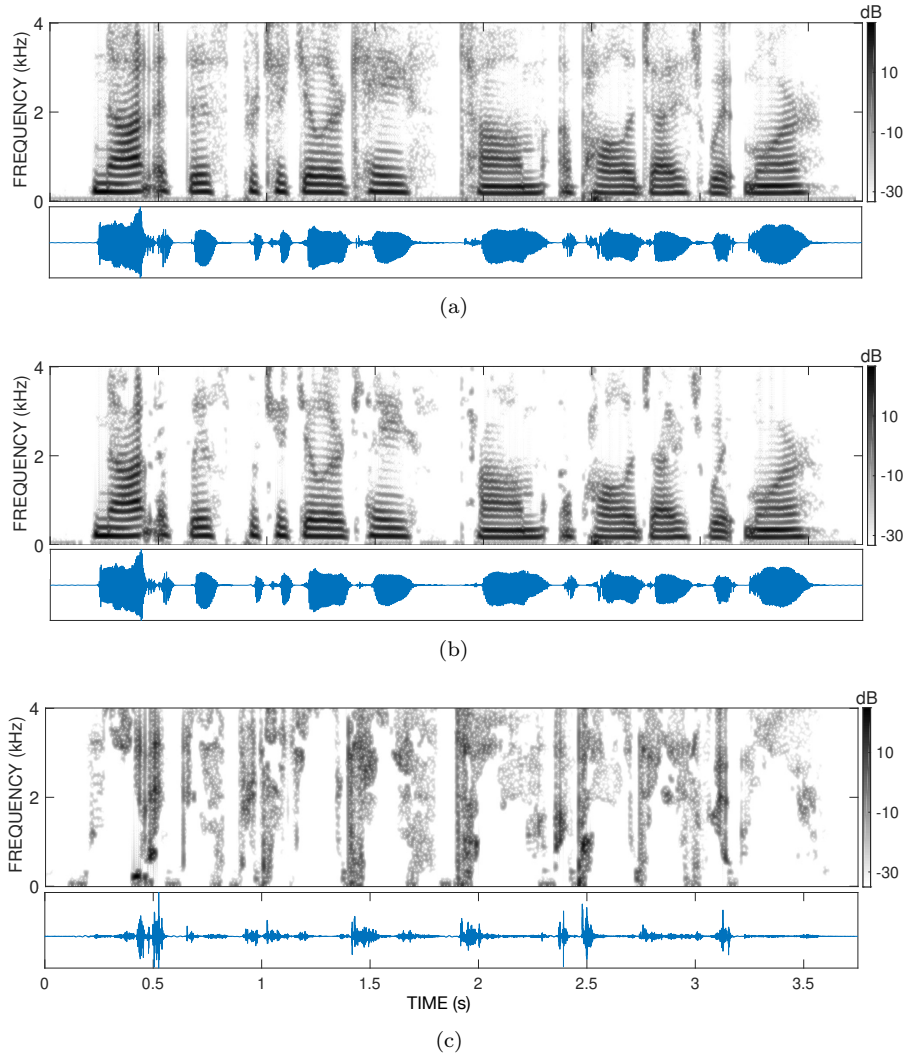


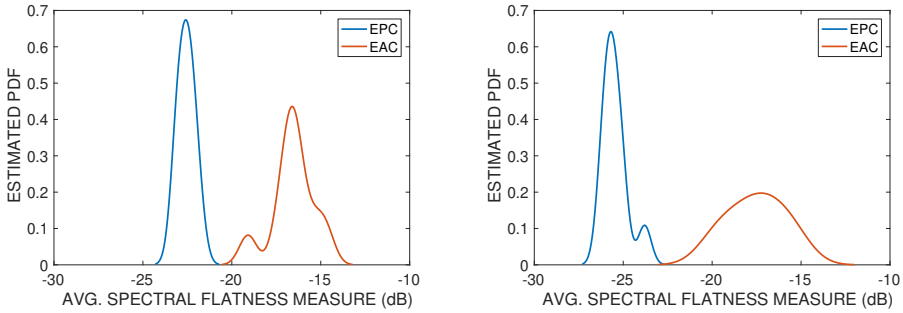
Figure 3.10: (a) Original spectrogram and the speech waveform; (b) spectrogram and the waveform of the periodic component; and (c) spectrogram and the waveform of the aperiodic component for the speech utterance, "*Not at this particular case, Tom apologized Whittemore," spoken by a female speaker.*

aperiodic t-f bins. The masked STFT is then used to reconstruct the signal using standard overlap-add procedure. This gives the aperiodic signal. A similar procedure with the complementary mask gives the periodic component. Figure 3.10 shows the spectrograms and the corresponding aperiodic and periodic waveforms. The

absolute error between the original waveform and the signal obtained by adding the estimated periodic and aperiodic components was found to be of the order of 10^{-16}. As expected, the spectrogram of the aperiodic component does not have any pitch structure/harmonics (Figure 3.10(c)). In contrast, the harmonic structure is preserved in the spectrogram of the periodic component (Figure 3.10(b)). We also compute the spectral flatness measure (SFM) [101] for the estimated periodic component (EPC) and the estimated aperiodic component (EAC). Spectral flatness measure also known as the tonality coefficient or Wiener entropy is a measure of how tone-like a signal is as opposed to being noise-like. The SFM is the ratio of the geometric mean to the arithmetic mean of the power spectrum:

$$\text{SFM} = \frac{\sqrt[K]{\prod_{k=1}^{K} |X[k]|^2}}{\frac{1}{K}\sum_{k=1}^{K} |X[k]|^2}, \tag{3.4}$$

where $|X[k]^2|$ is the power spectrum of the signal, k denotes the DFT index and K denotes the DFT length. By definition, SFM lies in the range from 0 to 1, being high for noise-like signals and low for tone-like signals. The SFM is expected to be higher for the estimated aperiodic component than the periodic component. We show that this is indeed the case. We choose one male (bdl) and one female (slt) speaker from the CMU-ARCTIC database and randomly select 50 speech utterances for each speaker. For every speech signal, the average of the frame-wise SFMs is computed for the aperiodic component as well as the periodic component. A kernel density estimate of the average SFMs (in dB) for EAC and EPC is shown in Figure. 3.11. This figure clearly shows that the average SFM is lower for the periodic component than for the aperiodic component irrespective of the speaker's gender. Table 3.1 reports the mean of the average SFMs for both the speakers. The mean SFM is lower for the female speaker than the male speaker, which is because of the relatively higher pitch for female speaker. Fewer harmonics for a female speaker than a male



(a) (b)

Figure 3.11: The kernel density estimates of average SFMs over 50 speech utterances spoken by (a) male speaker, and (b) a female speaker. From the figure, we observe that the average SFM is always lower for the estimated periodic component (EPC) than the estimated aperiodic component (EAC) irrespective of the gender.

Table 3.1: The mean of the average SFMs (dB) for estimated periodic and aperiodic components (EPC and EAC, respectively) across 50 speech utterances.

	Male (bdl)	Female (slt)
EPC	-22.57 ± 0.43	-25.46 ± 0.70
EAC	-16.56 ± 1.2	-17.54 ± 1.6

speaker result in a lower SFM value for a female speaker. The spread of the SFMs of aperiodic component for the female speaker is also broader than for the male speaker.

3.4 Chapter Summary

We examined the spectrotemporal properties of the carrier spectrogram for two broad classes of speech sounds. We showed that the carrier spectrogram exhibits distinguishable spectrotemporal patterns for the sound classes. The carrier spectrogram was used to compute two novel spectrotemporal maps: (1) the coherencegram; (2) the orientationgram; and the spectrogram is used to compute the tracegram,

which is an energy representation. The coherencegram and the orientationgram, and the energy information contained in the tracegram was used to arrive at a feature representation for periodic and aperiodic signals. The new feature representation results in a spectrotemporal mask, which is used to decompose a speech signal into periodic and aperiodic components. We provided supporting illustrations on both male and female speaker samples.

Chapter 4

Voiced/Unvoiced Segmentation and Quantification of Speech Aperiodicity

In Chapter 3, we analyzed the problem of splitting a speech signal into its periodic and aperiodic components, by analyzing the t-f regions using both coherence and orientation. For speech synthesis applications, it is important to quantify the extent to which a signal is periodic or aperiodic. The widely used speech vocoders STRAIGHT [44] and WORLD [45] use band-wise aperiodicity parameters. These vocoders model the stochastic part of voiced speech by coloring the spectrum of white noise using a transfer function derived from the aperiodicity parameters. Could one develop a numerical measure to specify the degree of aperiodicity in a signal based on the spectrotemporal analysis considered here. For instance, the measure could be 0 for a pure tone and 1 for white noise. Previous studies [88] have shown that the speech reconstructed without modeling the aperiodicity explicitly sounds unnatural. A good perceptual fusion of the periodic and aperiodic components is important even for voiced speech segments so that the hoarseness and breathiness can be preserved, which renders the synthesized speech natural and of high quality.

In this chapter, we develop methods to not only identify the voiced and unvoiced speech segments, but also quantify degree of aperiodicity of the voiced speech segments. Reliable voiced/unvoiced (V/UV) segmentation of a speech signal is essential for high-quality speech reconstruction [44,45,102]. V/UV segmentation has also been used for automatic speech recognition [103]. We develop a new V/UV segmentation technique based on novel features derived from the coherencegram.

We begin by describing the terminology behind V/UV segmentation in Section 4.1. Next, we discuss the principle behind modeling the aperiodicity of sinusoidal signals in Section 4.2 and extend it to model the aperiodicity of voiced speech segments (Section 4.3) and further to derive the band-aperiodicity parameters (Section 4.3.2). In Section 4.4, we show the effectiveness of the proposed band-aperiodicity parameters and the V/UV decisions for the task of speech reconstruction based on the spectral synthesis model within the WORLD vocoder framework.

4.1 Voiced/Unvoiced Speech Segmentation

Effectively, the objective is to determine whether the speech sound production involves the vibrations of the vocal folds or not. The vocal fold vibrations produce a quasi-periodic excitation to the vocal tract for voiced speech whereas pure transient and/or turbulent noises have a voiceless excitation. If neither of these is present, then we have a *speech pause/silence*. On the other hand, if both sources of excitation are present simultaneously, i.e., we have *mixed excitation*, and the speech is considered to be voiced because the vocal fold vibration is part of sound production. Before proceeding further, we review the prior art in this area.

4.1.1 *Prior Art*

Effectively, the V/UV decision task is a binary segmentation problem and machine learning approaches for training a classifier based on a set of acoustic features have been developed. The training could be either supervised or unsupervised and the acoustic features must be sufficiently discriminative. Guidelines for selecting an appropriate set of acoustic features have been provided early on [103–106]. Typically, the following features have been used.

(1) **Log energy** of signal $s[n]$ of length N given by

$$E_s = 10 \times \log_{10}\left(\epsilon + \frac{1}{N}\sum_{n=0}^{N} s^2[n]\right) \tag{4.1}$$

where ϵ is a small value to prevent singularity at the origin. This feature discriminates speech versus silence.

(2) **Zero-crossing rate**, which is the number of zero-crossings per frame. This feature is a good measure of the dominant frequency.

(3) **Normalized autocorrelation coefficient** computed from the signal and its right-shifted version defined as

$$C_1 = \frac{\sum_{n=1}^{N} s[n]s[n-1]}{\sqrt{\sum_{n=1}^{N} s^2[n] \cdot \sum_{n=0}^{N-1} s^2[n]}}. \tag{4.2}$$

Note that the range of summation is different in the denominator terms.

(4) **Energy ratio** between a high-frequency band to a low-frequency band. This is motivated by the observation that voiced speech sounds have energy concentrated below 1 kHz and unvoiced sounds above 2 kHz.

(5) **Energy in the linear prediction (LP) residue** given by

$$E_p = E_s - 10 \times \log_{10}\left(10^{-6} + \left|\sum_{k=1}^{p} \alpha_k \phi(0,k) + \phi(0,0)\right|\right) \tag{4.3}$$

where E_s is the log-energy of the signal (4.1) and

$$\phi(i,k) = \frac{1}{N}\sum_{n=1}^{N} s[n-i]s[n-k] \tag{4.4}$$

is the (i,k) term of the covariance matrix, α_ks are the LP coefficients, and p denotes the order of the predictor (typically $p = 12$) [107].

(6) **First-order LP coefficient** estimated using the covariance method [107].

One could design the classifier by assuming certain statistics on the features or without. Atal and Rabiner assumed a multidimensional Gaussian distribution for the acoustic features derived from the training data and used a Bayesian decision rule for

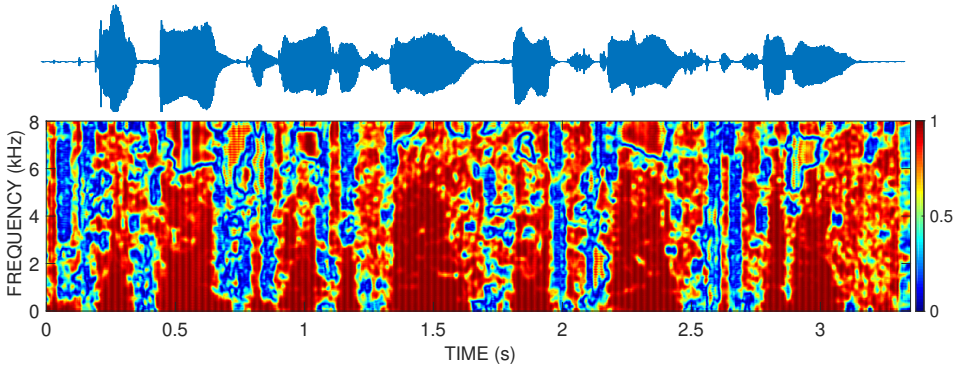


Figure 4.1: Coherencegram and the corresponding speech waveform

classification. However, they considered a three-class classification problem: voiced, unvoiced, and silence, which was relevant for telephony applications. Siegel [105] employed a linear discriminant function, whereas Siegel and Bessey [108] considered a three-class problem with the mixed excitation category comprising the third class. Rabiner and Samuer [109] used a spectral distance measure, more specifically, a combination of the LP spectrum and log energy to compute the spectral proximity with the pre-stored class templates obtained using training data, and obtained significant improvements. Recent trends for V/UV segmentation of speech include approaches based on machine learning and unsupervised learning [110–117].

The proposed approach relies on the coherencegram, which is computed from the carrier spectrogram. Unlike a waveform-based approach, the proposed method has a fair degree of spectrotemporal smoothing inbuilt, which makes it robust to variations of the order of a glottal cycle.

4.1.2 *Coherence Features for Voiced/Unvoiced Segmentation*

Figure 4.1 displays the coherencegram along with the corresponding speech utterance. The coherence changes from being high to low or vice versa at the voiced/unvoiced transitions boundaries in the speech waveform. However, the coherence could also go high for silent segments. This ambiguity can be overcome by considering short-time

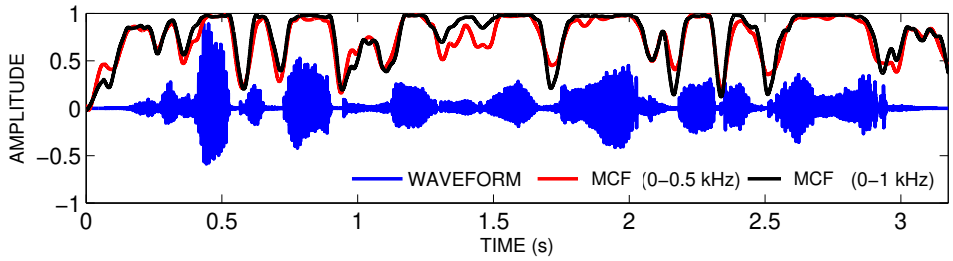


Figure 4.2: The variations of mean coherence features (MCFs) corresponding to the two frequency subbands: 0-0.5 kHz and 0-1 kHz. We observe that the MCFs are relatively high for voiced segments and relatively low for unvoiced segments.

energy. The short-time energy of a windowed speech signal at time instant t_i is given by

$$E(t_i) = \int_{-\infty}^{\infty} s(t)w(t - t_i)\,\mathrm{d}t, \tag{4.5}$$

and its normalized version is given by

$$E_n(t_i) = \frac{E(t_i)}{\max\limits_i E(t_i)}. \tag{4.6}$$

A 2-D coherencegram must be used to arrive at the decision boundaries.

The mean of the coherence $C(t, f)$ over a subband $(0, f_h)$ (in Hz) is given by

$$M_h(t) = \frac{1}{f_h}\int_0^{f_h} C(t, f)\,\mathrm{d}f. \tag{4.7}$$

We further smooth $M_h(t)$ using a 21-point, fourth-order Savitzky-Golay filter. The resulting feature is referred to as the *mean coherence feature (MCF)*. Figure 4.2 displays the MCFs corresponding to a speech waveform computed from the subbands: 0-0.5 kHz and 0-1 kHz. The MCF is high for voiced segments, and low for unvoiced segments. The question next is the choice of the subband. Since the low-frequency regions have higher energy, choosing a low-frequency subband would result in greater robustness.

We examine the accuracy of the V/UV classifiers separately trained on the MCFs computed from a set of subbands. For a speech signal sampled at 8 kHz, we compute

MCFs for seven subbands (in kHz): 0-0.5, 0-1, 0-1.5, 0-2.5, 0-3, 0-3.5 and 0-4. Combining $E_n(t_i)$ with MCF $M_h(t_i)$ gives a 2-D feature vector $[M_h(t_i)\ E_n(t_i))]^T$ for the i^{th} speech frame. We train a binary classifier that uses logistic regression with a cross-entropy loss function [100]. The parameters of the classifier are optimized by employing the gradient-descent algorithm without any regularization. We use CMU-ARCTIC database that has parallel recordings of electroglottograph (EGG) signals from which the ground truth V/UV time-stamps are marked by employing the algorithm described in [118]. The training data consists of a total of 140 speech utterances from the two speakers: 70 male (bdl) and 70 female (slt). The speech waveforms in the database are downsampled to 8 kHz. A total of seven classifiers are trained corresponding to each of the subbands mentioned above. The classifier performance is objectively evaluated on the test data, which consists of 100 speech utterances corresponding to each of the male and female voice from the database. The accuracy of the binary classifier is evaluated using the **Average Detection Rate**, which is the proportion of V/UV regions correctly classified. It is given by

$$\text{AVG. DR } (\%) = \frac{1}{2}\left(\frac{N_{V\to V}}{N_V} + \frac{N_{UV\to UV}}{N_{UV}}\right) \times 100, \tag{4.8}$$

where

N_V:	total number of voiced samples
N_{UV} :	total number of unvoiced samples
$\frac{N_{V\to V}}{N_V}$:	fraction of voiced samples classified as voiced, and
$\frac{N_{UV\to UV}}{N_{UV}}$:	fraction of unvoiced samples classified as unvoiced.

Table 4.1 shows the performance comparison of the classifiers for different frequency subbands. We observe slight degradation in the average accuracy of the classifier over wider subbands. Based on these results, we rely on the first two subbands (0-0.5 kHz and 0-1 kHz) for computing the MCFs. Let $M_l(t_i)$ and $M_h(t_i)$ denote the MCFs corresponding to the subbands 0-0.5 kHz and 0-1 kHz for i^{th} speech frame respectively. We combine the MCFs of these bands with normalized

Table 4.1: Average detection rate (in %) for various frequency subbands on CMU-ARCTIC database.

Frequency subbands (kHz)	0-0.5	0-1	0-1.5	0-2	0-2.5	0-3	0-3.5	0-4
Male (bdl)	**91.88**	91.77	91.51	90.90	91.05	90.85	90.74	90.24
Female (slt)	**93.72**	93.51	93.35	93.33	93.06	93.20	93.16	93.39
Male (jmk)	**88.45**	88.31	88.05	87.05	87.37	86.84	86.53	86.10

Table 4.2: Objective scores (in %) for the V/UV segmentation of speech using different features (CMU-ARCTIC database)

	MCF + STE	ZC + STE	NAC + STE	LELPRES + STE
	(a) Male speaker (bdl)			
V→V	91.02	82.80	84.37	83.00
V→UV	8.98	17.20	15.62	17.00
UV→UV	97.29	97.67	93.12	94.87
UV→V	2.71	2.33	6.90	5.13
Avg. DR	**94.16**	90.23	88.74	88.98
	(b) Female speaker (slt)			
V→V	96.60	93.70	94.60	93.30
V→UV	4.38	6.32	5.40	6.70
UV→UV	91.70	93.84	87.75	91.33
UV→V	8.30	6.15	12.25	8.67
Avg. DR	**94.00**	93.67	91.18	92.31
	(b) Male speaker (jmk)			
V→V	80.67	75.00	77.42	71.42
V→UV	19.32	25.00	22.60	28.62
UV→UV	99.70	99.44	97.31	97.15
UV→V	0.26	0.56	2.70	2.84
Avg. DR	**90.21**	87.2	87.36	84.26

short-time energy to obtain the feature vector $[M_l(t_i)\ M_h(t_i)\ E_n(t_i)]^T \in \mathbb{R}^3$, which is used for training the binary classifier. In the following section, we show that the combined features have a higher accuracy.

Table 4.3: Objective scores (in %) for the V/UV segmentation of speech using different features for CSTR-FDA database

	MCF + STE	ZC + STE	NAC + STE	LELPRES + STE
	(a) Male speaker (RL)			
V→V	85.45	70.85	73.31	73.18
V→UV	14.54	29.15	26.70	26.82
UV→UV	88.16	94.73	83.00	88.31
UV→V	11.84	5.27	17.14	11.70
Avg. DR	**86.81**	83.00	78.10	80.94
	(b) Female speaker (SB)			
V→V	86.17	69.55	72.26	73.50
V→UV	14.00	30.45	27.73	26.53
UV→UV	96.30	98.96	95.18	96.64
UV→V	3.70	1.03	4.81	3.35
Avg. DR	**91.24**	84.26	83.72	85.10

4.1.3 *Performance Evaluation*

We evaluate the performance of the coherence-based feature against the benchmark features: (1) zero crossing rate (ZCR), (2) normalized autocorrelation coefficient (NAC), and (3) log energy of linear prediction error (LELPRES). For a fair comparison, each of these features is also combined with the normalized short-time energy (STE). Four classifiers are trained in a supervised fashion for each of the combinations. We display the results in the form of a confusion matrix using the following notations.

Table 4.2 shows the performance comparison. We observe that the performance

V→V: fraction of voiced speech samples classified as voiced
V→UV: fraction of voiced speech samples classified as unvoiced
UV→UV: fraction of unvoiced speech samples classified as unvoiced
UV→V: fraction of unvoiced speech samples classified as voiced

of (MCF + STE) feature for both male and female speakers is superior than the

state-of-the-art. In addition, from Table 4.1 and Table 4.2, we also observe that a binary classifier trained on the joint MCFs of the first two frequency subbands is more accurate than a classifier trained individually on the MCFs of either of the subbands. These results show that the average detection rate of such a classifier is higher by about 2% and 1% for male and female speakers, respectively, in comparison to a classifier trained on individual MCFs. These results show that it is more reliable to combine the MCFs of the first two frequency subbands.

To assess the performance on voices that are not part of the training, we choose CSTR-FDA database [119]. This database has a total of 100 speech utterances (50 male and 50 female) along with the parallel EGG recordings. The male and female speakers in the database are named as "RL" and "SB," respectively. All the speech waveforms were downsampled to 8 kHz. Table 4.3 displays the performance of the proposed features against the existing features evaluated on the whole CSTR-FDA database. We observe that the proposed features outperform the existing ones indicating their generalizability to unseen data.

Next, we describe modelling of speech aperiodicity from the carrier spectrogram. In Section 4.4, we show that accurate V/UV decisions and aperiodicity estimate along with other analysis parameters directly benefit the reconstruction of perceptually high-quality speech.

4.2 Speech Aperiodicity

The aperiodicity in speech essentially quantifies its randomness, which is due to jitter, shimmer and turbulent noise at the glottis. Jitter is caused by perturbations in the periodic structure from one laryngeal cycle to another, whereas shimmer reflects variations among epochal amplitudes across laryngeal cycles. While unvoiced sounds are random with a high degree of aperiodicity, there is strong evidence that even voiced sounds possess some degree of aperiodic components [90, 91, 120–122].

Figure 4.3(a) shows a voiced speech segment and Figure 4.3(b) shows its power

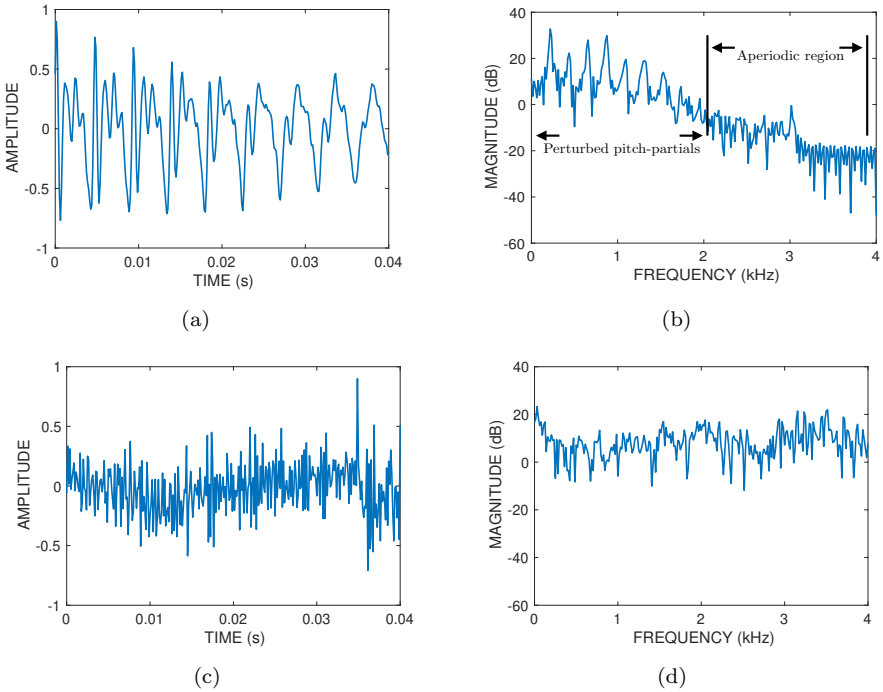


Figure 4.3: (a) A voiced speech segment, (b) its power spectrum, (c) unvoiced speech segment, and (d) its power spectrum.

spectrum. While the low-frequency part of the spectrum displays strong periodicity in terms of the pitch harmonics, the high-frequency region is mostly noise-like and reflects the aperiodicity. On the contrary, consider an unvoiced segment shown in Figure 4.3(c). Its spectrum shown in Figure 4.3(d) does not indicate the presence of periodicity in the signal. Speech aperiodicity, therefore, is not only time-dependent, but also frequency-dependent.

We have seen in Chapter-3 that there are "devoiced spectrotemporal patches," which depend to a large extent on the talker and the conditions of phonation. The variations in the pitch harmonics of the carrier spectrogram due to speech aperiodicity can be quantified using the frequency modulation of the primary sinusoid of a voiced speech segment. According to Quatieri et al. [42, 123], a primary sinusoid is the component of a voiced speech that oscillates at the fundamental frequency with

mild frequency modulation around it. One could define its 2-D counterpart: A 2-D primary sinusoid is the component of the spectrotemporal representation of voiced speech that oscillates at the grating frequency with mild frequency modulation around it. The carrier spectrogram facilitates both ways of analysis: 1-D analysis corresponding to a column (Section 4.3), or 2-D perspective considering patches (Section 4.3.3). Before proceeding further, we review previous attempts in the literature for modelling speech aperiodicity.

4.2.1 *Prior Art in Speech Aperiodicity Modeling*

A significant contribution for estimating speech aperiodicity comes from the literature on speech vocoders where the aperiodicity parameters play a crucial role in modelling the stochastic component of the speech signal. Vocoders based on source-filter theory of speech production follow an *analysis-by-synthesis* approach. The analysis aims to obtain high-resolution estimates of the fundamental frequency of the speaker, vocal tract envelope, voiced/unvoiced decisions and the band aperiodicity parameters, while the synthesis focuses on combining the parameters to reconstruct speech. The whole Nyquist band is divided into subbands and aperiodicity parameters are estimated for each subband since aperiodicity is a function of both time and frequency (Section 4.2).

A seminal contribution for the estimation of speech aperiodicity was made by Kawahara [124] who developed the STRAIGHT vocoder [44]. Following the working principles of STRAIGHT, Morise *et al.* [45] developed the WORLD vocoder. Kawahara and Morise [125] proposed an LPC-based framework which utilized the energy of linear prediction residual in a frequency band for estimating the strength of the aperiodic component. Subsequently, they proposed an alternative using temporally stable minimum-interference power spectra using which one can quantify the aperidicity [126]. Other methods for aperiodicity estimation rely on the group delay [124,127], complex-valued wavelet analysis [128], etc.

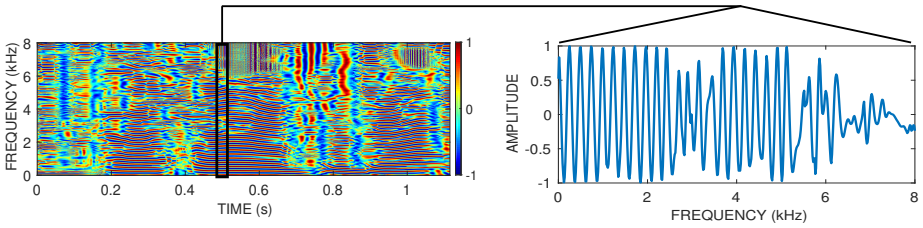


Figure 4.4: A carrier slice (right) taken from the carrier spectrogram (left) within a region enclosed by the rectangle.

4.3 A New Speech Aperiodicity Measure

4.3.1 *Synthetic Signals*

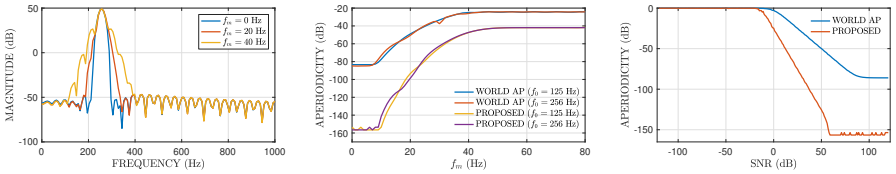
Consider a vertical slice of the carrier spectrogram corresponding to a voiced segment. It resembles a nearly constant amplitude sinusoid comprising only the quasi-periodic excitation as shown in Figure 4.4. Henceforth, we refer to time-localized carrier slices as the "carrier sinusoids". A carrier sinusoid is akin to a primary sinusoid.

Consider a narrowband frequency modulated sinusoid $c(t)$:

$$c(t) = \cos\left(\omega_0 t + k_f \int_{-\infty}^{t} m(\tau)\,\mathrm{d}\tau\right), \tag{4.9}$$

where $\omega_0 = 2\pi f_0$, $m(\tau)$ is the modulating signal, k_f is the modulation constant, $|k_f m(t)| \ll \omega_0$, and f_0 denotes the carrier frequency. In general, frequency modulated sinusoids have infinite bandwidth, but considering narrowband FM permits us to confine the bandwidth only to the first sideband. Frequency modulation is the cause of aperiodicity in voiced speech segments. Consider the windowed signal $c_w(t) = c(t)w(t)$, where $w(t)$ represents a window. Specifically, we use the Nuttal window [129] due to its high side-lobe attenuation (≈ 100 dB). Without FM, the spectrum is simply the spectrum of Nuttal window centered at ω_0. With FM, the spectrum spills over to the side-bands. Consider the Fourier transform $\hat{c}_w(\omega)$ of $c_w(t)$ and the normalized power spectrum

$$P_c(\omega) = \frac{|\hat{c}_w(\omega)|^2}{\int |\hat{c}_w(\omega)|^2 \mathrm{d}\omega}. \tag{4.10}$$



(a) (b) (c)

Figure 4.5: (a) The Fourier magnitude spectrums of a windowed frequency-modulated sinusoids at different carrier frequencies, (b) the behavior of the proposed aperiodicity measure with respect to the modulating frequency, and (c) aperiodicity measure versus signal-to-noise ratio.

A measure of aperiodicity using $P_c(\omega)$ is given by

$$A = 1 - \int P_c(\omega) M_{\omega_0}(\omega; \omega_n)\, \mathrm{d}\omega, \tag{4.11}$$

where ω_n denotes the bandwidth of the Nuttal window and

$$M_{\omega_0}(\omega; \omega_n) = \begin{cases} 1, \, |\omega - \omega_0| < \frac{\omega_n}{2}, \\ 1, \, |\omega + \omega_0| < \frac{\omega_n}{2}, \\ 0, \, \text{otherwise}, \end{cases} \tag{4.12}$$

represents a spectral mask for selecting the dominant subbands. The aperiodicity measure (AP) takes continuous values between 0 and 1.

Now, let us analyze the effect of FM on the aperiodicity measure. Consider the signal in Equation (4.9) with

$$m(t) = \sin(2\pi f_m t), \tag{4.13}$$

i.e., a single-tone modulation, where f_m denotes the modulating frequency varied from 0 to 80 Hz in steps of 20 Hz. The duration of the signal is set to 0.1 seconds obtained by multiplying with a Nuttal window, and $k_f = 0.15$. The aperiodicty measure is computed for each value of f_m using Equation (4.11). In order to account for the low and high pitch cases, the carrier frequency f_0 is set either to 125 Hz or 256 Hz (the average pitch values for the male and female speakers, respectively). The Fourier magnitude spectra of the windowed signal are displayed in Figure 4.5(a)

corresponding to $f_m = 0$, $f_m = 20$, and $f_m = 40$ Hz with $f_0 = 256$ Hz. The spectral spread increases as f_m increases. The aperiodicity measure increases with the modulating frequency till $f_m \approx 40$ Hz and then it saturates. The saturation point is determined by the main lobe width of the Nuttal window which is approximately given by $4/T$ Hz, where T denotes the window duration. For comparison, we also show the aperiodicity measure using the state-of-the-art WORLD vocoder that uses D4C [127] algorithm for aperiodicity estimation. Both D4C and the proposed aperiodicity show the expected trend: increasing aperiodicity with modulating frequency for a fixed value of f_0. From Figure 4.5(b), we observe that proposed aperiodicity measure is lower than the D4C-aperidocity by about 80 dB at $f_m = 0$, which indicates that the proposed measure is more effective than D4C.

Next, we consider aperiodicity of sinusoid (without FM) in additive white Gaussian noise and SNR varying from 0 to 120 dB. The corresponding aperiodicity measure as a function of SNR is shown in Figure 4.5(c). For very low SNR (below 0 dB), the aperiodicity is nearly constant and high value (mostly noise) and for very high SNR (above 60 dB), the aperiodicity is again nearly flat, but a low value (mostly signal). It is also lower than the D4C-aperiodicity by about 40 dB at high SNR. However, both measures are close to 0 dB at very low SNR (almost noise) indicating high aperiodicity.

4.3.2 *Band Aperiodicity Parameters for Real Speech Signal*

We now describe the estimation of band aperiodicity parameters (BAP) in voiced speech segments from the carrier spectrogram. Unvoiced speech segments are identified by using V/UV decisions (cf. Section 4.1) and the aperiodicity measure is set to 1 in these regions.

Consider a voiced speech segment at any time instant and the corresponding 1-D carrier signal sliced from the carrier spectrogram (cf. Figure 4.4). We compute BAP from 1-D carrier signal by processing it on short-time basis where we divide

the full Nyquist frequency band into smaller subbands. Each segment of the carrier sinsuoid is multiplied by a Nuttal window and an overlap of 50% between consecutive subbands. We choose three subands (in kHz): 0-4, 2-6, 4-8 in the Nyquist frequency range from 0 to 8 kHz for a speech signal sampled at 16 kHz. We model the windowed carrier signal within a subband as a frequency modulated cosine signal with additive noise. For the b^{th} subband at any instant

$$\hat{c}_w^{(b)}(\omega) = \hat{w}(\omega)\hat{c}(\omega), \tag{4.14}$$

where $\hat{w}_{t_i}(\omega)$ is the Nutall window centered on the subband and

$$\hat{c}(\omega) = \cos\left(T_0\omega + k_f \int_{-\infty}^{\omega} \hat{m}(\mu)\mathrm{d}\mu\right) + \hat{n}(\omega), \tag{4.15}$$

where T_0 is a constant, $\hat{m}(\omega)$ represents the modulating function, and $\hat{n}(\omega)$ is white Gaussian noise. The corresponding normalized power spectrum is given by

$$P_c(\tilde{t}) = \frac{|c_w^{(b)}(\tilde{t})|^2}{\int |c_w^{(b)}(\tilde{t})|^2 \,\mathrm{d}\tilde{t}}$$

where $\tilde{t}$ denotes the quefrency and $c_w^{(b)}(\tilde{t})$ is the inverse Fourier transform of $\hat{c}_w^{(b)}(\omega)$. The corresponding aperiodicity measure is defined in terms of $P_c(\tilde{t})$ as

$$A^{(b)} = 1 - \int P_c(\tilde{t}) M_{T_0}(\tilde{t}; T_n)\,\mathrm{d}\tilde{t}, \tag{4.16}$$

where

$$M_{T_0}(\tilde{t}; T_n) = \begin{cases} 1, & |\tilde{t} - T_0| < \frac{T_n}{2}, \\ 1, & |\tilde{t} + T_0| < \frac{T_n}{2}, \\ 0, & \text{otherwise}, \end{cases}$$

represents the cepstral mask, T_n denotes the main lobe width of the Nuttal window on the quefrency axis. For each subband, the value of T_0 is given by the location of the maximum peak in the normalized power spectrum.

To obtain the aperiodicty for the full band, the band aperiodicity parameters are linearly interpolated on the log scale. The aperiodicity value at zero frequency is

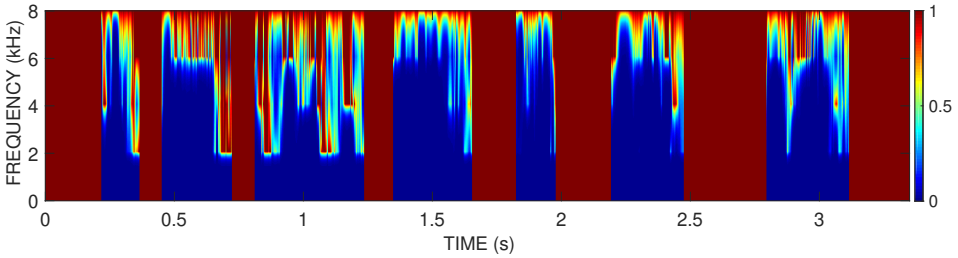


Figure 4.6: Time-frequency map of bandwise aperiodicity parameters by processing the carrier spectrogram on frame-by-frame basis. The speech utterance is "*Author of the danger trail, Philip Steels, etc.*," spoken by a female speaker.

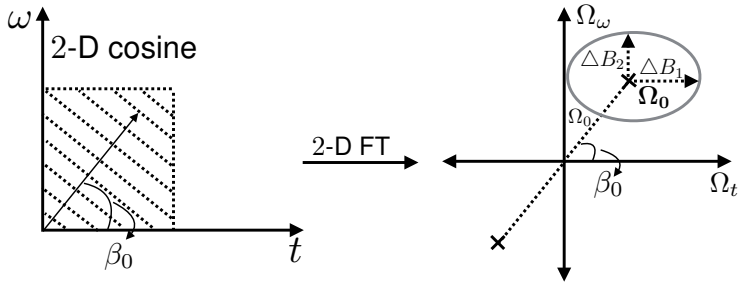


Figure 4.7: Schematic of a 2-D sinusoid (left) and its Fourier transform (right). The spatial frequency and orientation of sinusoid are denoted by Ω_0 and β_0, respectively. A mask around the peak located at $\boldsymbol{\Omega}_0 = (\Omega_0 \cos\beta_0, \Omega_0 \sin\beta_0)$ is illustrated by an ellipse.

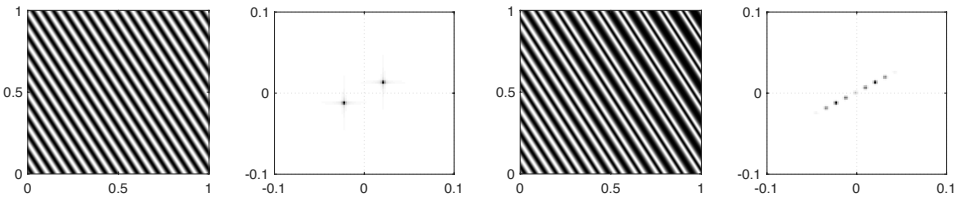
set to -60 dB. The resulting t-f aperiodicity map is shown in Figure 4.6 along with V/UV decisions. We observe that the aperiodicity for voiced sounds is relatively high roughly above 3 kHz.

4.3.3 *Aperiodicity in 2-D*

The aperiodicity estimation framework can be extended to 2-D using localized patches taken from the carrier spectrogram. Consider the 2-D FM sinusoid model $C(\boldsymbol{\omega}) : \mathbb{R}^2 \to \mathbb{R}$ given by

$$C(\boldsymbol{\omega}) = \cos\big(\Omega(\boldsymbol{\omega})(t\cos\beta_0 + \omega\sin\beta_0)\big), \tag{4.17}$$

where β_0 and $\Omega(\boldsymbol{\omega}) = \Omega_0 + k_f\Delta\Omega(\boldsymbol{\omega}) : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ represent the local orientation and grating frequency of the cosine, respectively. The strength of the FM is given



(a) (b) (c) (d)

Figure 4.8: (a) A 2-D cosine with constant frequency Ω_0, (b) its GCT, (c) a cosine with frequency modulation $k_f \Delta\Omega(\boldsymbol{\omega}) = 0.015\cos(10\pi\Phi_0(\boldsymbol{\omega}))$, and (d) its GCT.

by the modulation constant k_f. A schematic of a 2-D cosine with no frequency modulation ($k_f = 0$) and its Fourier transform is shown in Figure 4.7. We consider a 2-D sinusoid to be aperiodic if $k_f \Delta\Omega(\boldsymbol{\omega}) \neq 0$ and $\|\nabla_{\boldsymbol{\omega}} \Delta\Omega(\boldsymbol{\omega})\| \ll \Omega_0$. Let $\Phi_0(\boldsymbol{\omega}) = t\cos\beta_0 + \omega\sin\beta_0 : \mathbb{R}^2 \to \mathbb{R}$, corresponding to which

$$\begin{aligned} C(\boldsymbol{\omega}) &= \cos\big((\Omega_0 + k_f\Delta\Omega(\boldsymbol{\omega}))\Phi_0(\boldsymbol{\omega})\big) \\ &= \cos\big(\Omega_0\Phi_0(\boldsymbol{\omega})\big)\cos\big(k_f\Delta\Omega(\boldsymbol{\omega})\Phi_0(\boldsymbol{\omega})\big) - \sin\big(\Omega_0\Phi_0(\boldsymbol{\omega})\big)\sin\big(k_f\Delta\Omega(\boldsymbol{\omega})\Phi_0(\boldsymbol{\omega})\big). \end{aligned} \tag{4.18}$$

Under the NBFM assumption $|k_f\Delta\Omega(\boldsymbol{\omega})\Phi_0(\boldsymbol{\omega})| \ll 1$, Equation (4.18) is approximated as follows:

$$C(\boldsymbol{\omega}) \approx \cos\big(\Omega_0\Phi_0(\boldsymbol{\omega})\big) - k_f\Delta\Omega(\boldsymbol{\omega})\Phi_0(\boldsymbol{\omega})\sin\big(\Omega_0\Phi_0(\boldsymbol{\omega})\big). \tag{4.19}$$

The first term in Equation (4.19) is a primary 2-D sinusoid of frequency Ω_0 and the second term is the side-band due to FM. For illustration, consider the modulating function to be a slowly-varying sinusoid, $\Delta\Omega(\boldsymbol{\omega}) = \cos\big(\Omega_m\Phi_0(\boldsymbol{\omega})\big)$ with $\Omega_m \ll \Omega_0$. Figure 4.8 shows the GCTs of a 2-D cosine one with a constant frequency and one with FM. The appearance of the side-band in Figure 4.8(d) indicates the spectral spread caused by FM.

Consider a windowed 2-D cosine $C_W(\boldsymbol{\omega}) = W(\boldsymbol{\omega})C(\boldsymbol{\omega})$, where $W(\boldsymbol{\omega})$ represents

a real 2-D window function. From Equation (4.19), we have

$$C_W(\boldsymbol{\omega}) \approx W(\boldsymbol{\omega}) \cos\big(\Omega_0 \Phi_0(\boldsymbol{\omega})\big) - \underbrace{k_f \Delta\Omega(\boldsymbol{\omega}) \Phi_0(\boldsymbol{\omega}) W(\boldsymbol{\omega})}_{F(\boldsymbol{\omega})} \sin\big(\Omega_0 \Phi_0(\boldsymbol{\omega})\big). \tag{4.20}$$

Denoting $\boldsymbol{\Omega}_0 = (\Omega_0 \cos\beta_0, \Omega_0 \sin\beta_0) \in \mathbb{R}^2$ and taking 2-D Fourier transform on both sides in Equation (4.20), we have

$$\hat{C}_W(\boldsymbol{\Omega}) \approx \hat{W}(\boldsymbol{\Omega} - \boldsymbol{\Omega}_0) + \hat{W}(\boldsymbol{\Omega} + \boldsymbol{\Omega}_0) + \mathrm{j}(\hat{F}(\boldsymbol{\Omega} - \boldsymbol{\Omega}_0) - \hat{F}(\boldsymbol{\Omega} + \boldsymbol{\Omega}_0)). \tag{4.21}$$

The normalized power spectral density is given by $P_C(\boldsymbol{\Omega}) = \frac{|\hat{C}_W(\boldsymbol{\Omega})|^2}{\int |\hat{C}_W(\boldsymbol{\Omega})|^2 \, \mathrm{d}\boldsymbol{\Omega}}$. The aperiodicity measure $\mathcal{A}_{\boldsymbol{\Omega}_0}$ of a 2-D sinusoid is defined in terms of $P_C(\boldsymbol{\Omega})$ as

$$\mathcal{A}_{\boldsymbol{\Omega}_0} = 1 - 2 \int P_C(\boldsymbol{\Omega}) M_{\boldsymbol{\Omega}_0}(\boldsymbol{\Omega}; \Delta\boldsymbol{B}) \, \mathrm{d}\boldsymbol{\Omega}, \tag{4.22}$$

where

$$\boldsymbol{M}_{\boldsymbol{\Omega}_0}(\boldsymbol{\Omega}; \Delta\boldsymbol{B}) = \begin{cases} 1, \, |\Omega_t - \Omega_{0_1}| < \Delta B_1, |\Omega_\omega - \Omega_{0_2}| < \Delta B_2, \\ 0, \text{ otherwise}, \end{cases} \tag{4.23}$$

represents the spectral mask in 2-D with $\Delta\boldsymbol{B} = (\Delta B_1, \Delta B_2) \in \mathbb{R}^2_{>0}$ and $(\Omega_{0_1}, \Omega_{0_2}) = (\Omega_0 \cos\beta_0, \Omega_0 \sin\beta_0)$ as shown in Figure 4.7. ΔB_1 and ΔB_2 denote main-lobe widths of the window function $W(\boldsymbol{\omega})$ along Ω_t-axis and Ω_ω-axis, respectively. The aperiodicity measure $\mathcal{A}_{\boldsymbol{\Omega}_0}$ takes on continuous values between 0 and 1 with 0 corresponding to a 2-D sinusoid, which is perfectly periodic; and 1 corresponding to an aperiodic signal. The measure $\mathcal{A}_{\boldsymbol{\Omega}_0}$ quantifies the degree of aperiodicity within the patch.

Figure 4.9 illustrates the proposed aperiodicity measure for a synthetic 2-D cosine with single-tone modulation. The figure reflects the increasing aperiodicity with the modulating frequency. The saturation is because after a certain value of the modulating frequency, the normalized power spectral density within the bandwidth of 2-D mask $M_{\boldsymbol{\Omega}_0}(\boldsymbol{\Omega}; \Delta\boldsymbol{B})$ in Equation (4.22) does not change appreciably.

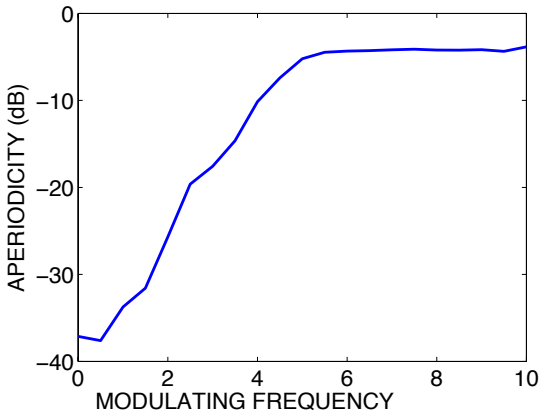


Figure 4.9: Aperiodicity measure of a synthetic 2-D cosine signal $F(\boldsymbol{\omega}) = \cos((\Omega_0 + k_f \Delta\Omega(\boldsymbol{\omega}))(t\cos\beta_0 + \omega\sin\beta_0))$, where $\Omega_0 = 40\pi$, $k_f = 0.015$, $\beta_0 = \pi/6$, and $\Delta\Omega(\boldsymbol{\omega}) = \cos(2\pi f_m(t\cos\beta_0 + \omega\sin\beta_0))$ with f_m varying from 0 to 10 in steps of 0.5.

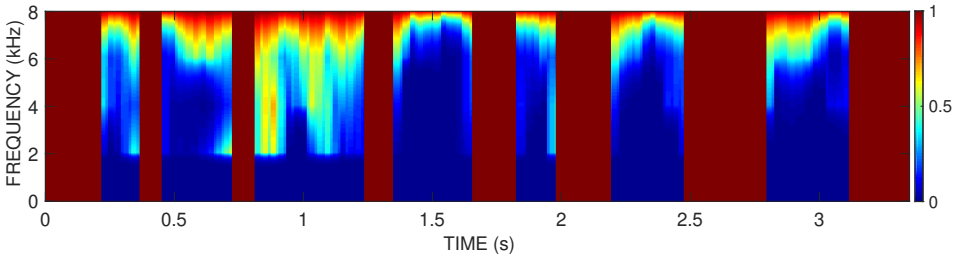


Figure 4.10: Time-frequency map of bandwise aperiodicity parameters by processing the carrier spectrogram on patch-by-patch basis. The speech utterance is "*Author of the danger trail, Philip Steels, etc.,*" spoken by a female speaker.

4.3.4 *Spectrotemporal Aperiodicity Map*

We divide the carrier spectrogram into patches of dimension 100 ms × 600 Hz and with an overlap of 75% along both the dimensions. Each patch is multiplied by a 2-D Nuttal window and the aperiodicity is computed using Equation (4.22). The aperiodicity value is constant for the entire patch assuming that it does not vary by large amounts within the patch. The patch-wise aperiodicity values are combined using 2-D overlap-add in least-squares sense (OLA-LSE) [130] that yields a spectrotemporal aperiodicity map, which is further used to obtain the t-f map

of the band-wise aperiodicity parameters. Considering a sampling rate of 16 kHz, the Nyquist frequency band is divided into three overlapping subbands, each of size 2 kHz with 50% overlap between consecutive subband. The aperiodicity in a subband is chosen to be the value from the spectrotemporal aperiodicity map at the middle of the subband. The band-wise parameters are linearly interpolated on the logarithmic scale. The aperiodicity value is set to -60 dB at zero frequency [127]. This gives another t-f map, which is then combined with V/UV decisions (Section 4.1), and the resulting t-f map is shown in Figure 4.10. This map is similar to that obtained using the 1-D approach (cf. Figure 4.6). Next, we compare the efficacy of the 1-D and 2-D approaches for the task of speech reconstruction by incorporating them in the state-of-the-art WORLD vocoder.

4.4 Application to Speech Reconstruction

The WORLD vocoder uses the spectral synthesis model for speech reconstruction. Recall that the spectrum of a voiced speech frame $\hat{s}_v(\omega)$ in the spectral synthesis model given by

$$\hat{s}_v(\omega) = \hat{v}(\omega)\Big(\sqrt{T_0}\big(1-\hat{a}(\omega)\big) + \hat{n}(\omega)\hat{a}(\omega)\Big), \tag{4.24}$$

where $\hat{v}(\omega)$, T_0, $\hat{n}(\omega)$, and $\hat{a}(\omega)$ denote the vocal-tract frequency response, pitch period, Fourier spectrum of zero mean and unit variance Gaussian noise and the aperiodicity parameters, respectively. The spectrum of an unvoiced speech frame is modeled as

$$\hat{s}_{uv}(\omega) = \hat{v}(\omega)\hat{n}(\omega)\hat{a}(\omega). \tag{4.25}$$

We rely on the PESQ score for comparing the performance between the 1-D and 2-D aperiodicity maps. We consider two male (bdl, ksp), and two female voices (slt, clb) from the CMU-ARCTIC database. A total of 200 speech utterances are used for evaluation where the speech waveforms are reconstructed by replacing the

Table 4.4: Average PESQ scores for the quality of reconstructed speech on CMU-ARCTIC database

	bdl	ksp	clb	slt
CRT-AP1D	3.34 ± 0.12	3.42 ± 0.12	3.34 ± 0.15	3.42 ± 0.12
CRT-AP2D	3.34 ± 0.12	3.33 ± 0.12	3.22 ± 0.14	3.37 ± 0.13

aperiodicity parameters in the spectral synthesis model of WORLD vocoder with the proposed ones. Table 4.4 shows the average PESQ scores for the two cases. The performance in both cases is comparable. Some reconstructed speech samples are provided online: `https://jitendradhiman.github.io/CRT1DAPCRT2DAP.html`. Informal listening tests showed that both the aperiodicity maps are on par with each other.

4.5 Chapter Summary

We addressed the problem of V/UV segmentation and the estimation of speech aperiodicity. We used the carrier spectrogram estimated using the Riesz transform approach for both the tasks. For V/UV segmentation, we derived a novel feature (MCF) from the coherencegram, the performance of which was objectively evaluated against state-of-the-art features. The MCF is also insensitive to the local temporal variations of the speech waveform and gives superior performance over the existing features. We also introduced a novel measure of aperiodicty base on the spectrum-spread property caused by frequency modulation. The aperiodicity was estimated from the carrier spectrogram using both frame-based and patch-based approaches. Following this, the band aperiodicity parameters were derived. The effectiveness of the proposed band aperiodicity parameters and the proposed V/UV decisions was assessed for the task of speech reconstruction by incorporating them in the state-of-the-art WORLD vocoder. Our results showed that the new aperiodicity map gives similar quality of speech reconstruction in both 1-D and 2-D variants although, computationally, the 2-D version is more involved.

Chapter 5

Pitch Estimation From the Carrier Spectrogram

Thus far, we have used the carrier spectrogram for periodic/aperiodic decomposition, voiced/unvoiced segmentation and for modelling the aperiodicity of a speech signal. In this chapter, we examine the carrier spectrogram for temporal evolution of the speaker's fundamental frequency (or pitch) and its harmonics. We discuss briefly the problem of pitch estimation and give the definition of pitch in Section **??**. The problem of pitch estimation has been addressed extensively in the literature. Section 5.1 introduces some of the widely used state-of-the-art pitch estimation algorithms over the past decade. In Section 5.2, we introduce the proposed techniques and highlight advantages over the existing ones. In Section 5.3, we present an objective comparison with the state-of-the-art methods.

The voiced speech sounds are characterized by the vibrations of the vocal folds yielding a quasi-periodic structure of the sounds in the time-domain [42]. In psychoacoustics, the perceptual correlate of the vocal-fold vibrations is referred to as the pitch [131]. In speech signal processing, the definition of pitch is often motivated from the speech production perspective. Pitch (or F0) measures the rate of vibration of the vocal folds during voicing and is the fundamental frequency of the vibrations. We will use this definition of pitch. The pitch of a speaker is the most prominent feature of the glottal excitation that directly affects the prosodic aspects of a speech waveform such as intonation, stress, tone, and rhythm. F0 information is useful in a variety of speech processing tasks. The estimation of F0-dependent spectral

envelope is also used to model the magnitude response of the vocal-tract filter [44]. F0 variation is inherently a continuous phenomenon as it varies often with a glottal cycle. Short-time analysis of the speech signal shows the time-varying F0. One could also estimate F0 without resorting to short-time processing such as the one reported in [132].

Accurate pitch estimation and tracking have been of considerable interest for applications such as synthesis [44,45,133], transmission [134], articulation training aids for the deaf [119,135], speaker recognition [136], prosody modification [137,138], foreign language training [139,140], etc. There is no universally optimal F0 estimation algorithm [141] that is preferred in all speech applications. For instance, a source filter theory-based speech reconstruction requires accurate pitch estimates at the resolution of one glottal cycle. On the other hand, speaker recognition applications in noise require robust pitch estimates. A comprehensive review of pitch estimation algorithms is given in [142]. We briefly review a few important methods while also highlighting the associated challenges.

5.1 Prior Art in Pitch Estimation

Pitch estimation techniques broadly operate in the time domain, frequency domain, or the time-frequency domain. Time-domain methods such as the autocorrelation technique exploit the repetitive structure of voiced speech sounds. On the other hand, frequency-domain methods such as the cepstrum exploit the harmonic structure of the Fourier spectrum. Time-frequency techniques such as YAAPT [143] consider both temporal and spectral structure for pitch estimation.

5.1.1 *Autocorrelation Method*

Autocorrelation is a measure of similarity of a signal with itself. Consider a discrete-time sequence $s[n]$ and a window $w[n]$, where n denotes the discrete-time index. The

autocorrelation of the windowed sequence $s_w[n] = s[n]w[m-n]$ at lag l is given by

$$r[l] = \sum_{n=-\infty}^{\infty} s_w[n]s_w[n+l]. \tag{5.1}$$

The autocorrelation is an even function and attains a maximum at $l = 0$. With a suitably chosen window duration, the autocorrelation sequence exhibits peaks at the $0, \pm P_0, \pm 2P_0, \ldots$ for a periodic signal with time period P_0. The window length must be more than two pitch periods for the peaks to occur in autocorrelation. Hence, regardless of the time origin of the waveform, an estimate of F0 is obtained by choosing highest peak at the non-zero lag within a search range of possible lags in the autocorrelation.

Although the technique is computationally faster, it is prone to estimation errors [144]. Due to the quasi-periodic nature of a speech waveform, it may exhibit more self-similarity at a lag equal to twice the pitch period. Consequently, the second maximum in the autocorrelation occurs at $\pm 2P_0$ and the F0 is underestimated by an octave – this is the *pitch halving*. *Pitch doubling* occurs when the F0 is overestimated by an octave. The method can show the dominant peaks at subharmonic levels. Consequently, it is often difficult to decide whether a peak belongs to the fundamental or its partial. Interactions between the vocal tract and glottal excitation can also lead to estimation errors. The vocal-tract resonances, typically the first formant, can emphasize harmonics other than the first, causing overestimation of F0. Nonlinear methods overcome the limitations of the autocorrelation method. The techniques include center clipping, lowpass filtering, and filter-bank approaches [107].

5.1.2 *Cepstrum Analysis for Pitch Estimation*

We now consider cepstrum-based F0 estimation [145]. Consider a sequence whose discrete-time Fourier transform (DTFT) $S(e^{\mathrm{j}\omega})$. The real cepstrum (or cepstrum) is

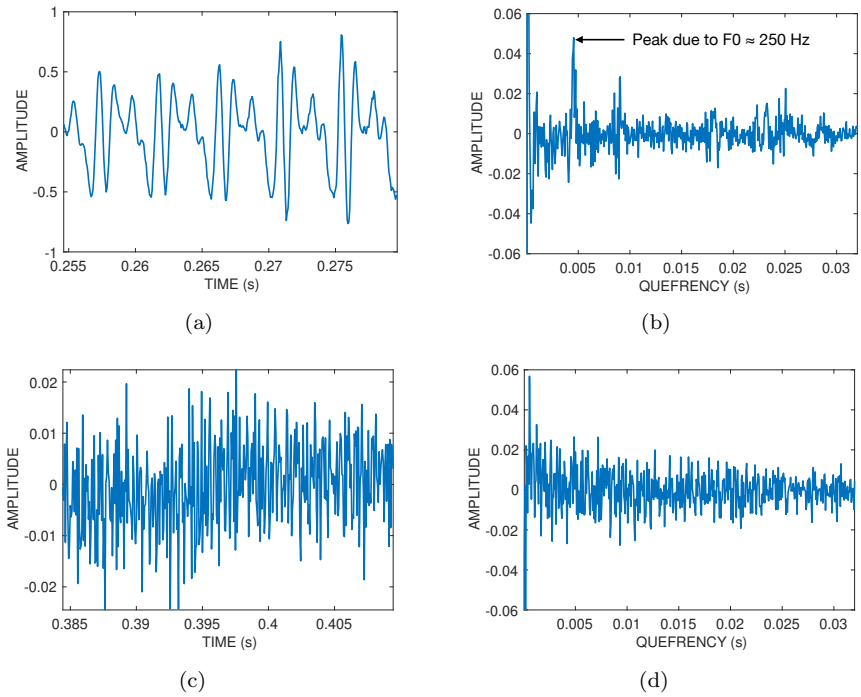


Figure 5.1: (a) A voiced speech segment, (b) its real cepstrum, (c) unvoiced speech segment, and (d) its real cepstrum. The speech utterance is "*Author of the danger trail, Philip Steels, etc.*," spoken by a female speaker having average F0 = 250 Hz. The cepstrum shows a dominant peak at fundamental frequency of the speaker, whereas such a peak is absent in the cepstrum of an unvoiced segment.

given by the inverse Fourier transform of the logarithm of the magnitude spectrum:

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |S(e^{\mathrm{j}\omega})| e^{\mathrm{j}\omega n} \, \mathrm{d}\omega, \tag{5.2}$$

where n denotes the time index on *quefrency* axis. The logarithm operation flattens the spectrum to a certain extent reducing the strength of the high-amplitude part, particularly near the first formant. Spectral flattening can also be achieved by inverse filtering [42]. The effect of the first formant that often dominates in the autocorrelation method [144, 146] is thus mitigated to some extent in cepstrum analysis. F0 estimation from cepstrum is based on the observation that the periodic oscillations in the log spectrum are reflected as a strong peak at the fundamental

period in the cepstrum of a voiced speech segment and no such peak appears for the unvoiced segments. Figure 5.1 displays the cepstrum corresponding to voiced and unvoiced segments. The peak in the cepstrum of voiced sounds is used to estimate pitch and can also be used to distinguish voiced segments from unvoiced ones. The peak is searched in the vicinity of the expected pitch period. If the peak is above a preset threshold, then the segment is voiced, else unvoiced. The reliability of F0 estimates can be improved by combining decisions from zero-crossing rate, energy and by forcing the pitch estimate to vary smoothly. One such algorithm was described by Noll [145]. Subsequently, a variety of cepstrum based F0 estimation algorithms have been developed [19,147].

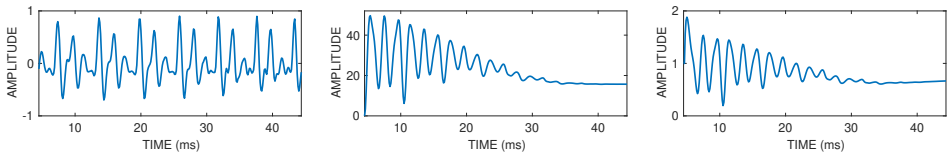
5.1.3 *YIN*

The YIN pitch estimation algorithm [148] was developed by Alain de Cheveigne and Hideki Kawahara. The difficulty with autocorrelation technique has been that the peaks occur at harmonics and sub-harmonics as well, it is sometimes challenging to determine which peak corresponds to the fundamental frequency or its partial. The authors attempt to overcome these limitations of autocorrelation in several ways. YIN is based on the difference function between the waveform and its delayed duplicate, unlike autocorrelation which uses the product.

An unknown period can be detected by forming the difference function

$$d_m[l] = \sum_{n=-\infty}^{\infty} (s_w[n] - s_w[n+l])^2, \tag{5.3}$$

where m and l denote position-index of the window and the time lag, respectively. If the signal $s_w[n]$ is perfectly periodic with period P_0, then the difference function exhibits dips at 0, P_0 and integer multiples of P_0. In order to reduce the errors in F0 estimation which might occur due to subharmonics, YIN uses a modified form of the



(a) (b) (c)

Figure 5.2: Illustration of (a) a voiced speech segment with period approximately 6 ms, the corresponding (b) difference function, and (c) the cumulative normalized difference function (CMNDF). CMNDF starts at value 1 and has the very first dip at the time period of voiced segment.

difference function; the cumulative mean normalized difference function (CMNDF)

$$\bar{d}_m[l] = \begin{cases} 1, & \text{if } \ l = 0, \\ \frac{d_m[l]}{\frac{1}{l}\sum_{j=1}^{l} d_m[j]}, & \text{otherwise.} \end{cases} \tag{5.4}$$

The values of this function at non-zero lags are obtained by dividing the value at the current lag by its average over the shorter lag-values. Figure 5.2 displays a voiced speech segment, its difference function and the CMNDF. Note that unlike the difference function, CMNDF starts at 1 and remains high until the period-dip. However, the CMNDF can occasionally exhibit more dominant secondary dips than the dip at the period. Consequently, YIN adopts a strategy which uses an absolute thresholding for unambiguous detection of the period-dip. However, such a pre-specified threshold might not be optimum across all the voiced frames of a speech signal. We shall see in Section 5.2.2 that the pitch estimation from carrier spectrogram does not require any such threshold-strategy. Other improvements of YIN includes parabolic interpolation of the detected period-dip which aids in improving the accuracy of the algorithm.

5.1.4 *Sum of Residual Harmonics (SRH)*

SRH [149] operates in the frequency domain and acts on the Fourier magnitude spectrum of the linear prediction residual [18] which makes it appealing for pitch

estimation under adverse conditions such as noise. The linear prediction residual exhibits peaks at the beginning of every glottal cycle, consequently the periodicity of the corresponding speech signal is preserved. For a wide variety of speech sounds, the short-time Fourier magnitude spectrum $E(f)$ of a windowed segment of residual signal has relatively a flat envelope and presents peaks at F0 and its harmonics. From this spectrum, the SRH is given by

$$SRH(f) = E(f) + \sum_{k=2}^{N_{harm}} [E(kf) - E((k-0.5)f)], \tag{5.5}$$

where N_{harm} denotes the number of harmonics within a pres-specified range of frequencies $[F_{0,min}, F_{0,max}]$ (assuming that the speaker's pitch will not exceed this frequency range), f denotes the frequency (in kHz). From Equation (5.5), it is expected that $E(f)$ is maximum at f = F0. However, the same is true for the harmonics. Hence, the subtraction by the term $E((k-0.5)f)$ reduces the relative importance of SRH maxima at even harmonics. For a voiced frame of residual signal, the estimated F_0^* is thus given by the maximum of SRH.

5.1.5 *Yet Another Algorithm for Pitch Tracking (YAAPT)*

Thus far, we have described F0 estimation methods which operate either in time or frequency domain. In contrast, YAAPT combines the F0 cues obtained by processing a speech signal independently in time and frequency domains. The kernel of this method is based on another algorithm known as $RAPT$ [141]. A combination of F0 cues obtained from both time and frequency domains makes the algorithm less sensitive to commonly occurred F0 estimation errors such as pitch doubling and pitch halving. This method was designed for robust pitch estimates for telephone and noisy speech, at signal to noise ratio varying from clean to very noisy speech. The full details of YAAPT are given in [143]. We briefly describe its basic steps.

In a preprocessing step, a bandpass filter (of bandwidth 50 to 1500 Hz) is applied to both the original acoustic waveform and a nonlinearly processed copy of the signal,

thus creating two versions of the signal. The nonlinear processing aids in partially restoring the fundamental frequency of the signal in case it is suppressed such as for a telephone speech. These two signals are then independently processed in time domain to obtain a Normalized Cross Correlation Function (NCCF) which is a modified form of the autocorrelation. In addition, the nonlinearly processed signal is also analyzed in frequency domain for estimating an initial F0 track based on spectral harmonic correlation (SHC) technique and the dynamic programming [107]. The SHC function makes use of multiple harmonic peaks of a voiced speech spectrum and exhibits a very prominent peak at the fundamental frequency. The frequency and the amplitude of each SHC peak above a preset threshold for each voiced speech frame are selected as spectral F0 candidates and merits, respectively. Additionally, the potential F0 candidates are also obtained from the time-domain NCCF's computed from both the versions of the signal. The F0 candidates obtained from time domain processing are refined by using the information from spectral domain. The final F0 track is estimated by employing dynamic programming technique which selects the lowest cost candidate among the F0 candidates. Though, this method gives reliable pitch estimates, it relies on several experimentally tuned design parameters for the accurate estimation of true F0 values.

5.1.6 *Harvest*

Harvest [150] shares some commonalities and foundations with another F0 estimation algorithm; *Time-domain Excitation extraction based on a Minimum Perturbation Operator (TEMPO)* [151, 152]. For instance, both the approaches use a bank of bandpass filters to process the speech signal in a preprocessing step for F0 estimation. These methods were especially developed for the vocoding application. The bandpass filters are uniformly placed on the log frequency axis, while their center frequencies are swiped over a range of possible pitch values. A speech signal is processed through each of the bandpass filters. The key idea is that the bandpass filter whose center

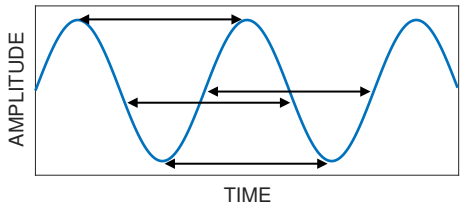


Figure 5.3: Four intervals used for the F0 estimation in Harvest from a bandpass-filtered voiced speech frame.

frequency coincides or nearly matches to the unknown pitch will generate an output waveform oscillating at the fundamental frequency. In the next step, each of the output waveforms of bandpass filters is marked with four intervals as shown in Figure 5.3. These intervals indicate the same value when the waveform is a sinusoid. The method proceeds by defining a basic pitch trajectory for a frequency sub-band (or channel) by computing the inverse of each of the intervals and taking the average of inverse values over the four intervals. This particular step is repeated for all the channels, this gives as many basic F0 candidates as the number of filters. Next, F0 candidates are obtained from the basic F0 candidates based on a bandwidth criterion. Consider a voiced speech frame and the mapping from the center frequency of the bandpass filter to the basic F0 candidate. If the basic F0 candidate comes from the fundamental component then the same value is observed over a certain bandwidth around the center frequency of a bandpass filter. In this case, a particular filter and its neighboring ones output nearly the same waveforms. Harvest selects the F0 candidates within a bandwidth on either side of the center frequency with 10% margin. These coarse F0 candidates are further refined by using the instantaneous frequency which is defined as the derivative of the short-term phase of the waveform. Additionally, each of the F0 candidate is also assigned a reliability score. Finally, the F0 candidate among the refined ones which gets the highest reliability score is chosen as the final estimate of the fundamental frequency for a voiced speech frame. In a post-processing step, the estimated F0 values are smoothed using a zero-lag Butterworth filter with a cut-off frequency of 30 Hz.

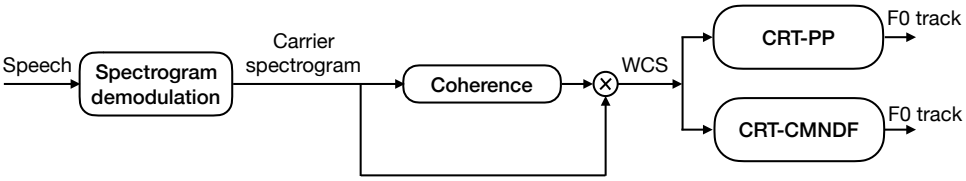


Figure 5.4: Block diagram for pitch estimation from the weighted carrier spectrogrm (WCS). Two approaches, CRT-Peak-Picking (CRT-PP) and CRT-CMNDF are proposed for the estimation of F0 track from WCS.

Other F0 estimation algorithms which employ filter-bank approaches can be found in [153–155]. In addition, data-driven statistical and machine learning approaches [102] have also been proposed.

5.2 The Proposed Technique

The carrier spectrogram is nearly free from the influence of the vocal-tract resonances. The pitch can be estimated from any t-f region of the carrier spectrogram more reliably than the standard spectrogram. A block diagram for the proposed pitch estimation method is shown in Figure 5.4. The key component is the weighted carrier spectrogram (WCS). In Chapter 3, we have shown that the carrier spectrogram reveals two distinct t-f patterns — coherent and incoherent (cf. Figure 3.1). The coherent regions are mostly in the low-frequency regions and correspond to voiced sounds, whereas the incoherent ones characterize unvoiced sounds (see Section 3.1). Prior to pitch estimation, it is desirable to separate the coherent t-f regions from the incoherent ones, for which we use the coherencegram as a weighting function. Figure 5.5 displays an example of the weighted carrier spectrogram (WCS). Ideally, the carrier in a coherent region is a sinusoid oscillating at the fundamental frequency. Recall from Chapter 4, Section 4.3, that a slice of the carrier spectrogram is referred to as the carrier sinusoid. F0 can be estimated from the carrier sinusoid either by picking the peak (denoted as CRT-PP) or by computing the cumulative mean normalized difference in Equation (5.4) (denoted as CRT-CMNDF).

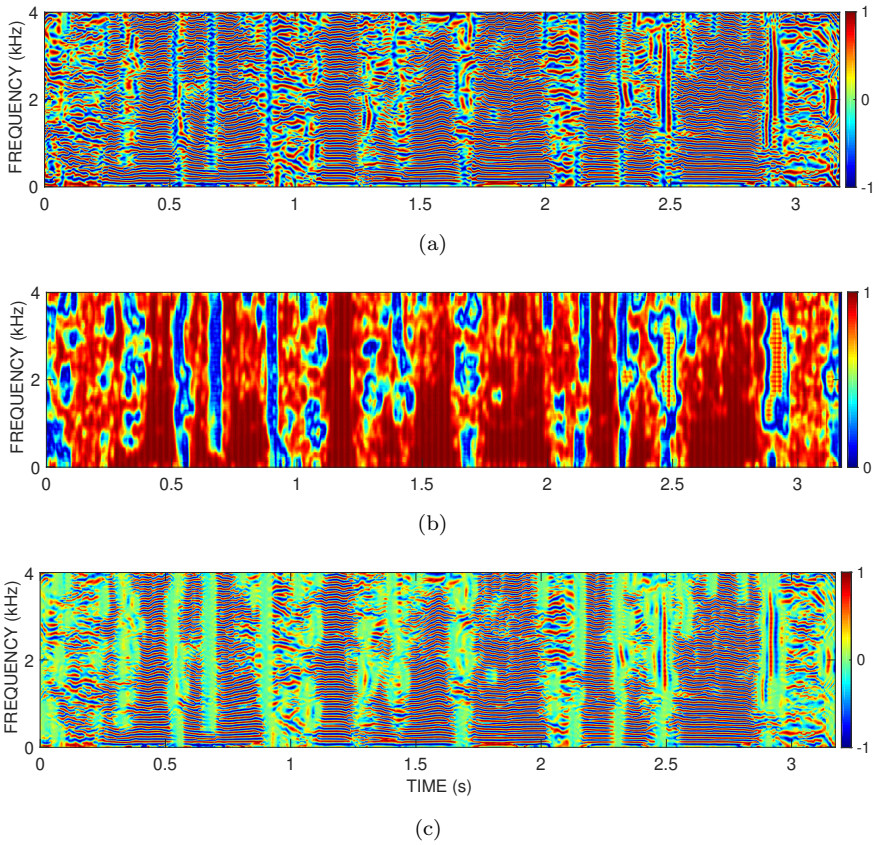


Figure 5.5: (a) The carrier spectrogram, (b) its coherence map, and (c) the weighted carrier spectrogram (WCS). The speech utterance is "*She had your dark suit in greasy wash water all year.*" spoken by a female speaker.

5.2.1 *Pitch Estimation Using Peak Picking*

Consider the carrier sinusoid within a frequency band from 0 to 1000 Hz, which covers the range of F0 values for both male and female speakers, even high-pitched ones. Figure 5.6 displays a slice of the coherence weighted carrier spectrogram. While most peaks are nearly of the same amplitude, we observe a smaller peak (around 0.7 kHz) between adjacent dominant carrier peaks. Such low amplitude peaks are caused by spectral leakage and imperfections in demodulation due to model mismatch. Since the carrier sinusoid is nearly symmetric, it suffices to work

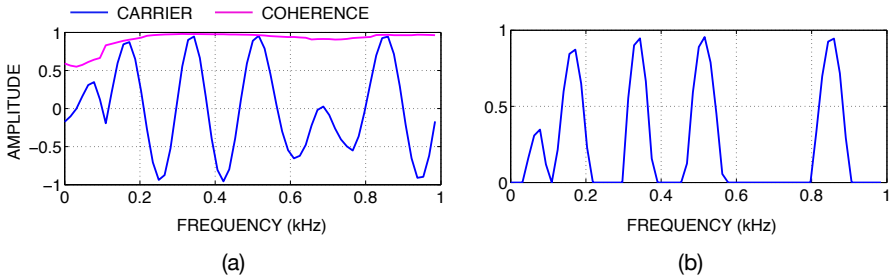


Figure 5.6: (a) A carrier-sinusoid multiplied by the corresponding coherence values, an undesirable peak exists around 0.7 kHz and (b) the carrier after thresholding operation which eliminates the undesired peak while retaining the dominant ones. The threshold value was 0.05.

with only the positive half. Empirically, a small positive threshold (0.05 in our experiments) was found to suppress small spurious peaks. Figure 5.6 displays a carrier sinusoid after thresholding.

The average pitch in the i^{th} speech frame is estimated as the harmonic mean:

$$\mathrm{F0}^{(i)} = \frac{f_s}{N} \times \frac{1}{\frac{1}{K_i}\sum_{k=1}^{K_i}\left(\frac{1}{d_{k+1}-d_k}\right)}, \tag{5.6}$$

where N and f_s denote the FFT size and sampling frequency, respectively, K_i is the number of peaks, and d_k denotes the location of k^{th} peak. Note that the arithmetic mean would effectively make use of the first and last peaks only. The estimates must be post-processed to suppress errors and get a smooth pitch contour. In nonstationary regions, such as voicing offset times when the vocal folds are relaxing and give rise to damped oscillations, most techniques fail to give reliable estimates and hence post-processing becomes necessary. Further, post-processing is applied to the segments of pitch contour corresponding to voiced segments of speech which is explicitly identified using either V/UV decisions from coherence (Chapter 4) or the EGG signal. In this chapter (Section 5.3), we use V/UV decisions from the EGG signal, which is done for a fair comparison of different pitch estimation algorithms. We employ the three-step post-processing technique as adopted in WORLD [45].

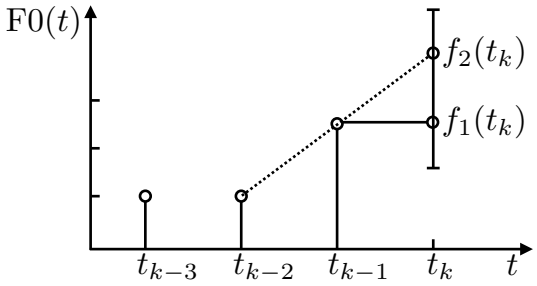


Figure 5.7: Illustration of frequencies $f_1(t_k)$ and $f_2(t_k)$ which are used to remove the rapid fluctuations in the F0 contour.

(1) **Removing spurious F0 estimates**: With reference to Figure 5.7, consider frequencies $f_1(t_k)$ and $f_2(t_k)$ at time t_k using the estimated F0 at previous time instants t_{k-1} and t_{k-2} as follows:

$$f_1(t_k) = \text{F0}(t_{k-1}), \quad \text{and}$$
$$f_2(t_k) = 2\text{F0}(t_{k-1}) - \text{F0}(t_{k-2}).$$

F0 at t_k is set to zero if it is not included in the range $[f_1(t_k) - 8\%, f_2(t_k) + 8\%]$.

(2) **Removal of short contour segments and interpolation**: Voiced segments of duration less than 3 ms are counted as invalid because voiced segments cannot be so short – they are labelled as unvoiced. The F0 values of unvoiced segments less than 15 ms duration are obtained by linearly interpolating the F0 across neighboring segments.

(3) **Smoothing**: A smooth F0 contour is obtained by lowpass filtering using a second-order Butterworth filter with cut-off frequency of 30 Hz.

Figure 5.8 shows the estimated pitch contour using CRT-PP before and after post-processing together with the corresponding speech waveform and the carrier spectrogram. The effect of post-processing to obtain a smooth pitch contour is clear from the figure.

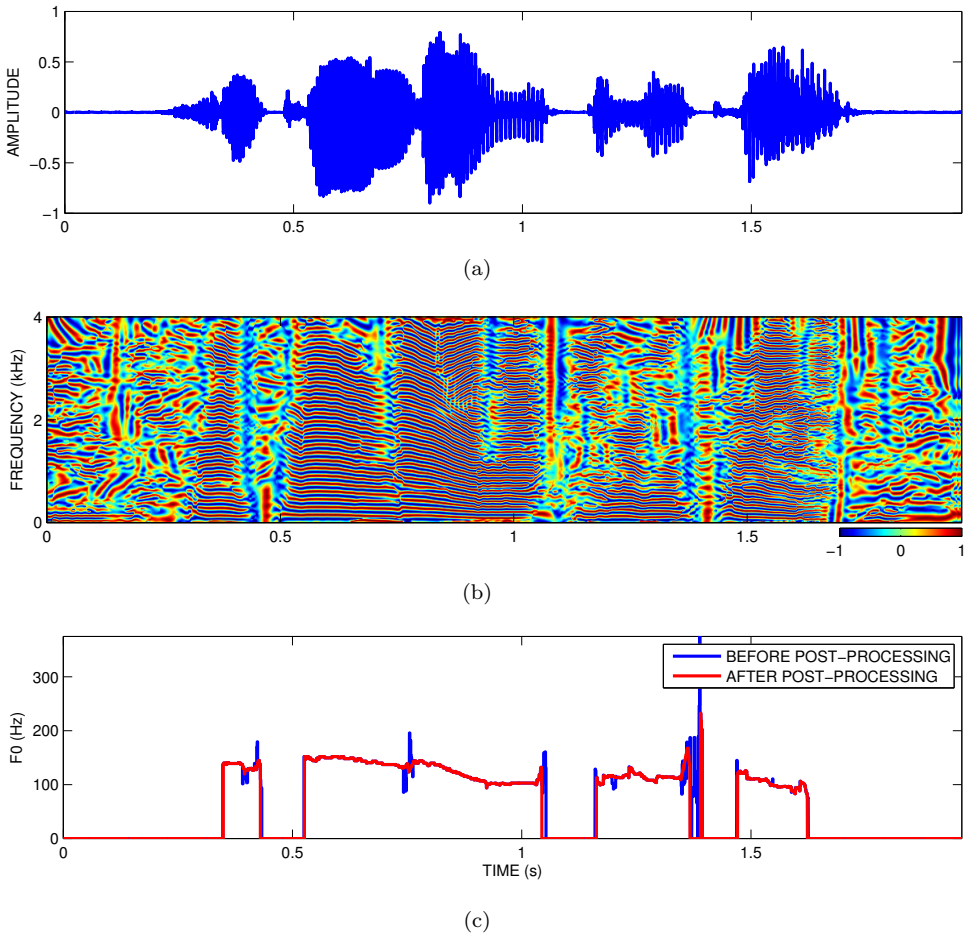


Figure 5.8: (a) A male speech utterance taken from CMU-ARCTIC database "arctic_a0036," (b) its carrier spectrogram, and (c) the estimated pitch using CRT-PP.

5.2.2 *Pitch Estimation Using CRT-CMNDF*

YIN computes CMNDF directly for a speech frame. In contrast, we compute it for a carrier-sinusoid from the WCS. Since the carrier is free from vocal-tract influence, the F0 estimates are likely to be more reliable. Figure 5.9 displays a Hamming-windowed carrier-sinusoid corresponding to a voiced speech frame and its CMNDF (computed using Equation (5.4)). The first dip in CMNDF occurs at F0 and subsequent ones at harmonics of F0. The first dip is a reliable estimator of F0. We use quadratic

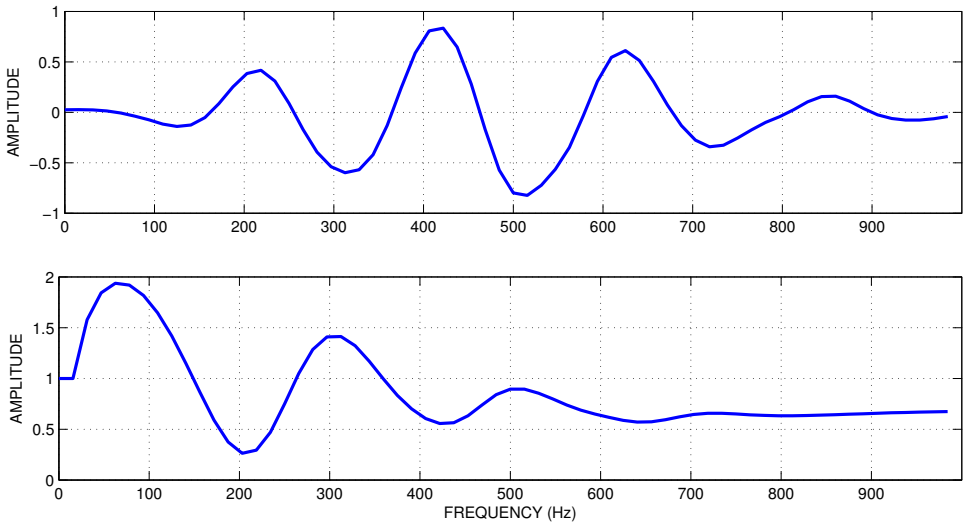
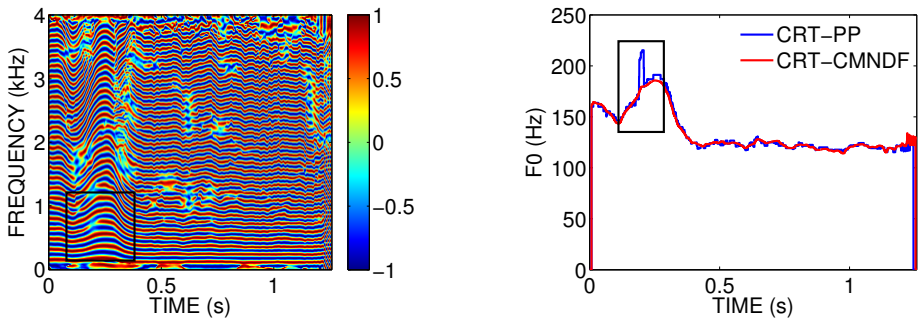


Figure 5.9: (a) A windowed carrier-sinusoid; and (b) its cumulative normalized difference function (CMNDF). The first dip in CMNDF occurs at F0 ($\sim$ 200 Hz in this example). The subsequent dips occur at harmonics of F0.

interpolation around the first dip to improve the accuracy. Repeating this process for all voiced speech segments gives an F0 contour. Unlike YIN and CRT-PP, CRT-CMNDF does not require a threshold. The post-processing step as in the case of CRT-PP is retained.

Between CRT-PP and CRT-CMNDF, the latter is more robust even without post-processing. To illustrate this point, consider the all-voiced speech utterance, "*Where were you while we were away?*". Figure 5.10 displays the carrier spectrogram and the F0 trajectories estimated using CRT-PP and CRT-CMNDF without post-processing. CRT-CMNDF gives a smoother trajectory than CRT-PP. The jump in F0 estimate in CRT-PP is caused by the loss of continuity of the harmonics as highlighted.



(a) (b)

Figure 5.10: (a) The carrier spectrogram of the all-voiced speech utterance "Where were you while we were away?" The black box shows the t-f region where the pitch harmonics have breaks. (b) The F0 trajectories estimated by CRT-PP and CRT-CMNDF without post-processing. This experiment illustrates that CRT-CMNDF gives a smoother estimate than CRT-PP.

5.3 Performance Evaluation

In this section, we compare the proposed F0 estimation algorithms against YIN, YAAPT, TEMPO, DIO and Harvest. We consider CMU-ARCTIC and CSTR-FDA databases for evaluation. From CMU-ARCTIC, we use data of three subjects: bdl (US male), slt (US female) and jmk (Canadian male). Both databases contain parallel EGG recordings. The ground-truth F0 trajectory and V/UV boundaries are derived from the EGG recordings employing the algorithm proposed in [118]. The speech signals are downsampled to 16 kHz. We compute a narrowband spectrogram with Hamming window of duration 40 ms and 1 ms frameshift. The DFT length was set to 1024-points.

For performance quantification, we use the **gross-pitch error (GPE)** and **fine-pitch error (FPE)**. Lower the GPE and FPE, better is the technique. GPE is the proportion of frames where the decisions for the algorithm and the ground-truth are both voiced, and for which the estimated F0 deviates from the reference by more

than 20%. The relative F0 error for the i^{th} speech frame is given by

$$\delta^{(i)} = \frac{|\hat{\text{F0}}^{(i)} - \text{F0}^{(i)}|}{\text{F0}^{(i)}} \times 100, \tag{5.7}$$

where $\hat{\text{F0}}^{(i)}$ and $\text{F0}^{(i)}$ denote the estimated F0 and the true F0, respectively. The GPE over N_v voiced frames is computed as

$$\text{GPE } (\%) = \frac{1}{N_v} \sum_{i=1}^{N_v} I_{(\delta^{(i)}>0.2)}, \tag{5.8}$$

where

$$I_{(A)} = \begin{cases} 1, \text{ if } A \text{ is true,} \\ 0, \text{ otherwise.} \end{cases} \tag{5.9}$$

FPE is computed as the standard deviation (in %) of the relative F0 error within 20% deviation from the reference F0.

Let

$$\Delta F^{(i)} = \delta^{(i)} I_{(\delta^{(i)}<0.2)}. \tag{5.10}$$

The FPE is given by

$$\text{FPE } (\%) = \sqrt{\frac{1}{N_v} \sum_{i=1}^{N_v} (\Delta F^{(i)} - \Delta \bar{F})^2}, \tag{5.11}$$

where $\Delta \bar{F}$ denotes the mean value of $\Delta F^{(i)}$ over the voiced frames.

5.4 Results

Tables 5.1 and 5.2 show the average scores over 200 clean speech files per speaker taken from the CMU-ARCTIC and CSTR-FDA databases, respectively. Although some existing algorithms make V/UV decisions as part of the method, to factor out the accuracy of V/UV decision-making, we have used the ground-truth V/UV decisions obtained from the parallel EGG recordings. We observe that CRT-CMNDF

is slightly superior than CRT-PP in terms of both GPE and FPE. These tables also show a comparison with the state-of-the-art pitch estimation algorithms. We conclude that the proposed pitch estimation methods (CRT-PP and CRT-CMNDF) are comparable to the state-of-the-art methods.

The performance in the presence of additive white Gaussian noise at 0 dB signal-to-noise ratio (SNR) is shown in Table 5.3 and Table 5.4 for CMU-ARCTIC and CSTR-FDA database, respectively. The proposed techniques are comparable with the state-of-the-art techniques. Between CRT-CMNDF and CRT-PP, the former is slightly superior.

Table 5.1: Objective evaluation of F0 estimation algorithms on CMU-ARCTIC database for clean speech.

Method →	CRT-PP	CRT-CMNDF	YIN	YAAPT	SRH	TEMPO	Harvest	DIO
	GPE (%) : Gross Pitch Error							
BDL	3.43 ± 1.95	2.29 ± 1.88	9.48 ± 3.81	1.33 ± 1.30	3.90 ± 2.60	$\mathbf{1.31 \pm 1.28}$	2.45 ± 1.80	3.31 ± 2.91
SLT	1.79 ± 1.90	1.63 ± 1.80	2.05 ± 2.86	$\mathbf{0.80 \pm 0.91}$	2.12 ± 2.38	0.98 ± 1.91	1.21 ± 1.84	1.34 ± 2.03
JMK	6.74 ± 2.62	6.24 ± 2.05	10.68 ± 2.45	$\mathbf{3.60 \pm 2.45}$	8.75 ± 3.93	4.81 ± 2.45	5.03 ± 2.67	5.44 ± 2.57
	FPE (%) : Fine Pitch Error							
BDL	3.03 ± 0.62	2.58 ± 0.46	$\mathbf{1.92 \pm 0.38}$	2.78 ± 0.48	3.38 ± 0.51	3.03 ± 0.42	3.28 ± 0.40	3.33 ± 0.46
SLT	2.40 ± 0.51	2.22 ± 0.49	$\mathbf{1.45 \pm 0.44}$	1.94 ± 0.40	2.77 ± 0.52	2.35 ± 0.45	2.37 ± 0.47	2.41 ± 0.45
JMK	3.87 ± 0.55	3.74 ± 0.59	$\mathbf{2.93 \pm 0.52}$	3.15 ± 0.47	3.99 ± 0.50	3.86 ± 0.50	3.85 ± 0.50	3.81 ± 0.47

Table 5.2: Objective evaluation of F0 estimation algorithms on CSTR-FDA database for clean speech.

Method →	CRT-PP	CRT-CMNDF	YIN	YAAPT	SRH	TEMPO	Harvest	DIO
	GPE (%) : Gross Pitch Error							
RL (female)	14.26 ± 3.15	$\mathbf{13.58 \pm 3.74}$	19.65 ± 5.94	15.40 ± 5.10	17.34 ± 6.18	14.98 ± 7.33	15.33 ± 5.52	15.44 ± 5.48
SB (male)	15.79 ± 5.05	7.89 ± 2.67	9.04 ± 3.54	6.14 ± 3.19	7.13 ± 2.12	7.79 ± 3.14	$\mathbf{6.22 \pm 2.82}$	6.80 ± 3.41
	FPE (%) : Fine Pitch Error							
RL (female)	4.39 ± 0.96	4.18 ± 1.04	$\mathbf{3.24 \pm 0.98}$	4.49 ± 0.89	5.29 ± 0.99	4.69 ± 0.91	4.92 ± 0.90	4.97 ± 0.89
SB (male)	9.39 ± 1.43	8.35 ± 1.64	$\mathbf{7.49 \pm 1.24}$	7.74 ± 0.94	8.81 ± 1.03	8.43 ± 1.03	8.62 ± 0.98	8.69 ± 1.02

Table 5.3: Objective evaluation of F0 estimation algorithms on CMU-ARCTIC database for noisy speech (SNR = 0 dB).

Method →	CRT-PP	CRT-CMNDF	YIN	YAAPT	SRH	TEMPO	Harvest	DIO
	GPE (%) : Gross Pitch Error							
BDL	13.97 ± 3.92	10.41 ± 2.09	46.95 ± 7.86	13.88 ± 6.13	$\mathbf{8.19 \pm 4.17}$	57.58 ± 14.17	10.19 ± 6.90	44.13 ± 14.82
SLT	24.74 ± 3.90	5.45 ± 3.69	24.49 ± 5.93	5.39 ± 4.06	5.02 ± 3.22	11.15 ± 7.81	6.56 ± 5.24	$\mathbf{3.70 \pm 4.19}$
JMK	24.33 ± 4.05	10.07 ± 3.45	43.06 ± 6.39	21.46 ± 6.03	12.56 ± 5.50	19.76 ± 8.15	9.76 ± 5.65	$\mathbf{9.60 \pm 4.31}$
	FPE (%) : Fine Pitch Error							
BDL	5.92 ± 0.53	4.56 ± 0.67	$\mathbf{1.42 \pm 0.37}$	2.81 ± 0.56	3.83 ± 0.65	3.41 ± 0.72	3.96 ± 0.49	5.36 ± 0.36
SLT	9.98 ± 0.83	6.65 ± 1.12	$\mathbf{2.82 \pm 0.80}$	3.89 ± 0.76	5.29 ± 0.97	4.90 ± 0.75	4.69 ± 0.84	6.16 ± 0.54
JMK	6.01 ± 0.47	4.66 ± 0.57	$\mathbf{2.16 \pm 0.80}$	3.30 ± 0.85	4.42 ± 0.63	4.21 ± 0.70	4.44 ± 0.67	4.60 ± 0.54

Table 5.4: Objective evaluation of pitch estimation algorithms on CSTR-FDA database for noisy speech (SNR 0 dB).

Method $\rightarrow$	CRT-PP	CRT-CMNDF	YIN	YAAPT	SRH	TEMPO	Harvest	DIO
	GPE (%) : Gross Pitch Error							
RL (female)	23.68 ± 7.29	22.81 ± 7.70	57.36 ± 9.34	$\mathbf{23.80 \pm 7.76}$	21.69 ± 7.19	59.67 ± 16.67	23.85 ± 9.54	46.79 ± 15.56
SB (male)	43.35 ± 14.41	14.41 ± 5.50	32.45 ± 7.57	12.25 ± 5.25	11.84 ± 4.18	31.68 ± 13.60	$\mathbf{28.39 \pm 10.80}$	31.12 ± 18.56
	FPE (%) : Fine Pitch Error							
RL (female)	6.49 ± 1.04	5.62 ± 1.12	$\mathbf{1.89 \pm 0.77}$	4.52 ± 1.07	4.53 ± 1.07	5.55 ± 1.52	4.95 ± 0.85	5.05 ± 0.73
SB (male)	12.58 ± 1.48	10.99 ± 1.51	$\mathbf{5.04 \pm 1.01}$	7.27 ± 0.95	8.92 ± 1.23	8.00 ± 1.23	8.21 ± 1.51	9.84 ± 1.04

5.5 Chapter Summary

We addressed the problem of pitch estimation from the carrier spectrogram. The F0 information in the carrier spectrogram was extracted by employing two different techniques, one based on peak-picking and the other based on cumulative normalized difference function of the carrier sinusoids. The latter approach was found to be more reliable even in the presence of noise. Performance comparisons with the state-of-the-art techniques carried out on the CMU-ARCTIC and CSTR-FDA databases, which have parallel EGG recordings that can be used to determine the ground-truth voiced/unvoiced decisions. The results showed that the proposed F0 estimation algorithms are on par with the state-of-the-art techniques for clean as well as noisy speech.

Chapter 6

Vocal-tract Filter Estimation and Speech Reconstruction

In the previous chapters, we focused on the analysis of the carrier spectrogram and addressed the problems of pitch estimation, demarcation of voiced/unvoiced regions, and delineation of periodic/aperiodic regions – all of these characterize the source or glottal excitation. The other important attribute in the source-filter model is the vocal-tract filter (VTF). In this chapter, we focus on the vocal-tract filter estimation. More precisely, we use 2-D AM as a spectrotemporal model for the magnitude response of the VTF. It shows the evolution of the vocal tract resonances (or formants). Formants are the most commonly used parameters in characterizing speech perception and intelligibility of voiced sounds.

In this chapter, we shall use the source and filter parameters together for the task of speech reconstruction without requiring the short-time phase of the speech waveform [44,45]. The window-length plays a crucial role in VTF estimation.

A fixed-length analysis window (typically 20 to 40 ms duration) is not optimal for estimating the VTF. The reason is as follows. Windowing in time is equivalent to convolution in frequency, which introduces interference between the fundamental and the harmonics (for voiced speech) and also smooths the spectral envelope. The purpose of the window is to obtain a time-slice of the signal during which the spectral characteristics are nearly constant. If the window is too long, it fails to capture the rapid variations of the spectrum. If it is too short, it smears the spectrum along the frequency dimension without commensurate improvement in time resolution. The

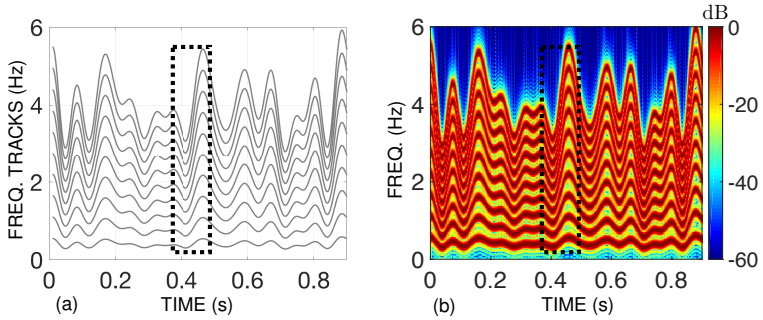


Figure 6.1: [Color online] (a) Harmonically related Instantaneous frequency (IF) tracks, and (b) spectrogram computed using a Hanning window of duration 30 ms. The regions of fast varying IF and the corresponding spectrogram region are indicated by using a rectangular box. The harmonic partials in the spectrogram are not clearly separable when IF changes rapidly. The frequency modulated synthetic signal is $s(t) = \sum_{k=1}^{10} \sin(2\pi k F_0 t + 0.25k \int_0^t \sum_{n=1}^{N} (a_n \sin(2\pi n\tau + \phi_n) + b_n \cos 2\pi n\tau)\ \mathrm{d}\tau)$, where $F0 = 400$ Hz, and a_i, b_i's are chosen from a uniform random distribution.

analysis window must be neither too short nor too wide. Ideally, it must be chosen based on the underlying frequency modulation [156]. Ideally, the window duration should be adapted to the rate of change of the spectrum. For a unit-modulus signal with phase $\phi(t)$, the optimal window duration Δ is inversely related to the rate of change of the instantaneous frequency (IF) [157]. Choosing $\Delta(t) = \sqrt{2}|\phi''(t)|^{-\frac{1}{2}}$ for a rectangular window reduces the spread of the ridge in the t-f plane. Consider a signal composed of ten harmonically related sinusoids, each with instantaneous frequency (IF) shown in Figure 6.1(a). Although the energy is concentrated along the harmonics, we observe smearing of the spectrum in t-f regions where the IF varies rapidly as shown in Figure 6.1(b) (see dashed box). This is because rapid IF variations reduce the localization accuracy. As a result, the instantaneous bandwidth of the signal is also overestimated [158]. These problems can be circumvented by adapting the duration of the analysis window to the pitch period of the speaker. STRAIGHT and WORLD vocoders have used this analysis methodology for estimating the VTF. Similarly, we adapt the duration of analysis window in proportion to the

inverse of F0 as the window slides with a constant frameshift over the duration of signal. Such a pitch-adaptive spectrogram trades-off the time-frequency resolution better than a fixed-window-length spectrogram. Further, short-time analysis results in wider formant bandwidths, which makes it unsuitable for high-quality speech reconstruction. To overcome this limitation, we propose a method for formant bandwidth correction. The formant bandwidth correction is applied to the 2-D AM obtained by employing CRT-based demodulation on the pitch-adaptive spectrogram. The curated 2-D AM is used for speech reconstruction along with CRT-based source parameters. The effectiveness of the proposed approach is shown in the context of (1) source-filter-based spectral synthesis model; and (2) a neural vocoder (WaveNet).

6.1 The Pitch-adaptive Spectrogram

In contrast to the fixed-window-length spectrograms (narrowband/wideband), a pitch-adaptive spectrogram is computed using a window whose length is varied in proportion to the fundamental frequency of the speaker.

The pitch-adaptive STFT of a voiced speech signal $s(\tau)$ at time t is given by

$$S(t,\omega) = \int_{-\infty}^{\infty} s(\tau) w_\mu(\tau - t) e^{-\mathrm{j}\omega\tau} \mathrm{d}\tau, \tag{6.1}$$

where the variable-length window $w_\mu(\tau)$ is supported over $\tau \in [-\mu T_0, \mu T_0]$, $T_0 = \frac{1}{\mathrm{F0}}$ being the fundamental period at time t, and μ is a parameter (typically, μ takes values $1, 1.5, 2, 2.5, 3, 3.5$ without violating the stationary assumption. The window duration is given by $\frac{2\mu}{\mathrm{F0}}$ (in seconds). The pitch-adaptive spectrogram is given by $S(\boldsymbol{\omega}) = |S(t,\omega)|^2$. Table 6.1 lists the window duration for various values of μ, considering that an adult speaker's pitch normally falls in the range $[100, 155]$ Hz for males, and $[165, 250]$ Hz for females. The window length varies between 8 ms and 70 ms, which corresponds to the highest and lowest pitch, respectively. We choose the value of μ that maximizes the demodulation accuracy, measured in terms of the

Table 6.1: Window duration (in milliseconds) as a function of μ

	Male		Female	
μ	F0 = 100 Hz	F0 = 155 Hz	F0 = 165 Hz	F0 = 250 Hz
1	20	12.9	12.1	8
1.5	30	19.4	18.2	12
2	40	26	24	16
2.5	50	32	30	20
3	60	39	36	24
3.5	70	45	42	28

Global Signal-to-Reconstruction Error Ratio (GSRER) given by

$$\text{GSRER} = 20 \log \frac{\|S(\boldsymbol{\omega})\|}{\|\tilde{S}(\boldsymbol{\omega}) - S(\boldsymbol{\omega})\|}, \quad \text{(dB)} \tag{6.2}$$

where $S(\boldsymbol{\omega})$ and $\tilde{S}(\boldsymbol{\omega})$ represent the original and reconstructed pitch-adaptive spectrogram, respectively.

6.1.1 *Choice of μ*

The optimum value of μ is the one that maximizes the GSRER. We consider a mono-component AM-FM model for the pitch-adaptive spectrogram and its reconstruction from the estimated AM and FM. We randomly selected 40 speech files from the CMU-ARCTIC database: 20 male (bdl) and 20 female (slt). The pitch-adaptive spectrogram is divided into t-f patches of dimension 100 ms $\times$ 600 Hz where the frequency dimension is chosen to include at least 2 pitch harmonics even for high pitch sounds. The overlap is 75% and 35% along time and frequency axes, respectively. The synthesized spectrogram is given by

$$\tilde{S}(\boldsymbol{\omega}) = \tilde{V}(\boldsymbol{\omega})(\alpha_0 + \cos \tilde{\Phi}(\boldsymbol{\omega})), \tag{6.3}$$

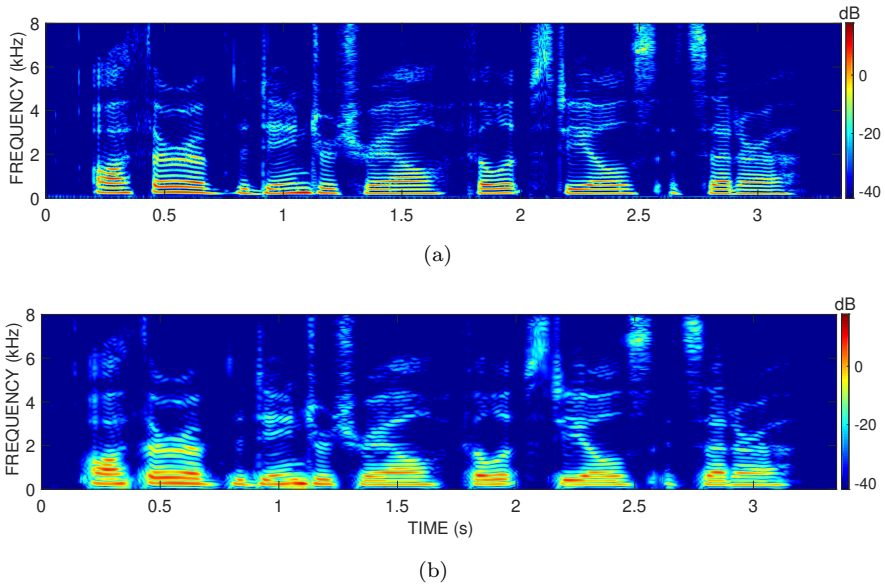


Figure 6.2: (a) A pitch-adaptive spectrogram; and (b) the reconstructed spectrogram from the estimated AM and FM using the monocomponent model. The spectrogram corresponds to the utterance "*Author of the danger trail, Philip Steels, etc.*" spoken by a female speaker taken from the CMU-ARCTIC database.

where $\tilde{V}(\boldsymbol{\omega})$ and $\tilde{\Phi}(\boldsymbol{\omega})$ are the estimated AM and FM, respectively. The value of α_0 in Equation (6.3) is obtained by solving the following least-squares regression:

$$\arg\min_{\alpha_0} \|S(\boldsymbol{\omega}) - \tilde{V}(\boldsymbol{\omega})(\alpha_0 + \cos\tilde{\Phi}(\boldsymbol{\omega}))\|_2^2. \tag{6.4}$$

Figure 6.2(a) and Figure 6.2(b) display a pitch-adaptive spectrogram and its reconstruction, respectively. Table 6.2 displays the average GSRER values for male and female speech utterances. The highest GSRER is obtained for $\mu = 2.5$, i.e., a window length of 5T0. Based on these results, the value of μ is fixed to 2.5 for computation of the pitch-adaptive spectrogram.

Table 6.2: Average GSRER (dB) for various values of μ corresponding to the speech utterances taken from CMU-ARCTIC database.

$\mu \rightarrow$	1	1.5	2	2.5	3	3.5
Male	14.96 ± 1.73	6.58 ± 0.74	13.54 ± 1.2	$\mathbf{15.7 \pm 1.61}$	9.91 ± 0.7	11.46 ± 0.91
Female	7.01 ± 1.4	5.98 ± 0.53	9.31 ± 1.46	$\mathbf{9.41 \pm 1.78}$	7.71 ± 0.73	8.44 ± 1.08

6.2 Formant Bandwidth Correction

Short-time analysis broadens the formants and the 2-D bandpass filter further affects the smoothness and formant bandwidths in the estimated spectral envelope. Larger formant bandwidths cause a synthesized speech signal to decay faster in every glottal cycle and makes it sound buzzy. On the other hand, narrow formants result in slow decay and result in tone-like perception. Hence, it is important to correct for formant bandwidth estimation errors.

Small fluctuations in the estimated envelope are due to residual harmonic interference or noise. They can be suppressed by applying a local smoother on the estimated AM:

$$V_{s,t_i}(\omega) = \frac{2}{\omega_0(t_i)} \int_{\omega-\omega_0(t_i)/4}^{\omega+\omega_0(t_i)/4} V_{t_i}(\lambda)\mathrm{d}\lambda \tag{6.5}$$

where $\omega_0(t_i)$ (radian/sec) is the fundamental frequency at time t_i. The smoothing considers a rectangular window of width $\omega_0(t_i)/2$. The width parameter was chosen experimentally.

The proposed formant bandwidth correction is based on the weighted central difference operator, applied to the smoothed AM on the log scale. The logarithm compresses the dynamic range and makes the correction more effective in high-frequency regions. A three-point weighted central difference of $V_{s,t_i}(\omega)$ is given by

$$X_{t_i}(\omega) = \sum_{k=-1}^{1} w_k \ln V_{s,t_i}(\omega - k\omega_0(t_i)), \tag{6.6}$$

where the weights sum to unity:

$$\sum_{k=-1}^{1} w_k = 1. \tag{6.7}$$

Equation (6.6) shows that $X_{t_i}(\omega)$ is the weighted sum of three terms: $\ln V_{s,t_i}(\omega)$, its right-shifted version $\ln V_{s,t_i}(\omega-\omega_0(t_i))$, and the left-shifted version $\ln V_{s,t_i}(\omega+\omega_0(t_i))$. We choose equal weights for the right-shifted and left-shifted versions, i.e., $w_{-1} = w_1$. Taking inverse Fourier transform on both sides of Equation (6.6) gives

$$\hat{X}_{t_i}(\tilde{t}) = \Big(w_0 + 2w_1 \cos\big(\omega_0(t_i)\tilde{t}\big)\Big)\hat{L}_{t_i}(\tilde{t}), \tag{6.8}$$

where $\tilde{t}$ is the dual of ω, which denotes the quefrency variable in the cepstral domain, and $\hat{L}_{t_i}(\tilde{t})$ is the inverse Fourier transform of $\ln V_{s,t_i}(\omega)$. The corrected spectral envelope is further subjected to lowpass filtering to suppress variations beyond F0 resulting in the following envelope:

$$V_{c,t_i}(\omega) = e^{\mathcal{F}\{\hat{X}_{t_i}(\tilde{t})\hat{S}_{t_i}(\tilde{t})\}}, \tag{6.9}$$

where

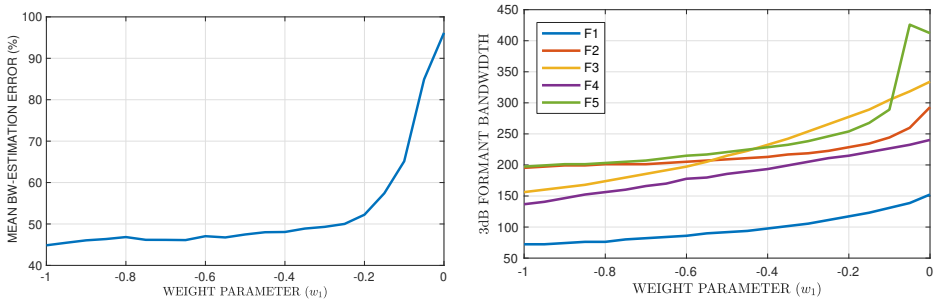
$$\hat{S}_{t_i}(\tilde{t}) = \frac{\sin(\pi \mathrm{F0}\tilde{t})}{\pi \mathrm{F0}\tilde{t}}.$$

A negative value of w_1 in Equation (6.6) ensures reduction in the formant bandwidths. The correction filter is parameterized by a single parameter w_1 since $w_0 = 1 - 2w_1$. We evaluate the effectiveness of the formant bandwidth correction for a synthetic vowel first and then for the real speech from VTR database [60].

In order to synthesize a vowel close to a natural one, we extract linear prediction-based envelope from a sustained real vowel "/o/" of duration 1 sec, which is spoken by a male speaker having an average pitch of 120 Hz — the average pitch is estimated by using an open source software *Praat* [84]. The linear prediction envelope from a real vowel is obtained by using short-time analysis with a Hamming window of duration 40 ms and frameshift of 5 ms. The vowel is synthesized using a 14^{th}-order

Table 6.3: The estimated 3-dB formant bandwidths (Hz) and the bandwidth estimation error (Hz) with respect to the ground-truth for a synthetic vowel. $\bar{\delta} = \frac{1}{5}\sum_{i=1}^{5}\delta^{(i)}$. Parameter $w_1 = -0.50$ was chosen based on the objective evaluation of the reconstructed speech waveforms (Section 6.4).

	$B^{(1)}\|\delta^{(1)}$	$B^{(2)}\|\delta^{(2)}$	$B^{(3)}\|\delta^{(3)}$	$B^{(4)}\|\delta^{(4)}$	$B^{(5)}\|\delta^{(5)}$	$\bar{\delta}$
Ground truth	113|–	86|–	279|–	125|–	213|–	–
After smoothing	152|34.48	293|240.91	334|19.58	240|92.19	412|93.58	96.15%
After correction ($w_1 = -0.50$)	90|20.69	207|140.91	205|26.57	180|43.75	217|1.83	46.75%
WORLD	109|3.45	225|161.36	262|6.29	199|59.38	248|16.51	49.40%



(a) (b)

Figure 6.3: (Color online) Influence of the weight parameter w_1 on the bandwidth estimation error: (a) Mean-bandwidth estimation error; and (b) the estimated formant bandwidths.

LP filter with the excitation

$$e(t) = \sum_{k=1}^{K_0} \cos(2\pi k\,\mathrm{F0}\,t), \tag{6.10}$$

where F0 = 120 Hz, $K_0 = \frac{F_s/2}{\mathrm{F0}} \approx 33$ (number of harmonics) with $F_s = 8000$ Hz. The envelope is obtained using CRT operating on the pitch-adaptive spectrogram. The envelope is subjected to formant bandwidth correction.

In order to examine the effect of the weight parameter w_1 on the formant bandwidths, we estimate the 3-dB formant bandwidths from the smoothed spectral envelope and the LP envelope that serves as the ground truth.

6.2.1 *Effect of weight parameter on bandwidth estimation*

We estimate 3-dB formant bandwidths from: (1) the ground truth LP envelope, (2) smoothed CRT envelope, (3) CRT envelope after correction, and (4) envelope obtained from WORLD vocoder. The relative estimation error for the i^{th} formant bandwidth is given by

$$\delta^{(i)} = \frac{\left|B^{(i)} - B_g^{(i)}\right|}{B_g^{(i)}} \times 100, \tag{6.11}$$

where $B^{(i)}$ and $B_g^{(i)}$ represent the estimated 3-dB bandwidth (Hz) and the ground truth, respectively. Table 6.3 compares the estimated bandwidths and the relative estimation error for five formants. We observe that the bandwidth correction filter reduces the average bandwidth estimation error from 96.15 % to 46.75 %. Figure 6.3(a) displays the mean bandwidth estimation error $\bar{\delta}$ with respect to weight parameter w_1. The error is small when $-1 < w_1 < -0.65$.

While Figure 6.3(a) displays an aggregated effect of w_1 on all the five formants, Figure 6.3(b) shows the effect of w_1 on the estimated bandwidths of the individual formants. Figure 6.4 displays the CRT-AM, and its corrected version ($w_1 = -0.50$) for a windowed segment of the synthetic vowel along with the ground truth LP envelope, WORLD envelope, and the STFT magnitude response. Figure 6.5 shows the corrected envelope for various choices of w_1. Large negative values of w_1 result in envelope fluctuations, which are undesirable. The objective evaluation in Section 6.3 confirms that the best quality of reconstructed speech (using a spectral synthesis model) is achieved for $-0.60 < w_1 < -0.50$.

6.2.2 *Effect of weight parameter on formant frequencies*

Ideally, the bandwidth correction step must not shift the formants. In this subsection, we check if there is a shift in the formant frequencies. The estimation error for the

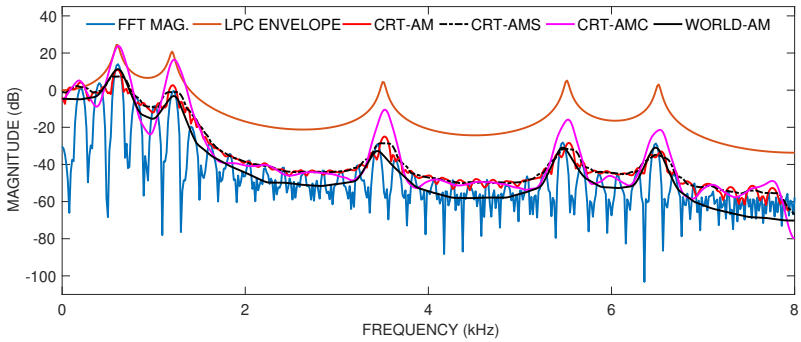


Figure 6.4: (Color online) This figure shows the STFT magnitude, envelope estimated using CRT (CRT-AM), envelope obtained after smoothing (CRT-AMS), envelope obtained after correction (CRT-AMC), and envelope obtained from the WORLD vocoder (WORLD-AM). The envelopes are shifted by introducing a bias only to aid visualization.

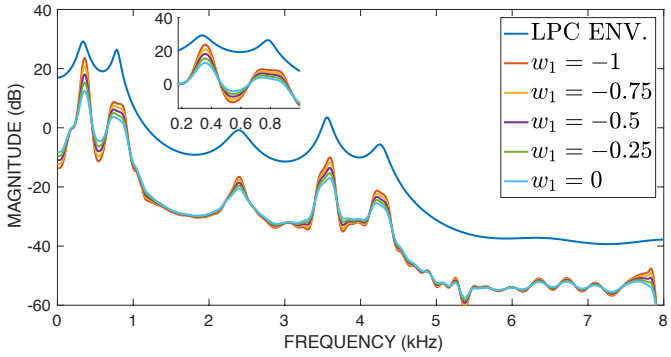


Figure 6.5: [Color online] The effect of weight parameter w_1 on the envelope estimated by the CRT demodulator. The formant bandwidths reduce as the weight parameter becomes more negative – this effect can be seen predominantly in the first formant shown in the zoomed-in portion.

i^{th} formant is given by

$$\gamma^{(i)} = \frac{|F^{(i)} - F_g^{(i)}|}{F_g^{(i)}} \times 100 \qquad (\%), \tag{6.12}$$

where $F^{(i)}$ and $F_g^{(i)}$ denote the i^{th} formant and the ground-truth formant frequencies, respectively. Table 6.4 shows the formant estimation error before and after applying bandwidth correction to smoothed CRT-AM. We observe that the correction

Table 6.4: Formants (Hz) and the formant estimation error (in %) after correction for a synthetic vowel. $\bar{\gamma} = \frac{1}{5}\sum_{i=1}^{5}\gamma^{(i)}$. Parameter $w_1 = -0.50$ was chosen based on the objective evaluation of the reconstructed speech waveforms (Section 6.4).

	$F^{(1)}\|\gamma^{(1)}$	$F^{(2)}\|\gamma^{(2)}$	$F^{(3)}\|\gamma^{(3)}$	$F^{(4)}\|\gamma^{(4)})$	$\tilde{F}^{(5)}\|\gamma^{(5)}$	$\bar{\gamma}$
Ground truth	343.83|–	781.44|–	2391.21|–	3561.42|–	4251.04|–	–
Smoothed CRT-AM	359.46|4.55	734.55|6	2391.21|0	3577.05|0.44	4204.15|1.1	2.42%
CRT-AM after correction ($w_1 = -0.50$)	359.46|4.55	734.55|6	2391.21|0	3592.67|0.88	4217.83|0.78	2.44%

mechanism causes only a minor perturbation in the formant frequencies.

For real speech, Figure 6.6 displays the envelope in the t-f plane estimated using the proposed technique and the CheapTrick [159] algorithm (used by WORLD vocoder) together with the corresponding pitch-adaptive spectrogram.

6.2.3 *Effect of formant bandwidth correction on real speech*

We analyze the effectiveness of the proposed CRT-based formant tracking and the bandwidth correction method on VTR database [60]. The database has 8 dialects of which we select a subset consisting of 2 female and 2 male utterances for each dialect, ending up with a total of 32 utterances for evaluation. The formants and their bandwidths are compared for STRAIGHT, WORLD and CRT envelope with respect to each other and in reference to the values provided in the VTR database. The standard objective measures for evaluation are as follows.

Gross Detection Rate (GDR) (in %): It is the number of formants detected within 20 % of the reference value or 300 Hz absolute deviation, whichever is smaller.

Mean Absolute Deviation (MAD) (in Hz): It is the mean value of the absolute deviation of the detected formants from the reference value.

A high value of GDR and a low value of MAD indicate better overall accuracy of a formant tracking algorithm. The envelopes from STRAIGHT, WORLD and CRT-AM (before formant bandwidth correction) are obtained for the 32 speech waveforms.

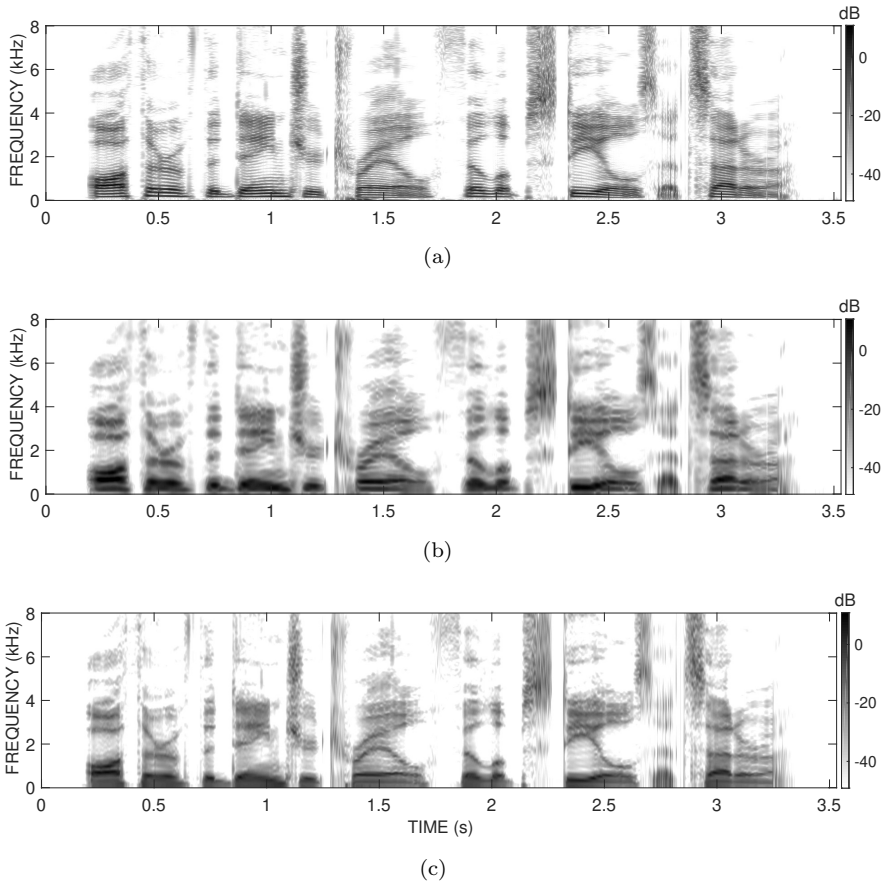


Figure 6.6: (a) A pitch-adaptive spectrogram; (b) CRT envelope after formant bandwidth correction; and (c) envelope obtained using CheapTrick (algorithm used by WORLD) for the speech utterance, "*Author of the danger trial, Philip Steels etc.*" spoken by a male speaker.

The GDR and MAD values are averaged across all the voiced speech frames and over the chosen utterances. The performance of three methods for formant tracking is reported using GDR and MAD. The estimation errors in formant bandwidths can only be reported using MAD. Table 6.5 and Table 6.6 report the average GDR and MAD scores for formant tracking for male and female speakers, respectively. The results show that the CRT-based approach is on par with STRAIGHT and WORLD. Figure 6.7 shows the estimated formant tracks superimposed on the ground-truth

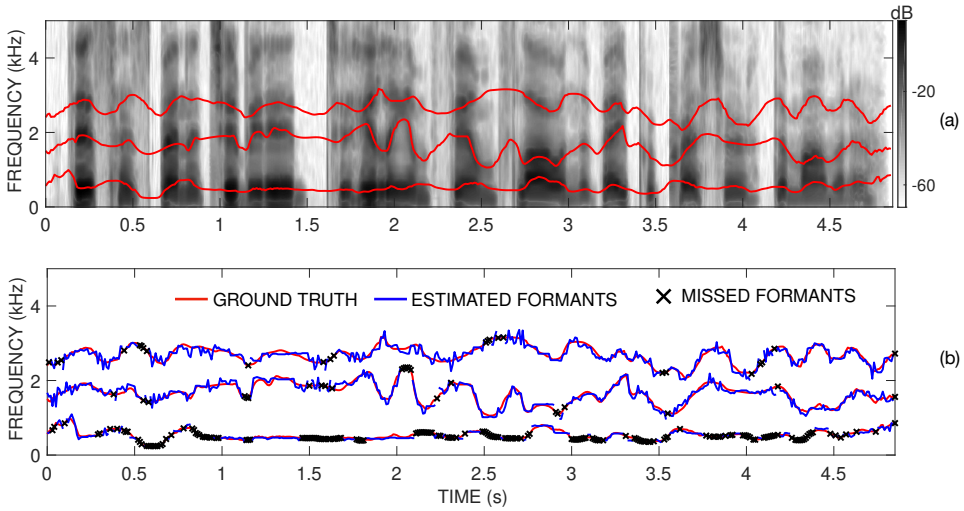


Figure 6.7: Illustration of the first three formant tracks from the VTR database and the estimated formant tracks using CRT. (a) Smoothed CRT envelope overlaid with the ground-truth formants; and (b) ground-truth formant tracks, their estimates, and missed formants. The speech utterance is "*This has been attributed to helium film flow in the vapor pressure thermometer.*" spoken by a female speaker, taken from the VTR database.

Table 6.5: Average GDR and MAD of formants F1, F2, and F3 for male speakers. Parameter $w_1 = -0.50$ was chosen for bandwidth correction based on objective evaluation of reconstructed speech (Section 6.4).

	GDR (%)			MAD (Hz)		
Method	F1	F2	F3	F1	F2	F3
CRT (before correction)	78.07	93.00	93.09	44.92	77.08	74.03
CRT (after correction)	76.88	**97.53**	98.04	46.25	**70.56**	72.58
STRAIGHT	**81.16**	96.27	99.37	45.03	78.95	**65.53**
WORLD	79.09	96.48	**99.38**	**44.76**	83.30	70.86

formant tracks provided in the VTR database. The VTR ground-truth is based on LP analysis followed by manual correction. The figure illustrates that the CRT envelope closely follows the ground truth formants.

Table 6.7 and Table 6.8 report the average values of MAD scores for formant bandwidths for male and female speakers, respectively, where we used the smoothed

Table 6.6: Average GDR and MAD of formants F1, F2, and F3 for female speakers. Parameter $w_1 = -0.50$ was chosen for bandwidth correction based on objective evaluation of reconstructed speech (Section 6.4).

	GDR (%)			MAD (Hz)		
Method	F1	F2	F3	F1	F2	F3
CRT (before correction)	79.01	81.27	81.41	**45.11**	92.06	**88.05**
CRT (after correction)	75.72	90.38	93.28	50.03	**89.97**	92.64
STRAIGHT	**80.73**	**96.76**	**94.47**	46.51	93.13	94.48
WORLD	76.53	94.08	92.56	45.06	93.68	97.76

Table 6.7: Average MAD of formant bandwidths (female speakers). Parameter $w_1 = -0.50$ was chosen for bandwidth correction based on objective evaluation of reconstructed speech (Section 6.4).

	MAD (Hz)		
Method	F1-BW	F2-BW	F3-BW
CRT (before correction)	162.61	157.23	721.33
CRT (after correction)	82.45	110.00	**208.00**
STRAIGHT	130.74	**101.78**	492.12
WORLD	**63.11**	122.35	556.31

Table 6.8: Average MAD for formant bandwidths (male speakers). Parameter $w_1 = -0.50$ was chosen for bandwidth correction based on objective evaluation of reconstructed speech (Section 6.4).

	MAD (Hz)		
Method	F1-BW	F2-BW	F3-BW
CRT (before correction)	104.00	162.00	483.95
CRT (after correction)	97.15	144.00	388.00
STRAIGHT	51.13	105.07	300.74
WORLD	**44.76**	**83.30**	**70.86**

CRT envelope without formant bandwidth correction. From the tables, it is clear that CRT overestimates the formant bandwidths in comparison to STRAIGHT and WORLD. Also, the factor by which the bandwidths are overestimated is more for

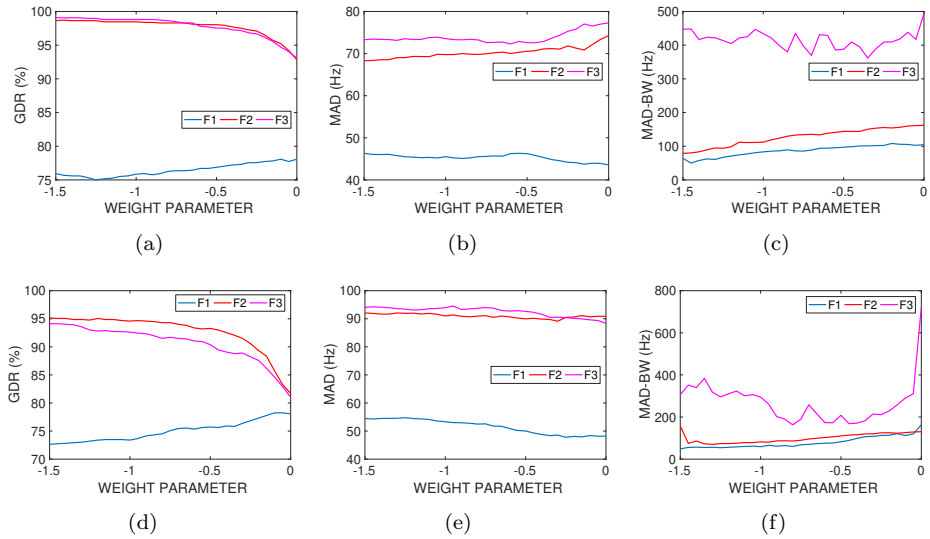


Figure 6.8: Average GDR and MAD for formants, and MAD scores for formant bandwidths (MAD-BW) with respect to the weight parameter used for formant bandwidth reduction. (First row: male speakers; Second row: female speakers).

female speakers than the male speakers, quite likely due to higher pitch of the female speakers.

Next, we compute the average GDR and MAD for various values of the weight parameter in Equation (6.6). From the results shown in Figure 6.8, we observe that average GDR exhibits an increasing trend for second and third formants, and a decreasing trend for the first formant for both the genders (see Figure 6.8(a) and Figure 6.8(d)). The average MAD scores for formant bandwidths exhibit a decreasing trend for the first two formants. However, the trend is a bit erratic for the third formant-bandwidth (see Figure 6.8(c) and Figure 6.8(f)). These results on real speech signals underscore the importance of formant bandwidth correction.

The Riesz transform based analysis of speech signals is summarized in the block diagram shown in Figure 6.9. The source parameters such as pitch, voiced/unvoiced segmentation, and aperiodicity are all estimated by employing CRT demodulation on a narrowband spectrogram. In particular, the 2-D FM is used for the estimation

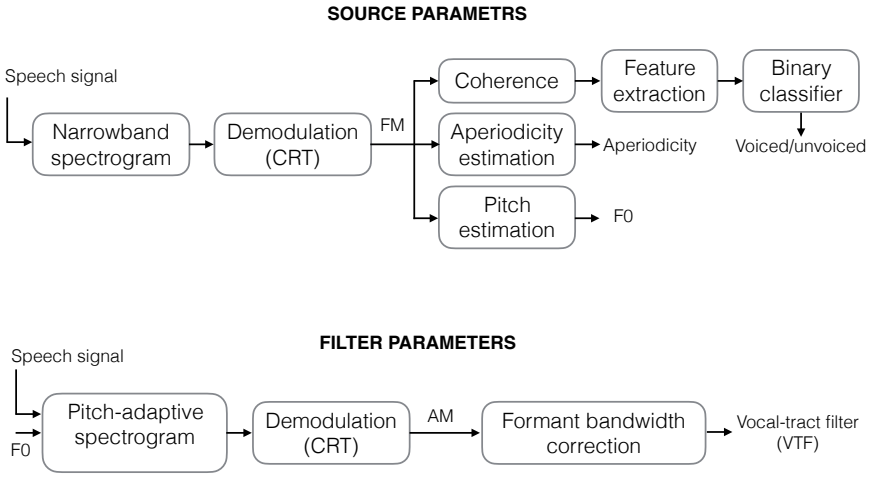


Figure 6.9: Block diagrams illustrating source and filter parameter estimation using CRT-based analysis of a speech signal.

of source parameters. The vocal-tract filter response is obtained by employing CRT-based demodulation of the pitch-adaptive spectrograms.

The rest of this chapter is devoted to speech reconstruction using the source and filter parameters estimated using the CRT approach. For comparison, we use acoustic features obtained from the state-of-the-art vocoders STRAIGHT and WORLD.

6.3 Speech Reconstruction Using the Spectral Synthesis Model

The spectral synthesis model requires four inputs: instantaneous F0, voiced/unvoiced decisions, aperiodicity parameters (AP), and the vocal-tract filter (VTF) response. In the spectral synthesis model, the spectrum of a voiced sound segment at instant t_i is given by

$$\hat{s}_{v,t_i}(\omega) = \hat{v}_{t_i}(\omega)\left(\sqrt{T_0}(1-\hat{a}_{t_i}(\omega))\hat{p}_{t_i}(\omega) + \hat{a}_{t_i}(\omega)\hat{n}_{t_i}(\omega)\right). \tag{6.13}$$

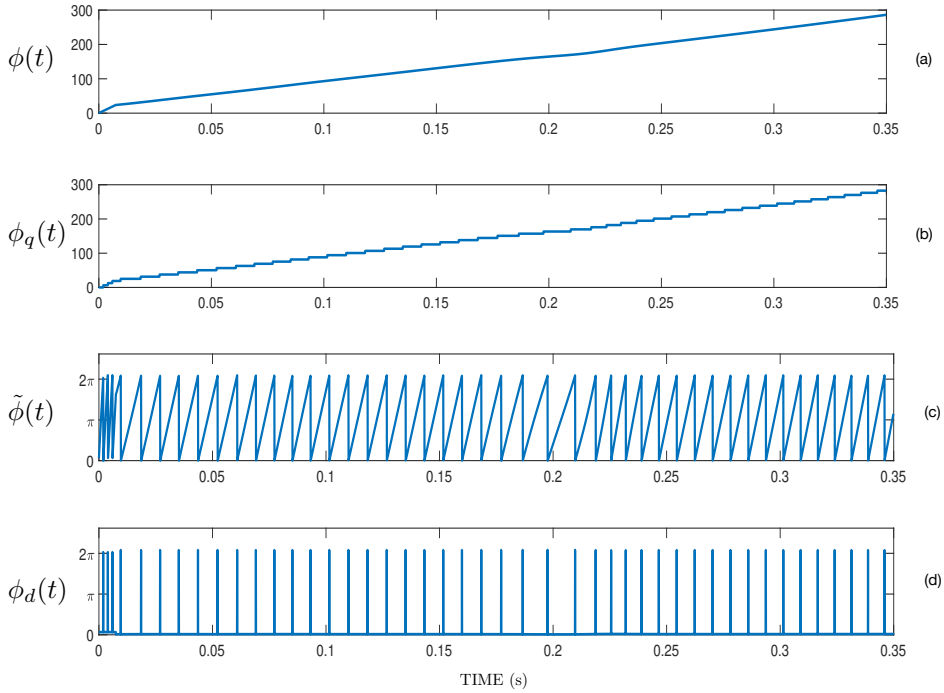


Figure 6.10: Illustration for obtaining the synthesis time instants from instantaneous phase. Synthesis time instants are given by the time-locations for which $\phi_d(t) > \pi$.

For unvoiced segments, the spectrum is given by

$$\hat{s}_{uv,t_i}(\omega) = \hat{v}_{t_i}(\omega)\hat{a}_{t_i}(\omega)\hat{n}_{t_i}(\omega), \tag{6.14}$$

where $\hat{v}_{t_i}(\omega)$, $\hat{a}_{t_i}(\omega)$, $\hat{n}_{t_i}(\omega)$, and T_0 denote the vocal-tract filter estimate, aperiodicity map, white Gaussian noise, the instantaneous pitch period at t_i. The quantities $\hat{v}_{t_i}(\omega)$ and $\hat{a}_{t_i}(\omega)$ are derived by a minimum-phase approximation of the estimated vocal-tract filter and aperiodicity parameters, respectively. At a given instant, the speech segment is identified either as voiced or unvoiced, the corresponding spectrum is computed, and then subjected to inverse Fourier transform. The resulting speech segments are overlapped and added in a pitch-synchronous fashion.

The synthesis instants are derived from F0 track of the speaker. In general, the analysis and synthesis time instants are not aligned. The filter response, AP, V/UV and F0 are copied from the nearest analysis instant to the current synthesis instant.

Table 6.9: Various configurations considered to assess the influence of the analysis parameters on the quality of speech reconstruction using the spectral synthesis model.

Parameters	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
V/UV	WORLD	CRT	CRT	CRT	CRT	CRT
AP	D4C [127]	D4C	CRT	CRT	CRT	CRT
F0	Harvest [150]	Harvest	Harvest	CRT	Harvest	CRT
VTF	CheapTrick [159]	CheapTrick	CheapTrick	CheapTrick	CRT	CRT

6.3.1 *Synthesis Time Instants*

The synthesis instants t_i in Equation (6.13) and Equation (6.14) are derived from the pitch contour $f_0(t)$ as follows. The procedure is identical to that followed in STRAIGHT and WORLD.

(1) Compute the instantaneous phase $\phi(t)$ from the instantaneous pitch $f_0(t)$ (cf. Figure 6.10(a)):

$$\phi(t) = 2\pi \int_0^t f_0(\tau) \mathrm{d}\tau. \tag{6.15}$$

A default value of 500 Hz is assigned to $f_0(t)$ in unvoiced segments.

(2) Construct a function $\phi_q(t) = 2\pi \left\lfloor \frac{\phi(t)}{2\pi} \right\rfloor$, which is discontinuous at the level-crossings of 2π and its integer multiples (cf. Figure 6.10(b)).

(3) Remove the contribution of $\phi_q(t)$ from $\phi(t)$ to obtain $\tilde{\phi}(t) = \phi(t) - \phi_q(t)$. The function $\tilde{\phi}(t)$ is bounded between 0 and 2π and retains the discontinuities in $\phi_q(t)$ (cf. Figure 6.10(c)).

(4) Highlight the singularities in $\tilde{\phi}(t)$ by computing its derivative $\phi_d(t) = \frac{\mathrm{d}\tilde{\phi}(\mathrm{t})}{\mathrm{d}t}$ (cf. Figure 6.10(d)).

(5) The synthesis instants are selected as the instants $\{t_i\}$ where $\phi_d(t) > \pi$.

Table 6.10: Average PESQ scores over 100 speech utterances for each of the speakers taken from CMU-ARCTIC database.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
bdl	3.47 ± 0.14	3.42 ± 0.16	3.34 ± 0.12	3.31 ± 0.16	2.93 ± 0.14	2.67 ± 0.15
ksp	3.42 ± 0.16	3.45 ± 0.13	3.42 ± 0.12	3.40 ± 0.16	2.73 ± 0.14	2.48 ± 0.11
clb	3.49 ± 0.12	3.40 ± 0.14	3.34 ± 0.15	3.45 ± 0.18	2.64 ± 0.14	2.30 ± 0.13
slt	3.58 ± 0.02	3.51 ± 0.15	3.43 ± 0.12	3.34 ± 0.16	2.91 ± 0.15	2.58 ± 0.18

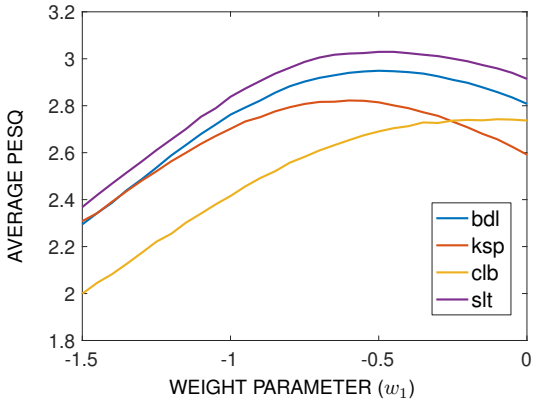


Figure 6.11: (a) PESQ scores versus bandwidth correction factor w_1, averaged over 100 speech waveforms for each speaker.

6.4 Results

Next, we analyze the effect of each of the parameters on the quality of reconstruction. The various scenarios considered are presented in Table 6.9. The reconstruction follows the spectral synthesis model in each case. Case 1 corresponds to the baseline WORLD vocoder, which uses D4C for aperiodicity estimation [127], Harvest for F0 estimation [150], and CheapTrick for vocal-tract filter estimation [159]. In Case 2, the AP used by WORLD is replaced by CRT aperiodicity (CRT-AP), and likewise for the other cases mentioned in the table.

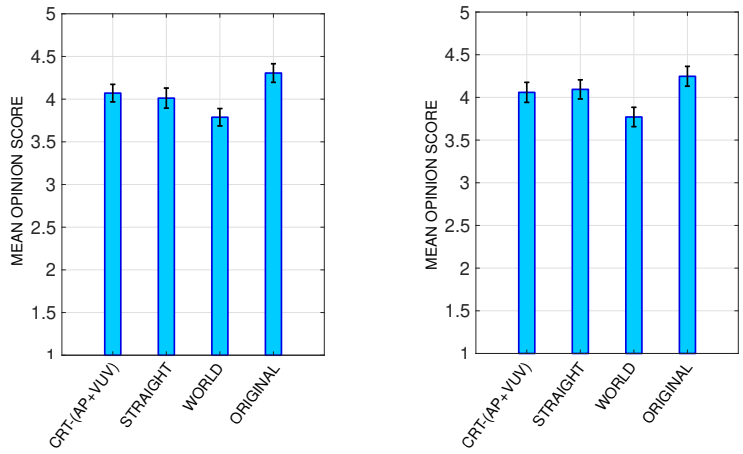


Figure 6.12: Mean opinion scores for the analysis/synthesis experiment: (a) male speakers (bdl, ksp); and (b) female speakers (clb, slt) taken from CMU-ARCTIC database. This figure compares the influence of CRT aperiodicity parameters and voiced/unvoiced decisions (CRT-(AP+V/UV)), WORLD, and STRAIGHT. The error bars show 95% confidence interval.

6.4.1 *Objective Evaluation*

For objective evaluation, we use the PESQ metric. We choose a total of 100 speech waveforms from CMU-ARCTIC database for each of the speakers: 2 male (bdl, ksp) and 2 female (slt, clb). Table 6.10 reports the average PESQ scores and the standard deviation. The scores show that the proposed parameters result in a performance comparable with the baseline except for Case 5 and Case 6, where the CRT envelope is used after formant bandwidth correction. The effect of the weight parameter w_1 on PESQ is determined by reconstructing speech signals using the configuration in Case 5. w_1 is varied from -1.5 to 0, with $w_1 = 0$ corresponding to no correction. Figure 6.11 displays the average PESQ over 100 speech waveforms for speakers bdl, ksp, slt, and clb. The highest PESQ scores are obtained around $w_1 = -0.5$. Informal listening tests also confirmed this observation. The optimal value of w_1 is set to -0.5. Reconstructed speech samples are available for listening at the link: `https://jitendradhiman.github.io/RZParametersImpactOnRecons.html`.

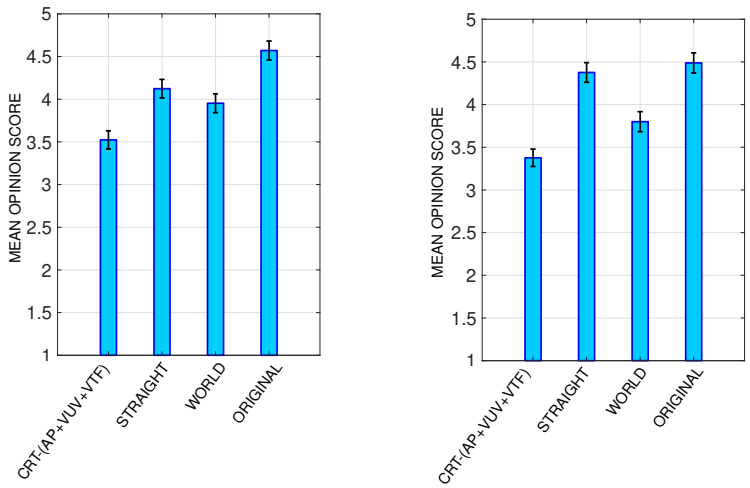


Figure 6.13: Comparison between CRT and WORLD/STRAIGHT in terms of the mean opinion score of the analysis/synthesis quality of speech: (a) male speakers (bdl, ksp); and (b) female speakers (clb, slt) taken from CMU-ARCTIC database. The error bars show 95% confidence interval.

6.4.2 *Subjective Evaluation*

We use standard MOS test employing 25 listeners in the age group of 21 to 30 years with normal hearing. We used 40 speech utterances (20 - male and 20 - female from CMU-ARCTIC database). The listening test was conducted in a soundproof chamber and the listeners were given a Sennheiser HD 650 headphone. Since the listening test is a time-consuming process, given the constraints due to Covid-19, we conducted the test only for Case 3 and Case 5 in Table 6.9. For comparison, we included STRAIGHT and WORLD vocoders. The tests were conducted in two sets. In the first set, we compared Case 3, STRAIGHT and WORLD. In the second set, we compared Case 5, STRAIGHT and WORLD. Case 3 indicates the joint impact of CRT-AP and CRT-V/UV; Case 5 also includes CRT-VTF. Figure 6.12 displays the results of MOS test for Case 3, the performance of CRT-AP and CRT-V/UV is on par with the state-of-the-art. Figure 6.13 displays the results of MOS test for Case 5, the performance of CRT-VTF is inferior to the existing methods.

6.5 Speech Reconstruction Using WaveNet

WaveNet [2] is a deep generative model for audio waveform generation. A speech waveform is time-series data and has both short-term and long-term dependencies. The temporal correlations are modeled using the *autoregressive model* in a probabilistic sense. The audio samples $\boldsymbol{x} = \{x_1, x_2, \ldots, x_T\}$ are modeled as follows:

$$p(\boldsymbol{x}) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1}), \tag{6.16}$$

where $p(\boldsymbol{x})$ represents the joint density. The sample x_t is therefore conditioned on the past samples up to instant $(t-1)$. The accuracy of generative models could be improved by suitably conditioning on an auxiliary feature $\boldsymbol{h}$ as follows:

$$p(\boldsymbol{x}|\boldsymbol{h}) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1}, \boldsymbol{h}), \tag{6.17}$$

where $\boldsymbol{h}$ are the mel-frequency cepstral coefficients in the standard WaveNet model. In our model, $\boldsymbol{h}$ corresponds to the acoustic features derived from CRT analysis. WaveNet could be interpreted as a statistical vocoder that is based on a nonlinear autoregressive model (Equation (6.17)) for sample generation given the past samples and acoustic features. By conditioning the model on acoustic features, one can guide WaveNet to produce realistic speech waveforms.

Figure 6.14 displays the architecture of WaveNet vocoder. The major blocks are as follows.

• **μ-law compression:** During inference, WaveNet produces speech samples as a function of time in an autoregressive fashion. The sampling rate and bit width determine the complexity of the synthesis. For instance, a 16 bit representation has $2^{16} = 65,536$ possible amplitudes, which means that at each instant, it is required to output one out of 65,536 possible outputs – this is more challenging than a typical classification problem. To address this issue, the speech waveform is subjected to

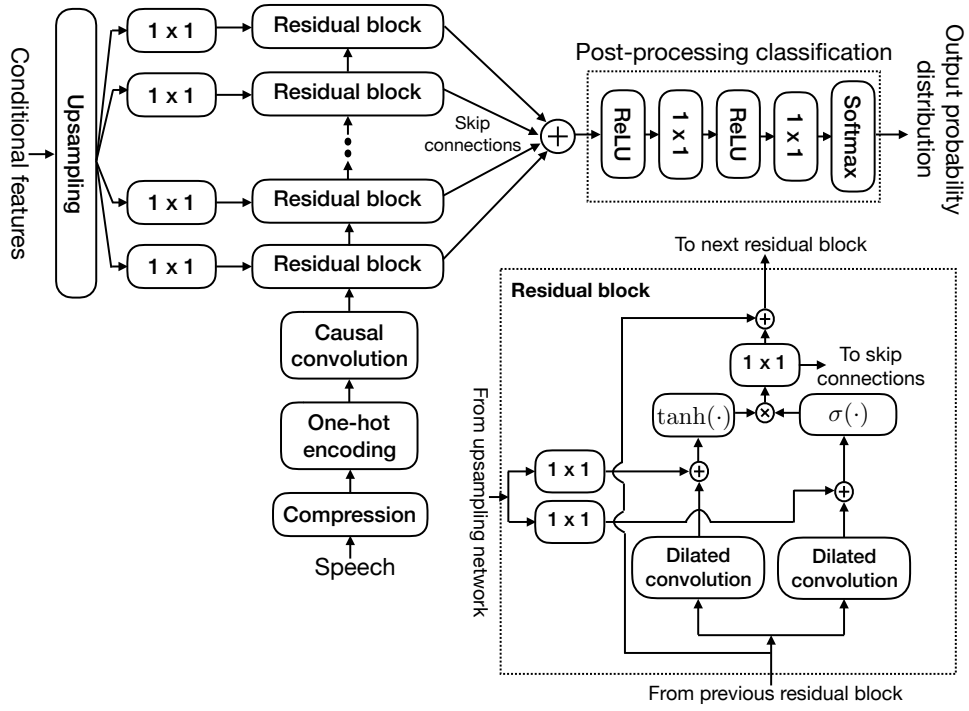


Figure 6.14: The WaveNet architecture [2].

μ-law compression to reduce the dynamic range:

$$y_t = \text{sign}(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}, \ |x_t| < 1, \tag{6.18}$$

where x_t and y_t denote the input and transformed waveform samples, respectively. The output lies between 0 and $\mu - 1$. The parameter μ is set to 256. After compression, there are only 256 categories to predict from, which is a relatively easier problem.

• **Dilated causal convolutions:** WaveNet uses a stack of dilated convolutional layers to model the long-range dependencies of audio samples in Equation (6.16). Stacking dilated convolution layers increases the temporal support of the network. The dilation factor is increased exponentially from one layer to the next and typically repeated after a certain limit. For example, a dilation depth of 3 and repeat of 2

Table 6.11: Values of the design parameters in WaveNet.

Dilation depth	Dilation repeat	Kernel size	Residual channels	Skip channels	μ
10	3	2	512	256	256

gives dilation factors: $1, 2, 4, 1, 2, 4$ in the stack. In addition, these layers use causal convolution, i.e., the prediction $p(x_t|x_1, \ldots, x_{t-1})$ emitted by the model at time t does not depend on the future time-steps $x_{t+1}, x_{t+2}, \ldots, x_T$.

• **Residual block:** WaveNet uses a stack of residual blocks (see Figure 6.14) to increase the model capacity in order to account for long-term dependencies. The main source of non-linearity in a residual block is the gated activation unit which is the same as the one used in PixelCNN network [160]. The conditional features are typically at a lower sampling rate than the signal. Hence, before feeding them to the residual block, the conditional features are upsampled to the same resolution as the speech samples to be generated. The output of a gated activation unit in the r^{th} residual block is given by

$$\boldsymbol{z} = \tanh\left(W_f^{(r)} * \boldsymbol{x}^{(r-1)} + V_f^{(r)} * \boldsymbol{y}\right) \odot \sigma\left(W_g^{(r)} * \boldsymbol{x}^{(r-1)} + V_g^{(r)} * \boldsymbol{y}\right), \qquad (6.19)$$

where $\boldsymbol{x}^{(r-1)}$ is the output of the previous residual block in the stack, $\boldsymbol{y}$ denotes the extended time series of the original features $\boldsymbol{h}$ at the time resolution adjusted to $\boldsymbol{x}^{r-1}$, $\sigma(a) = \frac{1}{1+e^{-a}}$, and $W_f^{(r)}, W_g^{(r)}, V_f^{(r)}, V_g^{(r)}$ are learnable linear transformations. Residual and skip connections [161] are used to facilitate faster convergence and training.

6.6 Experimental Results

We use 2 male (bdl, ksp) and 2 female (clb, slt) speakers from CMU-ARCTIC database. The database for each of the speakers consists of 1132 sentences out of which 1028 sentences were used for training the WaveNet, and 104 for testing (sampling rate is 16 kHz). We extract the following features: pitch (F0), voiced/unvoiced

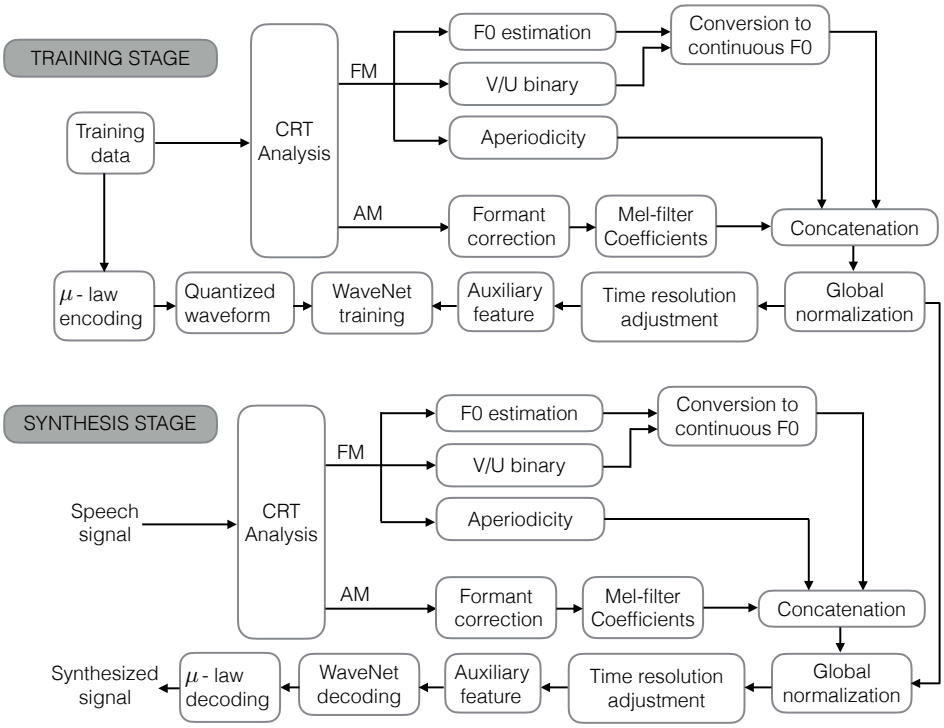


Figure 6.15: Block diagram illustrating the training and synthesis phases in a WaveNet vocoder using the acoustic features from CRT-based analysis.

decisions (V/UV), bandwise aperiodicity parameters (BAP), and the vocal-tract filter (VTF). The pitch contour is linearly interpolated in unvoiced segments. We compute 25-dimensional Mel-filterbank features from the VTF. The concatenation of F0 (1-dim), BAP (3-dim), VTF (25-dim) gives a 29-dimensional acoustic feature, which is used as a conditional feature ($\boldsymbol{h}$) in WaveNet. The three BAP parameters correspond to the frequency subbands: 0-4 kHz, 2-6 kHz, and 4-8 kHz. For comparison, we used 3 sets of acoustic features extracted using STRAIGHT, WORLD, and CRT and trained corresponding WaveNet vocoders in a speaker-dependent manner. Table 6.11 lists the parameter choices. Figure 6.15 shows the block diagram for WaveNet training and speech reconstruction. The feature vectors are normalized to have zero-mean and unit-variance ("Global normalization" block).

Table 6.12: Objective scores (averages and standard deviation) for speech reconstructed using the STRAIGHT, WORLD and CRT (proposed) features incorporated in a WaveNet vocoder. The scores were averaged over 104 test speech utterances for each speaker taken from CMU-ARCTIC database.

Method	bdl (male)	ksp (male)	clb (female)	slt (female)
	PESQ: Speech quality test			
STRAIGHT	3.14 ± 0.19	3.40 ± 0.14	3.38 ± 0.14	3.51 ± 0.17
WORLD	$\mathbf{3.46 \pm 0.14}$	$\mathbf{3.50 \pm 0.12}$	$\mathbf{3.49 \pm 0.22}$	3.26 ± 0.16
CRT (proposed)	3.37 ± 0.14	3.39 ± 0.13	3.41 ± 0.17	$\mathbf{3.65 \pm 0.15}$

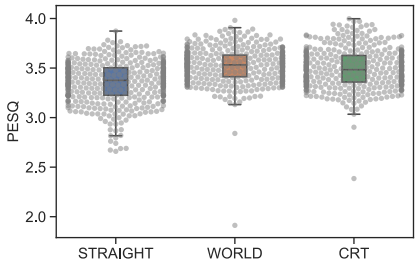


Figure 6.16: PESQ scores for the reconstructed speech waveforms corresponding to the speakers bdl, ksp, clb, and slt by using the acoustic features from STRAIGHT, WORLD, and CRT in the WaveNet vocoder. Both box and swarm plots are shown. A gray dot represents the objective score corresponding to a speech utterance.

6.6.1 *Objective Evaluation*

Table 6.12 shows the results of objective evaluation. We observe that the performance of CRT is on par with STRAIGHT and WORLD in the WaveNet setting. Figure 6.16 shows the PESQ scores for the reconstructed speech waveforms using acoustic features derived from STRAIGHT, WORLD, and CRT in the WaveNet setting. We observe that PESQ for STRAIGHT goes below 3.0 for some speech utterances in the dataset. On the contrary, PESQ for WORLD and CRT are concentrated above 3.0 barring a few outliers. Some synthesized speech samples are available for listening at the link: `https://jitendradhiman.github.io/waveNet.html`.

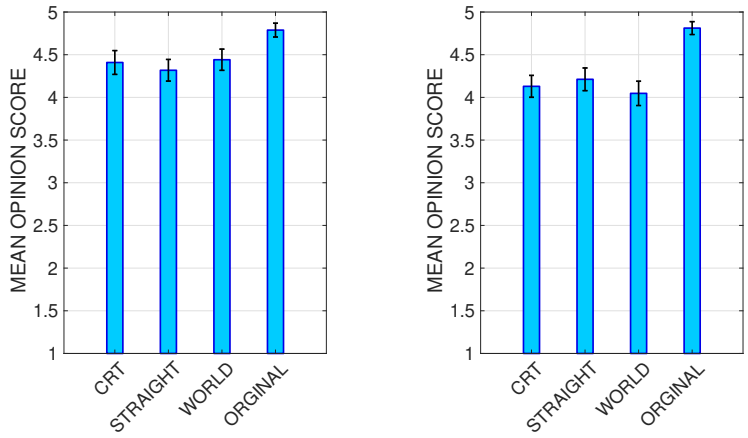


Figure 6.17: Mean opinion scores for assessing the synthesis quality of speech for (a) male speakers (bdl, ksp); and (b) female speakers (clb, slt) taken from CMU-ARCTIC database. The subjective evaluation compares CRT, STRAIGHT and WORLD features operating in a WaveNet setting. The error-bars show 95% confidence interval.

6.6.2 *Subjective Evaluation*

We conduct MOS test for subjective evaluation of the synthesized speech in the WaveNet setting. Seventeen listeners participated in the test and the test setup is the same as mentioned in Section 6.4. We select 2 male (bdl, ksp) and 2 female (clb, slt) from CMU-ARCTIC database. For each speaker, 5 randomly chosen speech utterances are used for MOS evaluation. In a single trial, the subject listens to 4 speech samples comprising speech synthesized using CRT, STRAIGHT, WORLD features in addition to the original waveform. As a result, a subject listens to a total of 80 speech utterances. The subjects were asked to rest for 10 minutes after 20 minutes of participating in the experiment. The order of presentation is random. Figure 6.17 displays the MOS scores for male and female speakers. We observe that the performance of CRT features is on par with the state-of-the-art for both genders. The MOS scores for individual speakers (bdl, ksp, clb, and slt) are displayed in Figure 6.18, from which the same conclusion can be drawn.

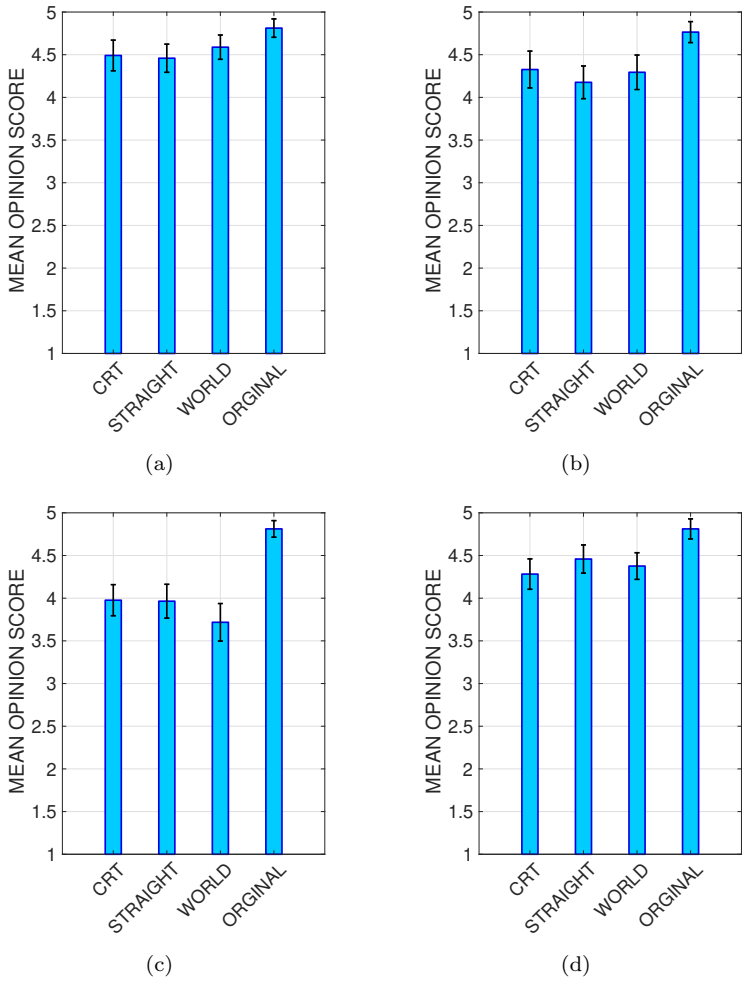


Figure 6.18: Mean opinion scores for assessing the synthesis quality of speech for (a) bdl, (b) ksp, (c) clb, and (d) slt speakers taken from CMU-ARCTIC database. The subjective evaluation compares CRT, STRAIGHT and WORLD features operating in a WaveNet setting. The error-bars show 95% confidence interval.

The statistical significance of the MOS scores was determined by conducting pairwise Mann-Whitney U-test [162] for the 6 pairs: CRT-ORIGINAL, CRT-STRAIGHT, CRT-WORLD, ORIGINAL-STRAIGHT, ORIGINAL-WORLD, STRAIGHT-WORLD. The corresponding MOS scores from the male and female speakers were pooled before computing the p-value for each pair. Table 6.13 shows

Table 6.13: p-values for different pairs in Mann-Whitney U-test to test the statistical significance of MOS values for speech reconstructed using WaveNet.

	ORIGINAL	STRAIGHT	WORLD
CRT	< 0.001	0.003	0.004
STRAIGHT	< 0.001	-·-	0.003
WORLD	< 0.001	-·-	-·-

the p-values — all pairwise differences were found to be statistically significant.

6.7 Chapter Summary

We considered the problem of vocal-tract filter estimation using the CRT approach. The formant bandwidths were found to be over-estimated. Since accurate formant bandwidth estimates are required for high-fidelity speech reconstruction, we proposed a correction method. The effectiveness of the correction was demonstrated on synthetic vowels and real speech. The accuracy of source and filter parameter estimates was analyzed for the task of speech reconstruction in the spectral synthesis model and WaveNet vocoder. Previous research has shown that the spectral synthesis model is sensitive to oversmoothing of the parameters. Our findings also confirmed that oversmoothing results in muffled speech. Among the parameters estimated using CRT, the vocal-tract filter estimate was found to be below par when incorporated in the spectral synthesis model. However, deep learning based WaveNet vocoder proved to be insensitive to oversmoothing. The quality of synthesized speech samples using WaveNet operating on CRT derived features was found to be on par with that of STRAIGHT and WORLD.

Chapter 7

Conclusions and Outlook

The key objective of this dissertation is systematic development of techniques for spectrotemporal analysis of speech signals. The idea is to divide a spectrogram into smaller patches, each of which is modeled using a 2-D AM-FM cosine. We showed that the proposed technique gave rise to several spectrotemporal representations that highlight both source and filter parameters. The key ingredient is *the complex Riesz transform* (CRT), which enables efficient demodulation of a narrowband pitch-adaptive spectrogram into 2-D AM and FM components that capture the filter and source attributes, respectively.

In this chapter, we summarize our findings, highlight the main contributions of this thesis, and discuss possibilities for further work.

7.1 Summary of the Contributions

In **Chapter 2**, we reviewed 2-D AM-FM cosines and discussed their Fourier-domain properties. The 2-D AM-FM model was applied to the voiced patches, in particular, a multicomponent one that uses a sum of weighted 2-D AM-FM cosines. The number of terms or the order must be adapted to the speaker's pitch. Extensive objective evaluation on standard speech databases showed that the multicomponent model has higher accuracy than the monocomponent model, particularly for high-pitched sounds. This observation indicates that a multicomponent AM-FM model is particularly better suited for female speakers. The estimation of the AM/FM

parameters requires one to solve the demodulation problem in 2-D, for which we used the complex Riesz transform.

The demodulation gave rise to the 2-D carrier spectrogram, which predominantly contains the fundamental frequency and the AM, which highlights vocal-tract resonances.

In **Chapter 3**, we examined the carrier spectrogram for different speech sounds and showed how it could be used to decompose a speech signal into its periodic and aperiodic components. The carrier spectrogram is particularly suited for detecting coherent versus incoherent t-f patterns. The coherent and incoherent t-f patterns were described using two t-f maps derived from the carrier spectrogram: *the coherencegram* and *orientationgram*. While the coherencegram aids in identifying the harmonic and inharmonic patterns, the orientationgram gives the local orientation. Taking the two t-f maps jointly proved to be useful to classify t-f bins as either periodic or aperiodic.

Chapter 4 dealt with the problem of voiced/unvoiced segmentation and aperiodicity estimation of speech using the carrier spectrogram. We derived novel features from coherencegram and used them for the task of voiced/unvoiced segmentation. Unlike the state-of-the art features which are typically obtained in the time-domain, coherence-based features were found to be relatively insensitive to the local variations of the speech signals. Objective evaluation showed that the coherence-based features outperformed the state-of-the-art. Next, we used the carrier spectrogram for estimating the speech aperiodicity and derived band-wise aperiodicity parameters, which were found to be useful for modeling the noise component in a spectral synthesis model for the task of speech reconstruction.

In **Chapter 5**, we addressed the problem of pitch estimation from the carrier spectrogram, which encodes the temporal evolution of pitch and its harmonics. We propose two methods and showed that their performance was on par with state-of-the-art techniques on both clean and noisy speech databases.

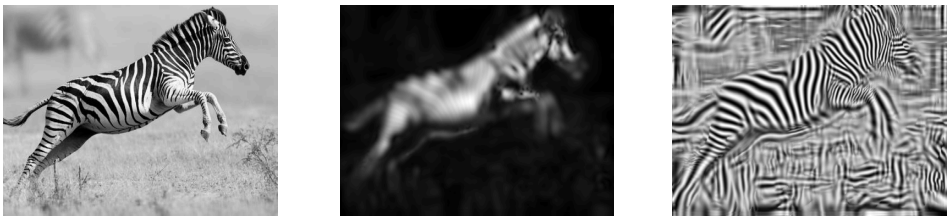


Figure 7.1: [Image taken from Google] (a) A grayscale image of zebra; (b) the AM component; and (c) the FM component.

In Chapters 3-5, the emphasis was on the demodulation of narrowband spectrogram and estimation of the source parameters from the resulting carrier spectrogram. In **Chapter 6**, we argued that the pitch-adaptive spectrogram is a better candidate for the estimation of vocal-tract filter. The bandwidths of formants have a direct impact on the quality of speech waveform reconstructed using a spectral synthesis model. Hence, we proposed a method to tune the formant bandwidths and consequently have a better control on the quality of synthesized speech. We then showed the effect of the source and filter parameters on the quality of speech reconstructed by using a spectral synthesis model and the WaveNet vocoder. The baseline features were extracted using STRAIGHT and WORLD, and compared with the proposed features.

Thus, we established that spectrotemporal analysis of speech signals using the complex Riesz transform is a promising alternative to short-time processing.

Out of sheer curiosity, we considered the demodulation of a zebra! Figure 7.1 shows the results of the demodulation – indeed the technique separated the zebra from its stripes.

7.2 Outlook

The advantages of the spectrotemporal analysis technique are accompanied by certain challenges. Consider a spectrogram computed using a frame-shift of 1 ms and 1024-length discrete Fourier transform. For a 3 s long signal at 16 kHz sampling rate, the spectrogram is of size 3000×512. A patch size of 600 Hz $\times$ 100 ms or 38×100 samples, and a hop size of 25 and 8 samples along time and frequency dimensions, respectively, gives a total of $3000/25 \times 512/8 \approx 7680$ patches. The demodulation takes approximately 200 s on a 20-CPU core machine. Developing fast algorithms for demodulation is open for further research.

Oversmoothing of the estimated parameters is a shortcoming of spectrotemporal analysis. Demodulation is effectively a process of separating a signal into slowly-varying (AM) and fast-varying (FM) components. Operating on large t-f patches invariably introduces additional smoothing, which affects the quality of synthesis, as it happened with the formant bandwidth estimates operating in the spectral synthesis model. Such degradation was not observed in the case of the WaveNet model. This indicates that the smoothing effects can be overcome by superior statistical modeling. This aspect requires further investigation.

Although we alluded to the multicomponent AM-FM model, we worked largely with the monocomponent model. Recall from Chapter 2 that the multicomponent model parameters are estimated by patch-wise least-squares regression. The t-f maps of $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_K$ may give rise to new insights into the model. The t-f maps of α_0 and α_1 are shown in Figure 7.2. We observe that α_0 is relatively higher for inharmonic t-f regions – this is because α_0 is a representative of the residual error after fitting the cosine term, and is therefore higher for inharmonic t-f regions. On the other hand, α_1 is higher in t-f regions containing a temporal discontinuity. For instance, α_1 could represent the occurrence of plosives or transient sounds.

This dissertation focused only on narrowband and pitch-adaptive spectrograms. AM-FM analysis of wideband spectrograms has been relatively unexplored and

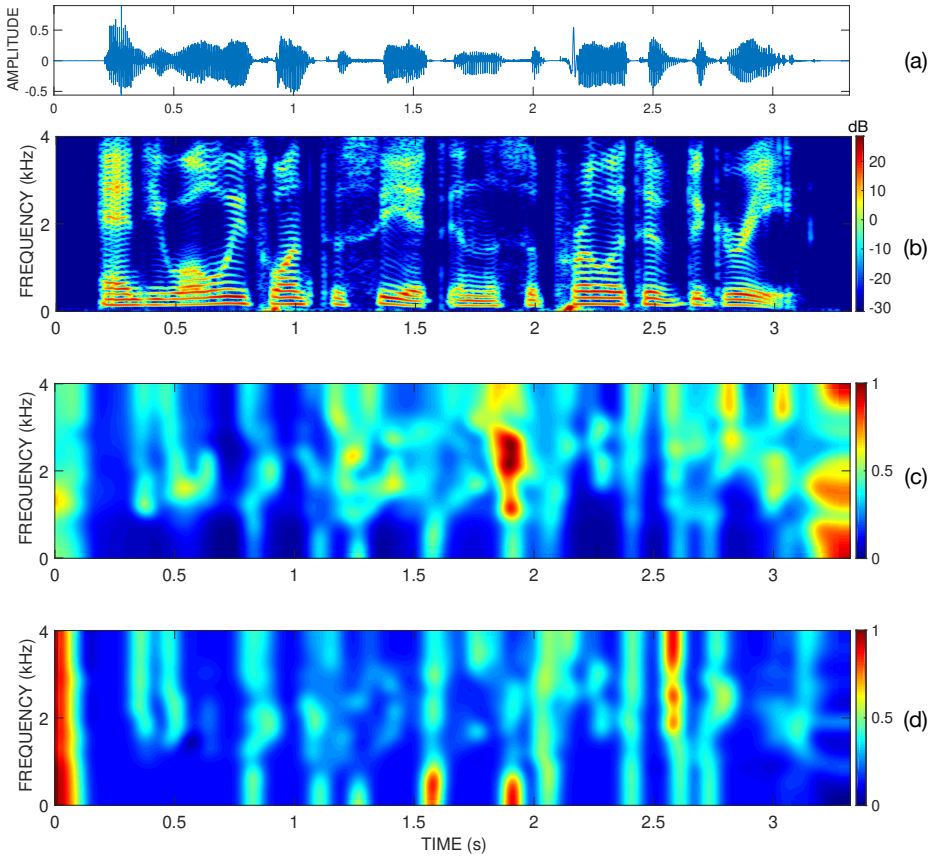


Figure 7.2: (a) A speech utterance "*And you always want to see it in the superlative degree,*" spoken by a female speaker; (b) its narrowband spectrogram; t-f maps of coefficients (c) α_0, and (d) α_1. The t-f maps of α_0 and α_1 are normalized between 0 and 1 for visualization.

might hold the key to interesting speech properties. Also, the analysis employed spectrotemporal patches of fixed dimension. Adapting the t-f patch size is an aspect that requires more investigation.

The proposed technique gives access to the source and filter parameters separately. Such segregated streams of information may be useful in applications such as voice conversion [163–165], singing voice synthesis [166], auditory chimaeras [167], etc.

Appendix A

Obtaining the Frequency Response From the Magnitude Response of a Minimum-Phase Sequence

Consider a real and minimum-phase sequence $h[n]$, and its discrete-time Fourier transform $\hat{h}(e^{\mathrm{j}\omega})$. We show that the frequency response $\hat{h}(e^{\mathrm{j}\omega})$ can be exactly obtained from its magnitude response $|\hat{h}(e^{\mathrm{j}\omega})|$.

Consider the real cepstrum of $h[n]$ given by

$$\tilde{c}[n] = \mathcal{F}^{-1}\{\ln|\hat{h}(e^{\mathrm{j}\omega})|\}, \tag{A.1}$$

and its complex cepstrum given by

$$c[n] = \mathcal{F}^{-1}\{\ln\hat{h}(e^{\mathrm{j}\omega})\}, \tag{A.2}$$

where $\mathcal{F}^{-1}$ denotes the inverse discrete-time Fourier transform. The real and complex cepstra are related as follows:

$$\tilde{c}[n] = \frac{c[n] + c[-n]}{2}. \tag{A.3}$$

The proof is as follows. Consider

$$\begin{aligned} c[n] + c[-n] &= \mathcal{F}^{-1}\{\ln\hat{h}(e^{\mathrm{j}\omega}) + \ln\hat{h}(e^{-\mathrm{j}\omega})\} \\ &= \mathcal{F}^{-1}\{\ln(\hat{h}(e^{\mathrm{j}\omega})\hat{h}^*(e^{\mathrm{j}\omega}))\} \\ &= 2\mathcal{F}^{-1}\{\ln|\hat{h}(e^{\mathrm{j}\omega})|\} \\ &= 2\tilde{c}[n], \end{aligned} \tag{A.4}$$

where we have used the Fourier property that $\hat{h}(e^{-\mathrm{j}\omega})) = \hat{h}^*(e^{\mathrm{j}\omega})$ for a real sequence $h[n]$. The second property that is used is that the complex cepstrum of a minimum-phase sequence is casual, *i.e.,*

$$c[n] = 0, \quad \text{for} \quad n < 0. \tag{A.5}$$

Using Equation (A.4) and Equation (A.5), the complex cepstrum of $h[n]$ can be expressed in terms of its real cepstrum as follows:

$$c[n] = \begin{cases} 0, & \text{for } n < 0, \\ \tilde{c}[n], & \text{for } n = 0, \\ 2\tilde{c}[n], & \text{for } n > 0. \end{cases} \tag{A.6}$$

Hence, in order to recover $\hat{h}(e^{\mathrm{j}\omega})$ from $|\hat{h}(e^{\mathrm{j}\omega})|$, one can compute the real cepstrum using Equation (A.1) and the complex cepstrum using Equation (A.6). The frequency response is obtained by plugging the complex cepstrum in Equation (A.2).

Appendix B

Least-Squares Overlap-Add in 2-D

The 2-D overlap-add method is used to reconstruct the spectrogram from its spectrotemporal patches. Let $S_W^{i,j}(t,\omega)$ denotes the (i,j)th windowed spectrogram patch. An approximation of the reconstructed spectrogram $\tilde{S}(t,\omega)$ from its windowed patches is obtained by solving the following optimization problem:

$$\arg\min_{\tilde{S}(t,\omega)} \sum_{i,j} \left(\tilde{S}(t,\omega) W(t - iT, \omega - jF) - S_W^{i,j}(t,\omega) \right)^2, \tag{B.1}$$

where $W(t,\omega)$ denotes the 2-D window, T and F denote the hop sizes along time and frequency axes, respectively. The cost function given in Equation (B.1) is minimized by computing its derivative with respect to $\tilde{S}(t,\omega)$ and setting it to zero:

$$\sum_{i,j} \left(\tilde{S}(t,\omega) W(t - iT, \omega - jF) - S_W^{i,j}(t,\omega) \right) W(t - iT, \omega - jF) = 0, \tag{B.2}$$

which gives

$$\tilde{S}(t,\omega) = \frac{\sum_{i,j} \tilde{S}_W^{i,j}(t,\omega) W(t - iT, \omega - jF)}{\sum_{i,j} W^2(t - iT, \omega - jF)}.$$

The above formula gives the reconstructed spectrogram by using 2-D overlap-add of the spectrotemporal patches in the least-squares sense (2-D OLA-LSE).

Publications

[1] **Jitendra K. Dhiman**, Nagaraj Adiga and Chandra Sekhar Seelamantula, "On the suitability of the Riesz spectro-temporal envelope for WaveNet based speech synthesis", *in Proc. Interspeech*, 2019.

[2] **Jitendra K. Dhiman** and Chandra Sekhar Seelamantula, "A spectro-temporal technique for estimating aperiodicity and voiced/unvoiced decision boundaries of speech signals," *in Proc. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2019.

[3] **Jitendra Dhiman**, Neeraj Sharma and Chandra Sekhar Seelamantula, "Multi-component 2-D AM-FM modeling of speech spectrograms," *in Proc. Interspeech*, 2018.

[4] **Jitendra K. Dhiman**, Nagaraj Adiga and Chandra Sekhar Seelamantula, "A spectrotemporal demodulation technique for pitch estimation," *in Proc. Interspeech*, 2017.

[5] Karthika Vijayan, **Jitendra K Dhiman**, and Chandra Sekhar Seelamantula, "Time-frequency coherence for periodic-aperiodic decomposition of speech signals," *in Proc. Interspeech*, 2017.

Other:

[1] Jishnu Sadasivan, **Jitendra K. Dhiman**, and Chandra Sekhar Seelamantula, "Musical Noise Suppression Using a Low-Rank and Sparse Matrix Decomposition Approach," in *Elsevier Speech Communication*, vol. 125, pp. 41-52, 2020.

Bibliography

[1] T. T. Wang and T. F. Quatieri, "Two-dimensional speech-signal modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1843–1856, 2012.

[2] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv:1609.03499*, 2016. [Online]. Available: https://arxiv.org/pdf/1609.03499.pdf

[3] A. Vafeiadis, E. Fanioudakis, I. Potamitis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Two-dimensional convolutional recurrent neural networks for speech activity detection," *in Proc. Interspeech*, 2019.

[4] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–15, 2015.

[5] A. Montalvo, Y. M. Costa, and J. R. Calvo, "Language identification using spectrogram texture," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2015, pp. 543–550.

[6] H. Mukherjee, S. Ghosh, S. Sen, O. S. Md, K. Santosh, S. Phadikar, and K. Roy, "Deep learning for spoken language identification: Can we visualize speech signal patterns?" *Neural Computing and Applications*, vol. 31, no. 12, pp. 8483–8501, 2019.

[7] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 6, pp. 871–879, 1988.

[8] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304,

1995.

[9] S. M. Schimmel, *Theory of modulation frequency analysis and modulation filtering, with applications to hearing devices*. Citeseer, 2007, vol. 68, no. 07.

[10] S. K. Nemala, K. Patil, and M. Elhilali, "A multistream feature framework based on bandpass modulation filtering for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 416–426, 2012.

[11] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[12] ——, "Effect of reducing slow temporal modulations on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2670–2680, 1994.

[13] S. Schimmel and L. Atlas, "Coherent envelope detection for modulation filtering of speech," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2005, pp. I–221.

[14] L. Atlas, Q. Li, and J. Thompson, "Homomorphic modulation spectra," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2004, pp. ii–761.

[15] S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-scale analysis for content identification," *IEEE Transactions on signal processing*, vol. 52, no. 10, pp. 3023–3035, 2004.

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

[17] S. Sukittanon and L. E. Atlas, "Modulation frequency features for audio fingerprinting," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2002, pp. II–1773.

[18] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[19] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE signal processing Magazine*, vol. 21, no. 5, pp. 95–106, 2004.

[20] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[21] D. A. Depireux, J. Z. Simon, J. Klein, David, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *Journal of Neurophysiology*, vol. 85, no. 3, pp. 1220–1234, 2001.

[22] S. A. Shamma, J. W. Fleshman, P. R. Wiser, and H. Versnel, "Organization of response areas in ferret primary auditory cortex," *Journal of Neurophysiology*, vol. 69, no. 2, pp. 367–383, 1993.

[23] F. E. Theunissen and J. E. Elie, "Neural processing of natural sounds," *Nature Reviews Neuroscience*, vol. 15, pp. 355–366, 2014.

[24] C. E. Schreiner and B. M. Calhoun, "Spectral envelope coding in cat primary auditory cortex: properties of ripple transfer functions," *Aud Neurosci*, vol. 1, no. 1, pp. 39–62, 1994.

[25] B. M. Calhoun and C. E. Schreiner, "Spectral envelope coding in cat primary auditory cortex: linear and non-linear effects of stimulus characteristics," *European Journal of Neuroscience*, vol. 10, no. 3, pp. 926–940, 1998.

[26] R. Christopher deCharms, D. T. Blake, and M. M. Merzenich, "Optimizing sound features for cortical neurons," *science*, vol. 280, no. 5368, pp. 1439–1444, 1998.

[27] S. A. Shamma, H. Versnel, and N. Kowalski, "Ripple analysis in ferret primary auditory cortex. i. response characteristics of single units to sinusoidally rippled spectra," Tech. Rep., 1994. [Online]. Available: http://hdl.handle.net/1903/5500

[28] D. J. Klein, D. A. Depireux, J. Z. Simon, and S. A. Shamma, "Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design," *Journal of Computational Neuroscience*, vol. 9, no. 1, pp. 85–111, 2000.

[29] N. Kowalski, D. A. Depireux, and S. A. Shamma, "Analysis of dynamic spectra in ferret primary auditory cortex. ii. prediction of unit responses to arbitrary dynamic spectra," *Journal of Neurophysiology*, vol. 76, no. 5, pp. 3524–3534, 1996.

[30] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824–839, 1992.

[31] N. R. Mahajan, N. Mesgarani, and H. Hermansky, "General properties of auditory spectro-temporal receptive fields," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. EL459–EL463, 2019.

[32] S. V. David, N. Mesgarani, and S. A. Shamma, "Estimating sparse spectro-temporal receptive fields with natural stimuli," *Network: Computation in Neural Systems*, vol. 18, no. 3, pp. 191–212, 2007.

[33] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[34] N. Mesgarani and S. Shamma, "Speech enhancement based on filtering the spectrotemporal modulations," in *Proceedings. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2005.*, vol. 1. IEEE, 2005, pp. I–1105.

[35] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Communication*, vol. 41, no. 2-3, pp. 331–348, 2003.

[36] S. V. Ravuri and N. Morgan, "Using spectro-temporal features to improve afe feature extraction for asr," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[37] D. Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[38] E. Bedrosian, "A product theorem for Hilbert transforms," *Proceedings of the IEEE*, vol. 51, no. 5, pp. 868–869, 1963.

[39] M. Unser and D. Van De Ville, "Higher-order riesz transforms and steerable wavelet frames," in *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2009, pp. 3801–3804.

[40] C. S. Seelamantula, N. Pavillon, C. Depeursinge, and M. Unser, "Local demodulation of holograms using the Riesz transform with application to microscopy," *Journal of the Optical Society of America*, vol. 29, no. 10, pp. 2118–2129, Oct. 2012.

[41] H. Aragonda and C. S. Seelamantula, "Demodulation of narrowband speech spectrograms using the Riesz transform," *IEEE/ACM Transactions on Audio, Speech, and Language Process.*, vol. 23, no. 11, pp. 1824–1834, Nov 2015.

[42] T. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2001.

[43] M. R. Schroeder, "Vocoders: Analysis and synthesis of speech," *Proceedings of the IEEE*, vol. 54, no. 5, pp. 720–734, 1966.

[44] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, vol. 27(3-4), pp. 187–207,

1999.

[45] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[46] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder." in *Proc. Interspeech*, 2017, pp. 1118–1122.

[47] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv:1802.08435*, 2018. [Online]. Available: https://arxiv.org/pdf/1802.08435.pdf

[48] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "FFTNet: A real-time speaker-dependent neural vocoder," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2251–2255.

[49] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5891–5895.

[50] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3617–3621.

[51] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[52] Y. Ai and Z.-H. Ling, "A neural vocoder with hierarchical generation of amplitude and phase spectra for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 839–851, 2020.

[53] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," *arXiv:1606.07947*, 2016.

[54] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *arXiv preprint arXiv:1807.03039*, 2018. [Online]. Available: https://arxiv.org/pdf/1807.03039.pdf

[55] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv:1406.2661*,

2014. [Online]. Available: https://arxiv.org/pdf/1406.2661.pdf

[56] J. Kominek and A. W. Black, "The CMU-Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[57] "Starkey Hearing Technologies. Open access stimuli for the creation of multi-talker maskers," http://www.starkeyevidence.com, (Last accessed: May 2021).

[58] G. Fairbanks, *Voice and Articulation Drillbook*. New York:Harper & Row., 1960, vol. 2.

[59] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. nist speech disc 1-1.1," *NASA STI/Recon Technical Report*, vol. 93, p. 27403, 1993.

[60] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. 2006 IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 1.

[61] L. Deng, L. J. Lee, H. Attias, and A. Acero, "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances," in *Proc. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–557.

[62] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application*. Wiley Online Library, 2006.

[63] I. Rec, "P. 800: Methods for subjective determination of transmission quality," *International Telecommunication Union, Geneva*, vol. 22, 1996.

[64] ITU-T Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[65] K. G. Larkin, D. J. Bone, and M. A. Oldfield, "Natural demodulation of two-dimensional fringe patterns. i. general background of the spiral phase quadrature transform," *Journal of the Optical Society of America (A)*, vol. 18, no. 8, pp. 1862–1870, 2001.

[66] K. G. Larkin, "Natural demodulation of two-dimensional fringe patterns. ii. stationary phase analysis of the spiral phase quadrature transform," *Journal of the Optical Society of America (A)*, vol. 18, no. 8, pp. 1871–1881, 2001.

[67] T. T. Wang and T. F. Quatieri, "Towards interpretive models for 2-D processing of

speech," *IEEE Transactions on Audio, Speech, and Language Process.*, vol. 20, no. 7, pp. 2159–2173, Sept 2012.

[68] T. Ezzat, J. Bouvrie, and T. Poggio, "Spectro-temporal analysis of speech using 2-D Gabor filters," in *Proceedings of the Eighth Annual Conference of the International Speech Communication Association*, 2007.

[69] ——, "AM-FM demodulation of spectrograms using localized 2D Max-Gabor analysis," in *Proc. 2007 IEEE International Conference on Acoustics, Speech and Signal Process. (ICASSP)*, vol. 4, 2007, pp. IV–1061–IV–1064.

[70] V. D. Madjarova, H. Kadono, and S. Toyooka, "Dynamic electronic speckle pattern interferometry (DESPI) phase analyses with temporal Hilbert transform," *Optics Express*, vol. 11, no. 6, pp. 617–623, 2003.

[71] R. Onodera, H. Watanabe, and Y. Ishii, "Interferometric phase-measurement using a one-dimensional discrete Hilbert transform," *Optical Review*, vol. 12, no. 1, pp. 29–36, 2005.

[72] P. Pavliček and V. Michalek, "White-light interferometry—envelope detection by Hilbert transform and influence of noise," *Optics and Lasers in Engineering*, vol. 50, no. 8, pp. 1063–1068, 2012.

[73] S. Wang, L. Xue, J. Lai, and Z. Li, "An improved phase retrieval method based on Hilbert transform in interferometric microscopy," *Optik*, vol. 124, no. 14, pp. 1897–1901, 2013.

[74] H. Taub and D. L. Schilling, *Principles of Communication Systems*. McGraw-Hill Higher Education, 1986.

[75] G. H. Granlund and H. Knutsson, *Signal Processing for Computer Vision*. Springer Science & Business Media, 2013.

[76] S. L. Hahn, "Multidimensional complex signals with single-orthant spectra," *Proceedings of the IEEE*, vol. 80, no. 8, pp. 1287–1300, 1992.

[77] ——, *Hilbert Transforms in Signal Processing*. Artech House Boston, 1996, vol. 2.

[78] M. Kaseb, G. Mercère, H. Biermé, F. Brémand, and P. Carré, "Phase estimation of a 2D fringe pattern using a monogenic-based multiscale analysis," *Journal of the Optical Society of America (A)*, vol. 36, no. 11, pp. C143–C153, 2019.

[79] V. Sierra-Vázquez and I. Serrano-Pedraza, "Application of Riesz transforms to the isotropic AM-PM decomposition of geometrical-optical illusion images," *Journal of the Optical Society of America (A)*, vol. 27, no. 4, pp. 781–796, Apr 2010.

[80] H. Knutsson, C.-F. Westin, and M. Andersson, "Representing local structure using tensors ii," in *Scandinavian Conference on Image Analysis*. Springer, 2011, pp. 545–556.

[81] G. Strang, *Introduction to Linear Algebra*. Wellesley-Cambridge Press Wellesley, MA, 1993, vol. 3.

[82] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE Transactions on Signal Processing*, vol. 49, no. 12, pp. 3136–3144, 2001.

[83] S. Bernstein, J.-L. Bouchot, M. Reinhardt, and B. Heise, *Generalized Analytic Signals in Image Processing: Comparison, Theory and Applications*. Basel: Springer Basel, 2013, pp. 221–246. [Online]. Available: https://doi.org/10.1007/978-3-0348-0603-9_11

[84] P. Boersma, "Praat: Doing phonetics by computer," http://www.praat.org/, (Last accessed: May 2021).

[85] W. M. Fisher, "Ther darpa speech recognition research database: specifications and status," in *Proc. DARPA Workshop on Speech Recognition, Feb. 1986*, pp. 93–99.

[86] G. Richard and C. R. d'Alessandro, "Modification of the aperiodic component of speech signals for synthesis," in *Progress in Speech Synthesis*. Springer, 1997, pp. 41–56.

[87] S. B. Jebara, "Periodic/aperiodic decomposition for improving coherence based multi-channel speech denoising," in *Proc. 2007 9th IEEE International Symposium on Signal Processing and its Applications*, pp. 1–4.

[88] G. Richard and C. d'Alessandro, "Analysis/synthesis and modification of the speech aperiodic component," *Speech Communication*, vol. 19, no. 3, pp. 221–244, 1996.

[89] G. Degottex, P. Lanchantin, and M. Gales, "A log-domain pulse model for parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 57–70, 2017.

[90] O. Fujimura, "An approximation to voice aperiodicity," *IEEE Trans. Audio Electroacoust.*, vol. 16, no. 1, pp. 68–72, Mar 1968.

[91] C. d'Alessandro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources," *IEEE Transactions on Speech and Audio processing*, vol. 6, no. 1, pp. 12–23, 1998.

[92] P. R. Cook, "Aperiodicities in the singer voice source," *Journal of the Acoustical Society of America*, vol. 91, no. 4, pp. 2434–2434, 1992.

[93] J. Hillenbrand, "A methodological study of perturbation and additive noise in

synthetically generated voice signals," *Journal of Speech, Language, and Hearing Research*, vol. 30, no. 4, pp. 448–461, 1987.

[94] D. J. Hermes, "Synthesis of breathy vowels: Some research methods," *Speech Communication*, vol. 10, no. 5-6, pp. 497–502, 1991.

[95] I. Sertcelik and O. Kafadar, "Application of edge detection to potential field data using eigenvalue analysis of structure tensor," *Journal of Applied Geophysics*, vol. 84, pp. 86–94, 2012.

[96] Y. Yuan, D. Huang, Q. Yu, and P. Lu, "Edge detection of potential field data with improved structure tensor methods," *Journal of Applied Geophysics*, vol. 108, pp. 35–42, 2014.

[97] S. K. Nath and K. Palaniappan, "Adaptive robust structure tensors for orientation estimation and image segmentation," in *International Symposium on Visual Computing*. Springer, 2005, pp. 445–453.

[98] S. Lefkimmiatis, A. Roussos, P. Maragos, and M. Unser, "Structure tensor total variation," *SIAM Journal on Imaging Sciences*, vol. 8, no. 2, pp. 1090–1122, 2015.

[99] T. Brox, J. Weickert, B. Burgeth, and P. Mrázek, "Nonlinear structure tensors," *Image and Vision Computing*, vol. 24, no. 1, pp. 41–55, 2006.

[100] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[101] N. S. Jayant and P. Noll, "Digital coding of waveforms: principles and applications to speech and video," *Englewood Cliffs, NJ*, pp. 115–251, 1984.

[102] T. Drugman, G. Huybrechts, V. Klimkov, and A. Moinet, "Traditional machine learning for pitch detection," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1745–1749, 2018.

[103] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 201–212, 1976.

[104] L. Rabiner, C. Schmidt, and B. Atal, "Evaluation of a statistical approach to voiced-unvoiced-silence analysis for telephone-quality speech," *Bell System Technical Journal*, vol. 56, no. 3, pp. 455–482, 1977.

[105] L. Siegel, "Features for the identification of mixed excitation in speech analysis," in *Proc. 1979 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pp. 752–755.

[106] ——, "A procedure for using pattern classification techniques to obtain a

voiced/unvoiced classifier," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 1, pp. 83–89, 1979.

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

[108] L. Siegel and A. Bessey, "Voiced/unvoiced/mixed excitation classification of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, pp. 451–460, 1982.

[109] L. Rabiner and M. Sambur, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 4, pp. 338–343, 1977.

[110] H. Deng and D. O'Shaughnessy, "Voiced-unvoiced-silence speech sound classification based on unsupervised learning," in *2007 IEEE International Conference on Multimedia and Expo*. IEEE, 2007, pp. 176–179.

[111] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 250–255, 1993.

[112] S. Kia and G. G. Coghill, "A mapping neural network and its application to voiced-unvoiced-silence classification," in *Proceedings 1993 The First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*. IEEE, 1993, pp. 104–108.

[113] T. Ghiselli-Crippa and A. El-Jaroudi, "Voiced-unvoiced-silence classification of speech using neural nets," in *IJCNN-91-Seattle International Joint Conference on Neural Networks*, vol. 2. IEEE, 1991, pp. 851–856.

[114] A. H. Shandiz and L. Tóth, "Voice activity detection for ultrasound-based silent speech interfaces using convolutional neural networks," *arXiv:2105.13718*, 2021. [Online]. Available: https://arxiv.org/abs/2105.13718

[115] M. Wang, Q. Huang, J. Zhang, Z. Li, H. Pu, J. Lei, and L. Wang, "Deep learning approaches for voice activity detection," in *The International Conference on Cyber Security Intelligence and Analytics*. Springer, 2019, pp. 816–826.

[116] K. Struwe, "Voiced-unvoiced classification of speech using a neural network trained with lpc coefficients," in *2017 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)*. IEEE, 2017, pp. 56–59.

[117] J. Zeremdini, M. A. B. Messaoud, and A. Bouzid, "Two-speaker voiced/unvoiced decision for monaural speech," *Circuits, Systems, and Signal Processing*, pp. 1–17,

2020.

[118] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals." in *Proc. Interspeech*, 2009, pp. 2891–2894.

[119] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching," *in Proc. Eurospeech*, pp. 1003–1006, 1993.

[120] S. R. Kadiri and B. Yegnanarayana, "Analysis of aperiodicity in artistic Noh singing voice using an impulse sequence representation of excitation source," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. 4446–4457, 2019.

[121] H. Kawahara, O. Fujimura, and Y. Konpaku, "Voice quality of artistic expression in Noh: An analysis-synthesis study on source-related parameters," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 3028–3028, 2006.

[122] V. K. Mittal and B. Yegnanarayana, "Study of characteristics of aperiodicity in noh voices," *Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. 3411–3421, 2015.

[123] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[124] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system (STRAIGHT)," in *Proc. Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2001.

[125] H. Kawahara and M. Morise, "Simplified aperiodicity representation for high-quality speech manipulation systems," in *Proc. 2012 IEEE 11th International Conference on Signal Processing*, vol. 1, pp. 579–584.

[126] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proc. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 3933–3936.

[127] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[128] H. Kawahara, Y. Agiomyrgiannakis, and H. Zen, "Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis," *arXiv:1605.07809*, 2016. [Online]. Available: https://arxiv.org/pdf/1605.07809.pdf

[129] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[130] T. T. Wang and T. F. Quatieri, "Towards co-channel speaker separation by 2-D demodulation of spectrograms," in *Proc. IEEE Workshop on Applications of Signal Process to Audio and Acoustics*, Oct. 2009, pp. 65–68.

[131] B. Gold, N. Morgan, and D. Ellis, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & Sons, 2011.

[132] N. K. Sharma and T. V. Sreenivas, "Time-varying sinusoidal demodulation for non-stationary modeling of speech," *Speech Communication*, vol. 105, pp. 77–91, 2018.

[133] P. N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 102–105, 2012.

[134] A. S. Spanias, "Speech coding: A tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994.

[135] A. M. Zimmer, B. Dai, and S. A. Zahorian, "Personal computer software vowel training aid for the hearing impaired," in *Proc. 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 6, pp. 3625–3628.

[136] P. Ghahremani, B. Baba Ali, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2494–2498.

[137] Z. Inanoglu, "Transforming pitch in a voice conversion framework," *St. Edmond's College, University of Cambridge, Tech. Rep*, 2003.

[138] D. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: Factors responsible for quality," in *Proc. 1985 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 10, pp. 748–751.

[139] M. Eskenazi, "Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype," *Language learning & technology*, vol. 2, no. 2, pp. 62–76, 1999.

[140] C. Marques, S. Moreno, S. Luís Castro, and M. Besson, "Musicians detect pitch

violation in a foreign language better than nonmusicians: behavioral and electrophysiological evidence," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1453–1463, 2007.

[141] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[142] D. Gerhard, *Pitch extraction and fundamental frequency: History and current techniques*. Department of Computer Science, University of Regina Regina, Canada, 2003.

[143] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *Proc. 2002 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 1, pp. I–361.

[144] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 1, pp. 24–33, 1977.

[145] A. M. Noll, "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[146] M. Sondhi, "New methods of pitch extraction," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 262–266, 1968.

[147] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 333–338, May 1999.

[148] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[149] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech*, 2011, pp. 1973–1976.

[150] M. Morise, "HARVEST: A high-performance fundamental frequency estimator from speech signals,," in *Proc. Interspeech*, 2017, pp. 2321–2325.

[151] H. Kawahara, A. d. Cheveigne, and R. D. Patterson, "An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: revised tempo in the STRAIGHT-suite," in *Proc. Fifth International Conference on Spoken Language Processing*, 1998.

[152] H. Kawahara, H. Katayose, A. d. Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of

F0 and periodicity," in *Proc. Sixth European Conference on Speech Communication and Technology*, 1999.

[153] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Proc. 35th Intl. Audio Engineering Society Conference: Audio for Games*. Audio Engineering Society, 2009.

[154] B. S. Lee, "Noise robust pitch tracking by subband autocorrelation classification," Ph.D. dissertation, Columbia University, 2012.

[155] S. Lin, "A new frequency coverage metric and a new subband encoding model, with an application in pitch estimation." in *Proc. Interspeech*, 2018, pp. 2147–2151.

[156] L. Cohen, *Time-frequency Analysis*. Prentice Hall PTR Englewood Cliffs, NJ, 1995, vol. 778.

[157] B. Boashash, "Heuristic formulation of time-frequency distributions." Elsevier, 2003.

[158] L. Cohen and C. Lee, "Instantaneous bandwidth for signals and spectrogram," in *Proc. 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2451–2454.

[159] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, vol. 67, pp. 1–7, 2015.

[160] A. v. d. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," *arXiv:1606.05328*, 2016.

[161] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. European Conference on Computer Vision*. Springer, 2016, pp. 630–645.

[162] P. E. McKnight and J. Najab, "Mann-whitney u test," *The Corsini encyclopedia of psychology*, pp. 1–1, 2010.

[163] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. 2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 266–273.

[164] S. Lee, B. Ko, K. Lee, I.-C. Yoo, and D. Yook, "Many-to-many voice conversion using conditional cycle-consistent adversarial networks," in *Proc. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6279–6283.

[165] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6790–6794.

[166] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, "WGANSing: A multi-voice singing voice synthesizer based on the Wasserstein-GAN," in *Proc. 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[167] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 6876, pp. 87–90, 2002.