ENHANCED SINUSOIDAL RECONSTRUCTION FOR A HIGH QUALITY, MID-RATE MBE SPEECH CODER

By

TABITHA JOY PARKER

Bachelor of Science in Computer Engineering Technology

University of Arkansas at Little Rock

Little Rock, Arkansas

1995

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May, 1998

ENHANCED SINUSOIDAL RECONSTRUCTION

FOR A HIGH QUALITY, MID-RATE

MBE SPEECH CODER

Thesis Approved:

Thesis Advisor

R. Ramakumar.

Wayne B. Powell

Dean of the Graduate College

ACKNOWLEDGMENTS

This thesis would not have been possible without the help of many people. Of those, none is more important than my Lord and Savior, Jesus Christ. I want to thank Him for His perpetual love and faithfulness to me and for teaching me that I can trust Him even when I don't understand where He's leading me or why. He truly deserves all the credit and honor for any good work that I have done.

Second, I want to thank my advisor, Dr. Keith A. Teague, who has been instrumental in making this thesis a reality. Thanks for giving me the opportunity to work for you and pursue this area of research. Thanks for believing in me and for having the patience to let all of us temporarily put aside research at times when other pressing things (tests, projects, and finals) had to be done. Thanks again for your patience, support, and humor. I know you'll continue producing quality students who have fond memories of OSU's electrical engineering faculty and the little room called "410."

I want to especially thank my family for all the love, support, and encouragement they have given me. I could not have lasted without the prayer support from all of you, especially Mom. Thank you for uplifting me in prayer and supporting my arms as Aaron and Hur did for Moses. They supported his arms when he was weak so that his hands remained steady until sunset. Bless God, the sun is finally setting on this master's degree, and you have helped me remain steady throughout. Thank you.

Next, I want to thank my church family for their prayers, support, and friendship. They have made my stay here a lot of fun, and they have been an oasis in the dryness and drudgery of what is commonly called schoolwork. Thanks too for all the "real food" and good meals you made for me. There are too many of you to list by name and you have been so good to me that I don't want to exclude any of you. So I'll just say thanks and may God bless you as richly as you've blessed me.

I also want to thank the Department of Defense for funding this research and especially Tina Kohler and Ron Cohn for their input into the EMBE 8.0 kbps coder. They have helped keep all of us motivated and on track throughout. Thanks for encouraging our work and going to bat for us.

Finally, I would like to thank the guys in the lab (Walt, Steve, and Guenter) and the EMBE 2.4 kbps predecessors (Bryce, Walt, and Buddy). Your work was a large stepping stone in the development of the EMBE 8.0 kbps. I also want to acknowledge my research partner, Edward Daniel, who has helped produce the other half of the EMBE 8.0 kbps coder. You have been a pleasure to work with, Ed. I wish you well in your pursuit of a higher degree.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Purpose

This thesis describes an enhanced sinusoidal reconstruction method for Multi-Band Excitation (MBE) based coders used to produce high quality speech at moderate bit rates of approximately 8,000 bits per second. The research and development for this reconstruction was spawned from prior work done on the low bit rate Enhanced Multi-Band Excitation (EMBE) 2.4kbps vocoder developed at Oklahoma State University by Dr. Keith Teague, Walter Andrews, Bryce Leach, and Buddy Walls [1-6].

In recent years, there has been a demand for low bit rate (approximately 2,400bps) coding of speech due to limitations in bandwidth for applications such as hand held cellular phones, low bandwidth/low power radio systems, secure communication systems, and telephone systems. Although current advancements in communication technology have produced more efficient use of bandwidth and, in many cases, the capability of handling higher bit rates, reduction of bit rates, as compared to the standard telephone quality bit rate (64kbps), is still needed. However, emphasis is currently being placed on increasing the quality of speech at the expense of slightly higher bit rates. The tradeoff of bit rate for increased quality is desirable in many applications. Therefore, it is advantageous to design a speech coder which can produce high quality speech at low to moderate bit rates.

MBE-based coders such as EMBE 2.4kbps and Improved Multi-Band Excitation (IMBE) model speech signals well in terms of pitch, voicing, and the magnitude spectrum. However, these coders were designed for low bandwidth, low bit rate applications. To reduce the bandwidth, the frame rate is reduced and phase is synthetically generated in the synthesizer using a linear excitation model with randomized jitter in the upper harmonics. Although the human auditory system is usually considered phase deaf, this synthetic phase model is not sufficient to produce particularly natural tonal quality for voiced and mixed excitation speech.

Therefore, the main goal of this thesis is to develop an enhanced reconstruction for MBE-based coders which makes use of the increased bit rate to improve the tonal quality of reconstructed voiced speech. The addition of a new phase model and a phase-based sinusoidal voiced reconstruction procedure will be central to the improvement of tonal quality. Methods presented in this thesis are based on sinusoidal-based coders such as EMBE, IMBE and STC, along with new enhancements developed specifically for the coder discussed here. Both MBE and STC are important in the development of the enhanced reconstruction and will be discussed in more detail in later chapters.

The enhanced sinusoidal reconstruction method introduced in this thesis has been fully implemented in a test coder to analyze the quality of reconstruction as compared to other speech coders currently available. The test coder used is the EMBE 8.0kbps vocoder [7].

Thesis Outline

The remainder of this thesis details the development of the enhanced sinusoidal reconstruction method. Chapter 2 provides a general background which lays the foundation for speech coding work. It covers briefly the categories of speech coders.

Then, more specifically, the basic speech model used for vocoders is presented. Finally, the two speech coding models used for this research, MBE and STC, are described.

Chapter 3 gives specific background concerning the MBE and STC models. The particular reconstruction methods used by each one are covered in more detail. The benefits and limitations of the reconstruction methods are discussed as an introduction to the work presented in the following chapters.

Chapter 4 describes the topics covered for the development of an enhanced sinusoidal reconstruction method. These include the basic synthesis structure, reintroducing phase as a parameter, smooth frame connection and, finally, parameter smoothing. In particular, Chapter 4 addresses the first topic, the basic synthesis structure, in detail.

Chapter 5 provides a more detailed discussion of the second topic, the reintroduction of phase as a parameter. Several methods of analyzing phase are considered in detail, and an alternate model for generating synthetic phases is presented.

Chapter 6 concludes the development of the synthesizer with an in depth discussion of two methods available for connecting frames. This also leads to a detailed discussion of issues involving parameter connection and smoothing necessary for one of the methods.

Chapter 7 summarizes the design of the enhanced reconstruction procedure and concludes with a brief discussion of the quality of the fully-implemented reconstruction procedure. In addition, potential research topics for further study are discussed.

Basic Speech Model

Before discussing sinusoidal vocoders, it is important to have a fundamental understanding of the speech characteristics utilized by these vocoders. Although the real speech process is quite complex, a sinusoidal model for speech production considers two distinct mechanisms for exciting the vocal tract. "Voiced" speech is produced by excitation of the vocal cords at a fundamental frequency and results in the production of harmonics—integer multiples of the fundamental frequency. Voiced speech may be successfully reproduced using a sum of harmonically related sinusoids weighted by a set of harmonic amplitudes (1.1).

$$s_v(t) = \sum_l A_l(t)\cos(\theta_l(t)) \tag{1.1}$$

"Unvoiced" speech is produced without excitation of the vocal cords and is the result of turbulent air flow caused by a constriction in the vocal tract. Unvoiced speech can be reproduced using band-limited white noise produced either in the frequency domain or the time domain. For example, a simple pseudo-random noise generator can serve as the source, followed by a suitable band-pass filter. Alternatively, a bank of sinusoidal oscillators having random phase can be used [17]. In the case of a sinusoidal representation, each peak in the frequency spectrum (regardless of whether it is harmonically related to a fundamental frequency) is considered to represent an underlying sinusoid. Regardless of which method of representation is used, both voiced and unvoiced speech are weighted by the response of the vocal tract.

Because speech is quasi-stationary, it must be divided into sections, or frames, of short duration for analysis. Many frames exhibit characteristics of both voiced and unvoiced speech. These frames are referred to as having mixed excitation. When analyzed in the frequency domain, some harmonics in these frames may be clearly present while others are not. The frequency ranges where harmonics are not present are noisy. Therefore, speech in such mixed frames can be described by a fundamental frequency with harmonics that are classified as either voiced or unvoiced, all shaped by the vocal tract response. Figure 1 shows three examples of speech spectra, each illustrating the basic types of speech frames—all voiced, all unvoiced, and frames with mixed excitation.

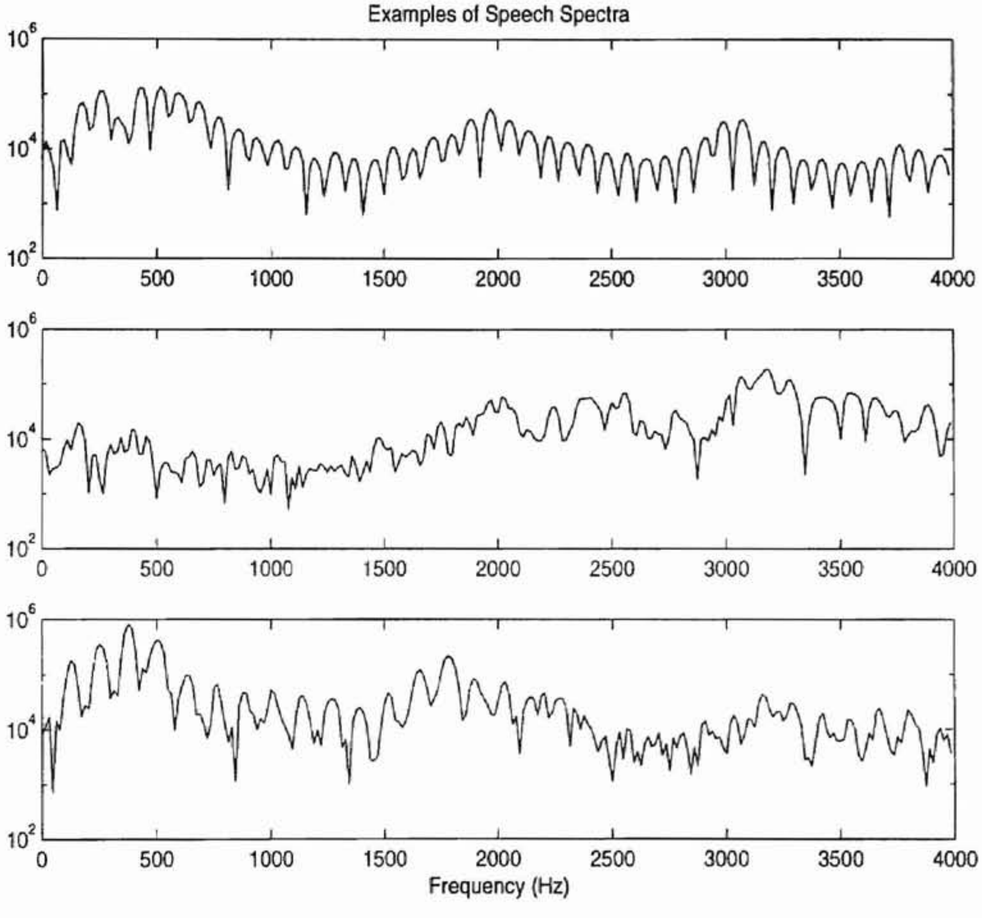


Figure 1. Examples of Speech Spectra

a) Completely voiced spectrum
b) Completely unvoiced spectrum
c) Mixed excitation spectrum

Again, since voiced and unvoiced speech are produced from independent sources, speech can be divided into separate voiced and unvoiced components. In this case, a speech signal is treated as the sum of voiced components produced using a harmonically weighted sum of sinusoids and a sum of unvoiced components produced by band-limited white noise shaped by the vocal tract.

In a more general sense, a speech signal can be treated as the sum of a set of arbitrary sinusoids, located at frequencies having significant amplitude. Those located at harmonics will be related by phase and thus contribute to the voiced components. Sinusoids not located at harmonics of the fundamental frequency will exhibit randomness in their phase and thus contribute to the unvoiced components of the overall speech signal. Such a representation can be thought of as being based on an under-sampled Discrete Fourier Transform (DFT) spectrum.

Sinusoidal Vocoders

Vocoders make maximum use of this simplified model of speech. They use the fundamental models of voiced and unvoiced components to describe the speech and then reconstruct it so that it retains these characteristics. To do so, the basic vocoder is broken into two parts—an analyzer and a synthesizer. The analyzer divides the original speech signal into frames, analyzes each frame to determine the characteristics (parameters) of that frame of speech, encodes and quantizes the parameters for storage or transmission, and sends these to a synthesizer. The synthesizer then decodes these parameters and reconstructs the speech on a frame by frame basis. Sinusoidal vocoders use a sum of sinusoids to model (at the very least) the voiced components of the speech in reconstruction.

This paper will be dealing with two specific types of sinusoidal speech coders. The first is Multi-Band Excitation (MBE), and the second is Sinusoidal Transform Coding (STC). Each uses a unique set of parameters while still reconstructing at least the voiced components of the speech as a sum of sinusoids.

Multi-Band Excitation (MBE)

MBE was developed by Daniel Griffin and Jae Lim in the mid- to late-1980's. MBE makes use of both the harmonic nature of speech and its voiced/unvoiced nature. Although not the first to make use of the harmonic nature of speech, MBE was one of the first to allow for multiple voicing decisions. Prior to the introduction of MBE, harmonic sinusoidal vocoders were usually given only a single voicing decision. The resulting reconstructed speech was often "buzzy," due in part to reconstructing unvoiced parts of mixed excitation frames with periodic sinusoids. MBE sought to eliminate this quality by introducing multiple voicing decisions per frame. For each frame, the spectrum is subdivided into frequency bands and a voicing decision is made for each band.

MBE as developed by Griffin and Lim requires four sets of parameters--pitch, harmonic amplitudes, harmonic phases, and voicing decisions for grouped harmonics [13]. After these four sets of parameters are determined, they are encoded and sent to the synthesizer where the voiced harmonics and unvoiced harmonics are reconstructed independently and summed. A general block diagram of an MBE vocoder is shown in Figure 2.

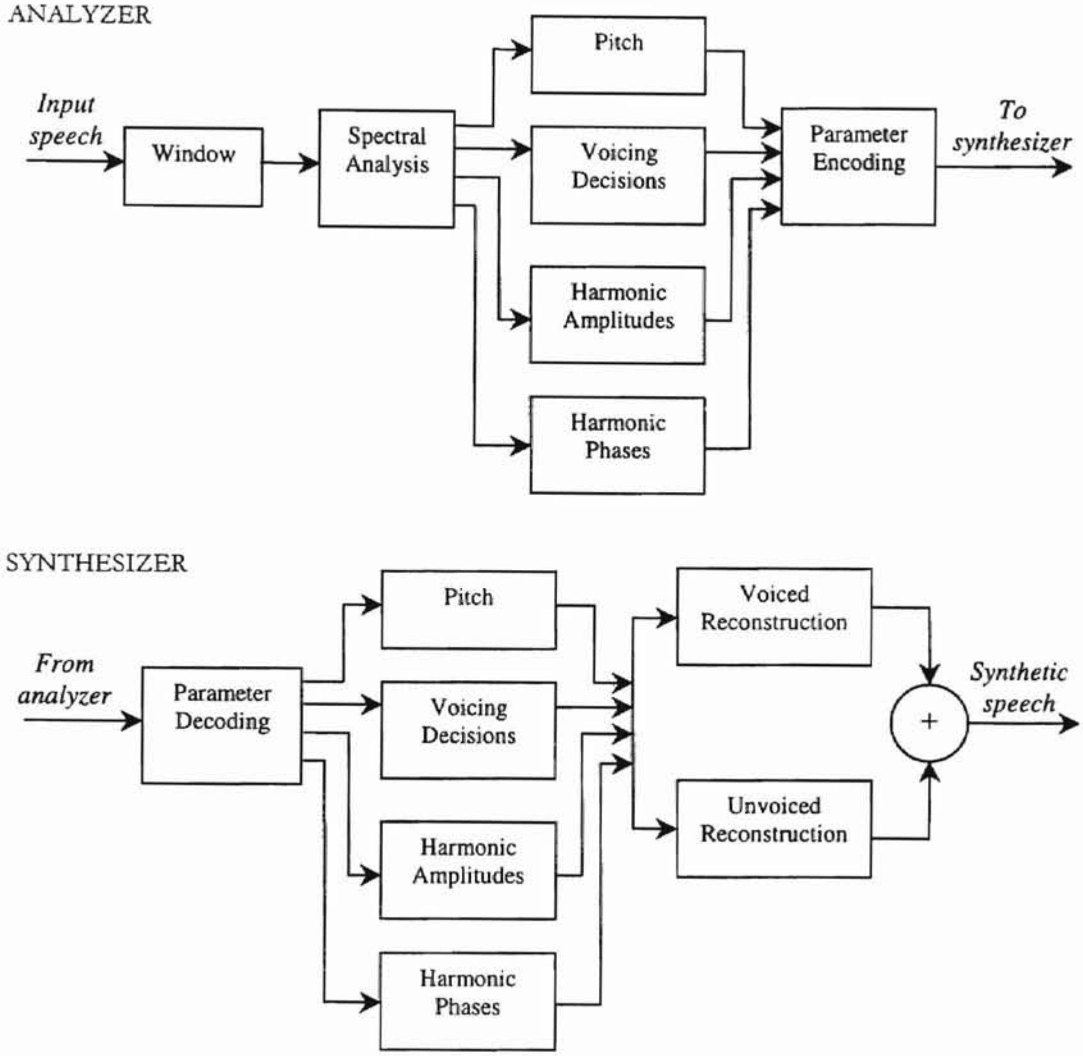


Figure 2. MBE analyzer and synthesizer block diagrams

Voiced bands are built using a sum of sinusoids, but unvoiced bands are built using band-limited white noise. This mixture of permitting voiced and unvoiced speech within a single frame effectively represents mixed excitation speech.

For implementing MBE as a mid-rate coder, Griffin and Lim were able to differentially encode and send only the first 12 harmonic phases. This was due to the limited number of bits and the inability to accurately predict the phases of higher harmonics using a linear excitation phase model [12]. As a result, they generated the upper harmonic phases synthetically in the receiver using the average fundamental frequency of the current and subsequent frames and, presumably, a linear excitation phase. The lack of phase information for synthesis was acknowledged as a contributing factor to the quality degradation of their mid-rate implementation.

Sinusoidal Transform Coding (STC)

During the approximate time period that MBE was being developed by Griffin and Lim, STC was being developed by Robert McAulay and Thomas Quatieri. Rather than using a harmonic model, STC approaches sinusoidal reconstruction by using a non-harmonic model based on arbitrarily located sinusoids. STC is based on peak picking in the frequency domain. Each peak in the spectrum is assumed to represent an underlying sinusoidal component with an associated frequency, amplitude, and phase. Figure 3 illustrates spectral peak picking as performed in STC analysis. Usually, only a fixed number of peaks (marked by the *x*'s in Figure 3) is selected.

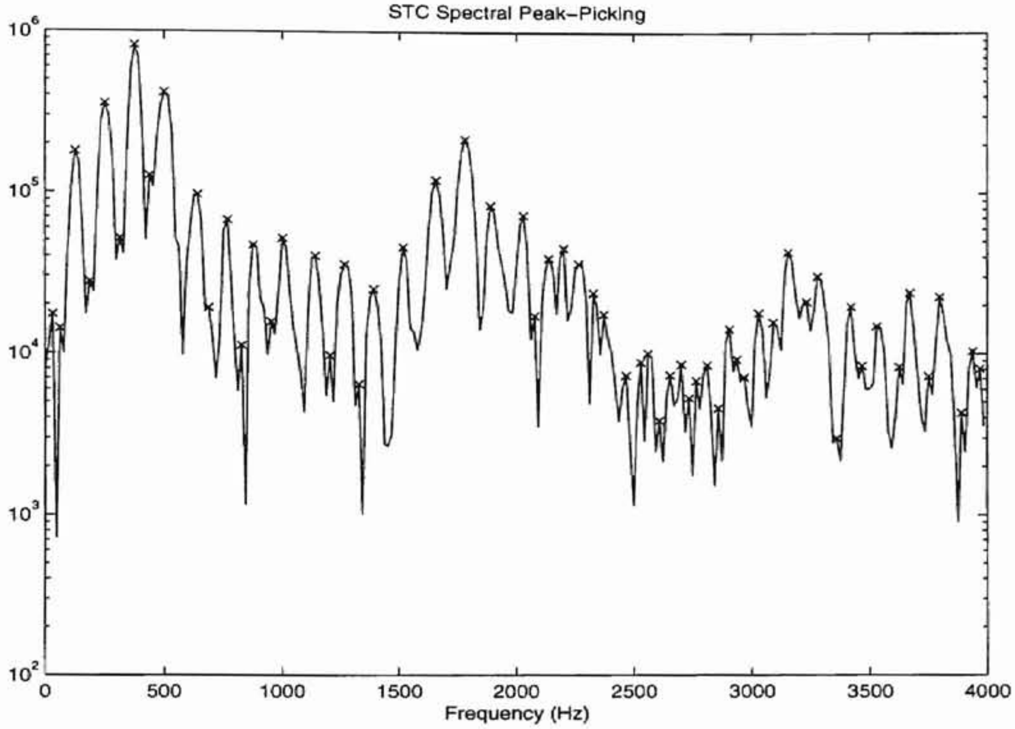


Figure 3. STC spectral peak-picking

The corresponding amplitudes, phases, and peak locations are encoded and sent to the synthesizer where they are decoded and used to reconstruct the speech using Equation (1.1) summed over the number of peaks. Figure 4 shows the basic STC analyzer and synthesizer.

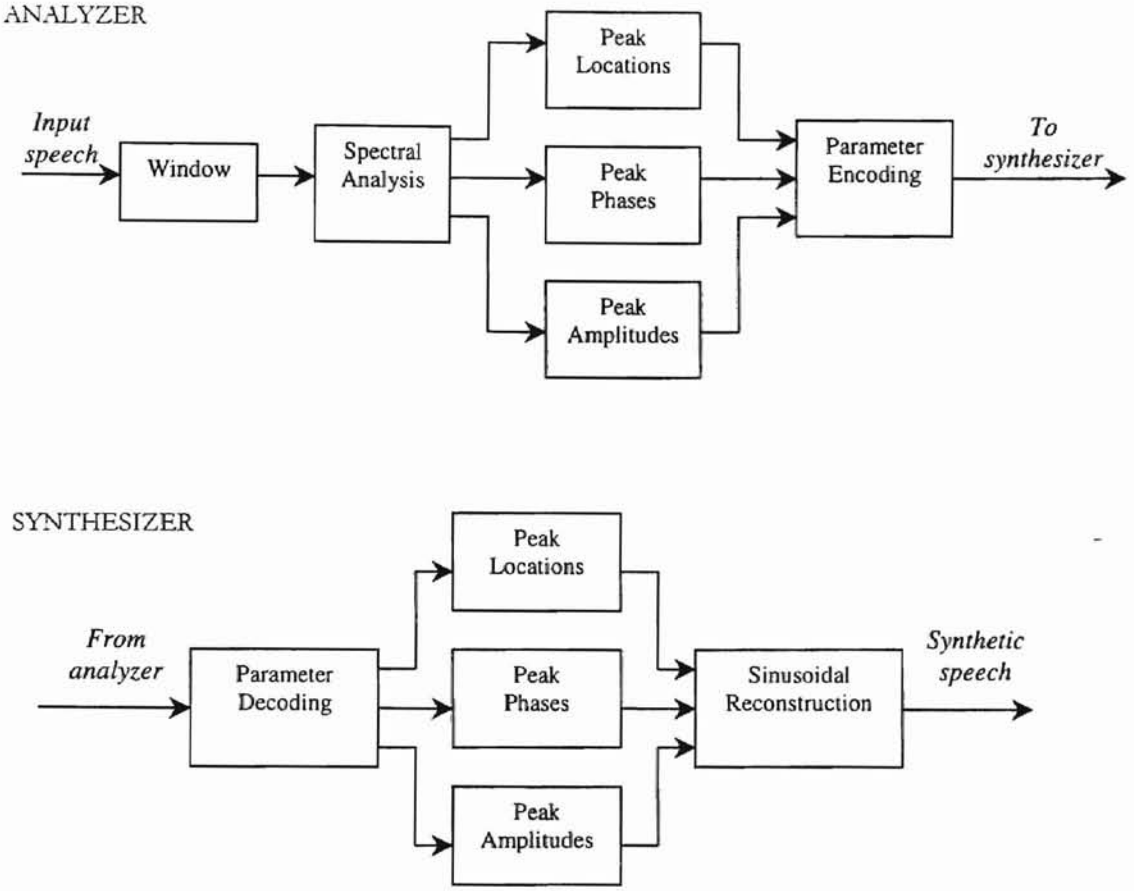


Figure 4. STC analyzer and synthesizer block diagrams

STC can be thought of as a generalized case of MBE. For voiced speech, MBE is only concerned with peaks that occur at a harmonic relation to the fundamental frequency. STC doesn't make that distinction and, as a result, is capable of representing arbitrary signals which aren't composed of harmonically related components. Both models are capable of representing speech with excellent intelligibility and tonal naturalness.

CHAPTER III

SPECIFIC BACKGROUND

MBE and STC Synthesis

Although parameters for both MBE and STC are sent for every frame of speech, these parameters must be connected smoothly from frame to frame so that the reconstructed speech has no discontinuities at frame boundaries. The reconstructed speech must evolve smoothly from frame to frame as it did in the original speech signal. Each harmonic or peak has its own set of parameters, and each must be smoothly connected to the parameters of the corresponding harmonic or peak in the next frame. In order to smoothly connect the parameters, a criterion for matching harmonics or peaks from frame to frame must be established. Thus, deciding which harmonics or peaks to connect from frame to frame for sinusoidal reconstruction becomes an important issue.

In STC, each frequency peak in a frame is assigned to a frequency track. The frequency of each track is matched to the peak with the closest frequency in the subsequent frame. If the current frequency track has no future match, then that frequency track must "die." Its amplitude in the future frame will be assigned a value of zero. If a future frequency track has no current match, then that frequency track must be "born." Its amplitude in the current frame will be assigned zero [17]. Figure 5 shows a simplified version of track matching with the three main cases of track birth, matching, and death illustrated. ω is the frequency in radians, k is the frame number, n is the number of the harmonic in the current frame, and m is the number of the harmonic in the next frame.

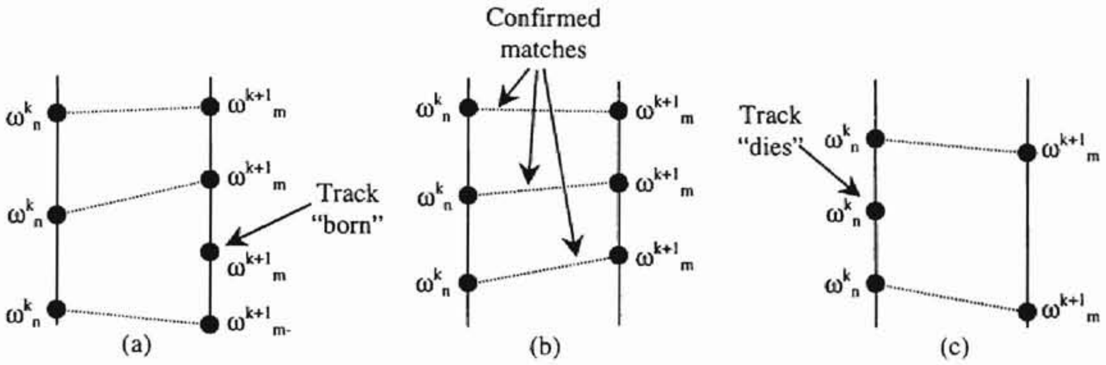


Figure 5. STC track matching (generalized)

a) Birth of an unmatched track
b) Confirmed track matches
c) Death of an unmatched track

MBE uses a simpler method of matching voiced harmonics. Voiced harmonics are matched by harmonic number unless the frequency change is greater than a threshold, usually about ten percent. In that case, the voiced harmonics are no longer matched, but are treated as if they are preceded and followed by unvoiced harmonics, similar to the birth and death cases of STC's frequency matching, respectively [13]. All unvoiced harmonics are assigned an amplitude of zero during voiced reconstruction so they will not be reconstructed using sinusoids.

Once the criterion has been established for matching harmonics or peaks, the set of parameters for each harmonic or peak must be smoothly interpolated between frames. MBE and STC vary in their interpolation methods, but the same set of three parameters is involved in representing the sinusoidal components. The first parameter is amplitude, the second is frequency, and the third is phase. Only on amplitude do both MBE and STC use somewhat similar interpolation schemes.

Both STC and MBE use linear amplitude interpolation in the time domain. Since STC builds all speech using sinusoidal reconstruction in the time domain, linear amplitude interpolation is used to smoothly connect the amplitudes of all peaks between frames [17]. However, in MBE, the linear amplitude interpolation previously described is used on voiced harmonics only. Unvoiced components are built in the frequency domain using a noise source typically based on band-limited white noise. The unvoiced components are then smoothly connected in the time domain using a weighted overlap addition method [13]. Equation (3.1) denotes triangularly tapered overlap addition for unvoiced reconstruction, where k is the frame number, t is the time in seconds, and T is the frame shift in seconds.

$$s_{uv}(t) = s_{uv}^{k}(t)\frac{(T-t)}{T} + s_{uv}^{k+1}(t-T)\frac{t}{T} \qquad t = [0:T] \tag{3.1}$$

After overlap addition, no further parameter interpolation for the unvoiced components is needed. Figure 6 shows a block diagram of typical MBE-based unvoiced synthesis.

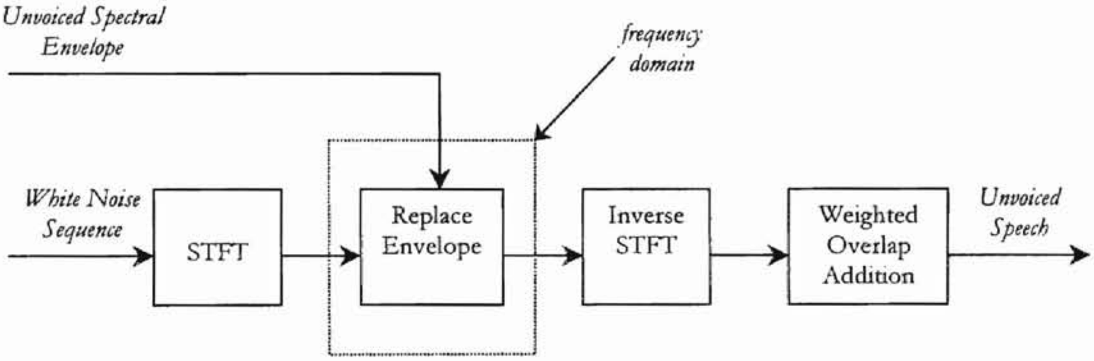


Figure 6. MBE unvoiced reconstruction block diagram

Phase and frequency interpolation for voiced harmonics are not as simple as amplitude interpolation for voiced harmonics. Because phase is the integral of frequency, phase and frequency are related. MBE assumes voiced harmonics exhibit a linear frequency variation in time. The phase is allowed to vary quadratically such that the frequencies and phases at the beginning and ending of each frame match the original frequencies and phases sent by the analyzer [13]. Equation (3.2) denotes the linear frequency interpolation in time for voiced harmonics, and Equation (3.3) denotes the resulting quadratic phase interpolation.

$$\omega_l(t) = \omega_l^k \frac{(T-t)}{T} + \omega_l^{k+1} \frac{t}{T} + \Delta\omega_l \tag{3.2}$$

$$\theta_l(t) = \int_0^t \omega_l(\xi) d\xi + \phi_0 \tag{3.3}$$

Note that l is the harmonic number, k is the frame number, t is the time in seconds, and T is the frame shift in seconds. $\Delta\omega$ and ϕ_0 are the frequency deviation and initial phase, respectively, such that the phase and frequency parameters match the measured phases and frequencies at frame boundaries. Again, since MBE treats unvoiced components differently, frequency and phase interpolation in time do not apply to the unvoiced reconstruction.

MBE-based coders such as IMBE and EMBE 2.4kbps do not send phase at all in order to further reduce the bit rate; phase is synthesized entirely in the receiver. Both IMBE and EMBE assume zero starting phase for voiced harmonics and generate subsequent phases by tracking the phase in time using a quadratic phase model. Within a frame, the phases for the voiced harmonics are calculated using a linear excitation phase model. The excitation phase is calculated using (3.4) where k is the frame number, ω is the pitch in radians, S is the number of samples per frame shift, and f_s is the sampling frequency. Equation (3.5) denotes the excitation phase for each harmonic, where l is the harmonic number.

$$\phi_0(k) = \phi_0(k-1) + \frac{\left(\omega_0^{k-1} + \omega_0^k\right)}{2}\left(\frac{S}{f_s}\right) \tag{3.4}$$

$$\phi_l^k = l\phi_0(k) \tag{3.5}$$

In IMBE, random jitter is added to the upper three-fourths of the phases of voiced harmonics to produce more natural sounding speech [16]. This is shown in Equation (3.6). Note that L_{uv} denotes the number of unvoiced "harmonics" in the frame, L is the total number of harmonics in the frame, and ρ is a random number in the interval [-π, π).

$$\theta_l^k = \begin{cases} \phi_l^k & for\,1 \le l \le \dfrac{L}{4} \\ \phi_l^k + \dfrac{L_{uv}\rho_l^k}{L} & for\,\dfrac{L}{4} < l \le L \end{cases} \tag{3.6}$$

STC treats frequency and phase differently than typical MBE models do; STC assumes cubic phase variation in time. The frequency of each sinusoidal component is allowed to vary quadratically such that the frequencies and phases at the beginning and ending of each frame match the original frequencies and phases sent [17]. Equation (3.7) denotes cubic phase interpolation in time as used by STC.

$$\theta^k(t) = \theta^k + \omega^k t + \alpha(M)t^2 + \beta(M)t^3 \tag{3.7}$$

Increasing the phase order from quadratic to cubic results in more freedom of variation for the sinusoidal frequency tracks. In addition, as shown in Figure 7, there is no longer a unique solution for the phase track from frame to frame. Finding the path with the least frequency variation (finding the optimal M and calculating $\alpha(M)$ and $\beta(M)$ for each track) requires a significant increase in computation.

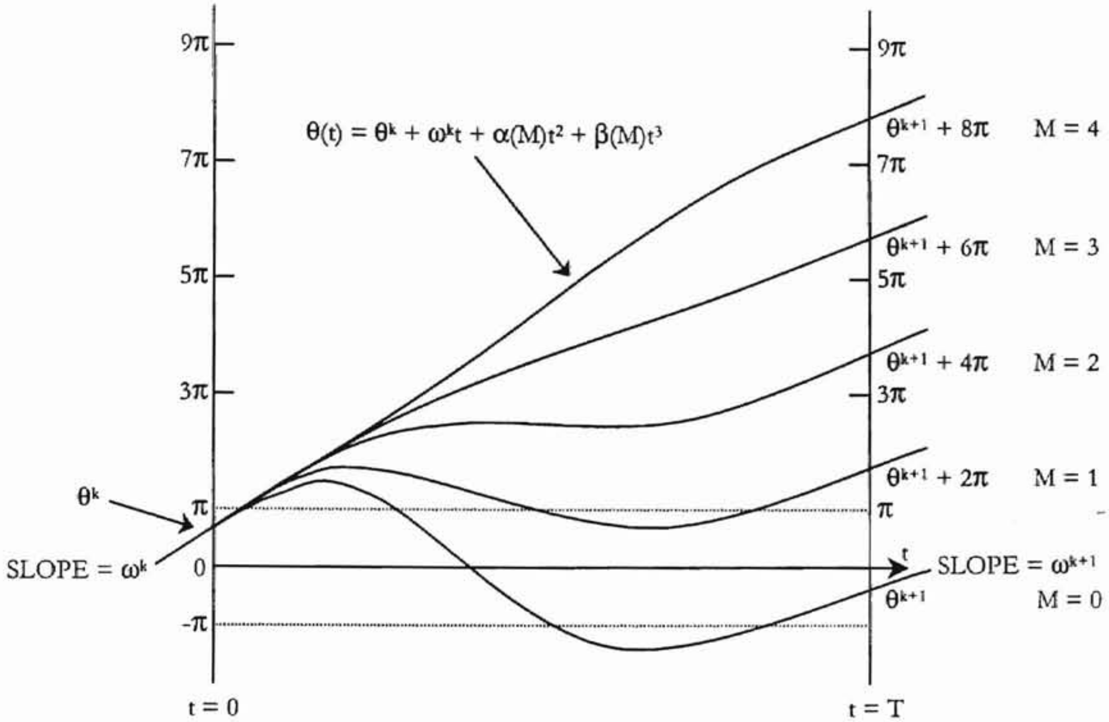


Figure 7. STC phase trajectories

Limitations of MBE and STC

MBE and STC have certain inherent strengths and weaknesses. MBE makes more efficient use of speech qualities such as harmonic relationships and voiced/unvoiced characteristics to reduce the amount of information that must be transmitted. Because the MBE sinusoidal reconstruction is based on frequency rather than phase, the phases can be generated synthetically to reduce the bit rate. This is the method used by IMBE, EMBE, and the mid-rate implementation of MBE by Griffin and Lim. As noted before, for IMBE the synthetic phase model is based on the excitation with added phase jitter at higher harmonics. Although this produces very intelligible speech, the speech often has a slightly synthetic, reverberant tone.

Sending both phase and peak location information for each sinusoidal component, as is done in STC, results in very high quality reconstruction. The reconstructed signal can be virtually indistinguishable from the original. STC is also extremely flexible in that signals without a harmonic structure can be successfully reconstructed. However, this is at the expense of significantly increasing the bit rate. Sending peak locations for highly periodic speech is very redundant and unnecessarily costly in bits. For this reason, mid-rate and low-rate implementations of STC typically assume harmonically related components for voiced speech. Under this constraint, the MBE and STC models are quite similar.

Thus, the MBE and STC methods of reconstruction are well established as high quality methods of reconstructing speech using sinusoidal reconstruction. However, both methods have weaknesses that leave room for improvement. The synthetic phase model used for lower bit rate MBE-based coders is not sufficient to produce high quality speech with natural tonal qualities. The STC model, although it fully incorporates measured phases and is capable of producing high quality speech with natural tonal qualities, requires sending too much information. As a result, the full STC model is not practical for low to moderate bit rate speech coding. Even the MBE model with all phases encoded is not practical for low to moderate bit rate speech coding. Therefore, there is a need to develop an enhanced sinusoidal reconstruction method aimed at low to moderate bit rate coders which utilizes the efficient speech modeling of MBE coders while reincorporating phase such that the tonal naturalness of coders which use measured phase is restored.

The development of a new, enhanced sinusoidal reconstruction method for high quality mid-rate speech coding is the topic of the rest of this paper. The next chapter will begin with the discussion of this development.

CHAPTER IV

OVERVIEW OF THE RECONSTRUCTION DEVELOPMENT

Overview of Topics

The development of a reconstruction method for any type of speech coder involves many topics. What underlying models will be used for voiced and unvoiced components? Will the models assume a harmonic structure like MBE or a non-harmonic structure like STC? What parameters are necessary for the models? Are new parameters introduced that must be analyzed and sent in the analyzer? How will all the parameters be smoothly connected between frames?

The questions are numerous. However, the development of the enhanced sinusoidal reconstruction approach can be broken down into three specific topics. These include i) the basic synthesis structure, ii) the analysis of any new parameters not included in the target coder model (MBE, in this case), and iii) the smooth connection of parameters from frame to frame.

Before any other topics can be discussed, the basic synthesis structure must be addressed. A decision must be made concerning whether to choose a harmonic structure like MBE-based coders use, or a non-harmonic structure like STC coders use. Assuming the simplified speech model of voiced and unvoiced components discussed in Chapter 2, choosing a harmonic structure such as the one MBE uses represents voiced components well, but does not represent unvoiced components well. Thus, the use of a harmonic structure raises the question of how to deal with speech containing unvoiced non-harmonically related components. Therefore, the end of this chapter deals not only with the basic structure, but also how to model speech components which are not modeled well by a harmonic structure.

Once the structure is established, the analysis of any parameter not included in typical MBE-based coders must be discussed. The analysis of pitch, harmonic amplitudes, and voicing decisions is already well-established in MBE-based literature. For specific information on these topics, see [21-28] for pitch, [27-28] for voicing, and [29-30] for spectral analysis. Analysis information for specific MBE-based coders is also provided in [1-3], [12-13], and [15-16]. Therefore, in this paper, it is assumed that these parameters are estimated accurately in the analyzer and are available for use in the synthesizer.

Although pitch, amplitude, and voicing are well covered in other literature, MBE-based vocoders largely ignore phase, choosing to generate it synthetically as a byproduct of pitch. As mentioned in Chapter 3, it appears that the phase models used by IMBE and EMBE are not sufficient to retain tonal naturalness in reconstructed speech. Therefore, the second topic of discussion is the reintroduction of phase as a parameter, covered in Chapter 5. Since the reintroduction of phase is central to the efforts to improve the reconstructed speech quality, the analysis of phase and an alternate form for generating natural-sounding synthetic phases will be discussed in depth.

After the synthesis structure and reintroduction of phase are established, the third and last topic to consider is how to smoothly connect all of the parameters from frame to frame. The frame-to-frame connection of parameters is a broad topic and will be divided into several subtopics in Chapter 6. The first is choosing whether to use overlapping or non-overlapping frames. Depending on this decision, further parameter smoothing subtopics follow. These include deciding which harmonics to connect from frame to frame and evaluating methods to smoothly connect the harmonic amplitudes, frequencies, and phases in time so that the resulting speech has no discontinuities at frame boundaries.

Indeed, the synthesis of speech is very involved. Now that a general overview has been given, the first of the three main topics, the basic synthesis structure, will be discussed in detail.

The Basic Synthesis Structure: Harmonic vs. Non-Harmonic

The structural model is the basis for synthesis. The structural model will determine what parameters are needed, how the parameters will be matched, and how the voiced and unvoiced components will be reconstructed.

There are two basic sinusoidal models to consider—a harmonic model and a non-harmonic model. The harmonic model, typically used in MBE-based coders, assumes that speech has a fundamental excitation frequency (pitch) with harmonics located at integer multiples of the pitch. It represents voiced components very well, but does not accurately represent unvoiced non-harmonically related unvoiced components. A non-harmonic model, such as the one used in STC, does not assume any fundamental excitation frequency. Instead, reconstruction is based on the locations of peaks in the spectrum, as discussed in Chapter 3. The STC model represents both voiced and unvoiced speech well, but models unvoiced speech more efficiently than voiced speech.

Ideally, using different models for different types of speech would be best. Voiced components could be modeled using the harmonic model and unvoiced components could be modeled using the non-harmonic model. However, because any use of a non-harmonic model requires sending the locations of peaks, using different models for different types of speech is not feasible for mid-rate speech coding. Using a harmonic model for the basic reconstruction model is the most realistic choice. In fact, a harmonic model is used not only in MBE-based vocoders such as IMBE and EMBE, but also in the 8kbps STC coder [18]. In the latter case, a harmonic model was required in order to reduce the bit rate to a moderate level.

Reconstruction of Non-Harmonic Components

The selection of a harmonic model introduces the problem of how to reconstruct non-harmonically related speech components. Unvoiced speech and unvoiced components in frames of speech with mixed excitation are not harmonically related to an underlying fundamental frequency. Hence, applying a harmonic model in such cases does not accurately represent the speech.

Although unvoiced components are not harmonically related, unvoiced areas can be reconstructed using a sinusoidal harmonic model if the phases of the sinusoids are randomized. This has been shown to work well if the frame update rate is at least once per 12.5ms so that the phases are randomized frequently [18] and the frequency spacing between unvoiced sinusoidal components is sufficiently small such that the Karhunen-Loeve expansion for noise-like signals [17] is satisfied. This requires that sinusoids used to reconstruct noisy areas be at most about 100 Hz apart in frequency. However, this presents another problem.

To use a sinusoidal harmonic model to reproduce unvoiced components in a frame with mixed excitation, the fundamental frequency must be less than or equal to 100

Hz to fulfill the frequency spacing requirements. In practice, this is rarely the case since the pitch of typical speakers ranges from 70Hz for a low-pitched male to 400Hz for a high-pitched female. The average pitch for most speakers is above 100 Hz. To overcome the insufficient frequency spacing problem, unvoiced areas can be resampled in the synthesizer. Specifically, the spacing of sinusoids for voiced bands will be determined by the fundamental frequency, while the spacing of sinusoids for unvoiced bands will be determined such that the frequency spacing is appropriate.

To resample the unvoiced components, a smooth curve is fitted through the spectral amplitudes of the harmonics so that the spectrum can be resampled at frequency locations in between the harmonics. The amplitude spectrum is then evenly resampled in unvoiced areas so that the sinusoids representing the unvoiced areas are not more than 100 Hz apart in frequency. The amplitudes of the resampled sinusoids must be rescaled by the ratio of the previous number of harmonic sinusoids in the unvoiced band to the number of new resampled sinusoids in the unvoiced band. This preserves the proper energy ratio between the voiced and unvoiced components in frames with resampled areas. Equations (4.1) and (4.2) reflect this change. Equation (4.1) is the sum of sinusoids from (1.1) with *l* restricted to be voiced harmonics only. Equation (4.2) is the rescaled sum of sinusoids in the unvoiced areas. Note that *N*(*b*) is the previous number of harmonic sinusoids in the unvoiced band, and *M*(*b*) is the number of new resampled sinusoids in the unvoiced band.

$$s_v(t) = \sum_l A_l(t)\cos(\theta_l(t)) \qquad l = voiced\ harmonic \tag{4.1}$$

$$s_{uv}(t) = \frac{N(b)}{M(b)}\sum_m A_m(t)\cos(\theta_m(t)) \tag{4.2}$$

As an alternative to using a sinusoidal model to reconstruct unvoiced components, MBE diverges from sinusoidal harmonic reconstruction completely to reproduce unvoiced components. Rather than using a sum of harmonically related sinusoids with random phases, unvoiced components are independently reconstructed using band-limited white noise shaped by the vocal tract response. Since unvoiced components and voiced components can be reconstructed separately and then summed together to produce the final synthetic speech, unvoiced components may be reconstructed using this method without altering the voiced sinusoidal harmonic reconstruction. However, using band-limited white noise to reconstruct the unvoiced components requires two completely separate reconstruction techniques, whereas using all sinusoidal reconstruction requires using only the sum of sinusoids given in (4.1) and (4.2).

Comparison of Unvoiced Reconstruction Methods

To perform a comparative study of the quality of synthesis resulting from the use of different types of reconstruction for unvoiced components, informal tests were run. The different reconstruction methods were implemented in the synthesizer of a fully functional EMBE-based speech coder. Speech files containing both male and female speakers in quiet environments were processed through the coder using different unvoiced reconstruction methods. The analysis method was held constant in each case. Three to four people were asked to listen to the reconstructed speech and give their opinion as to which version of reconstructed speech they preferred and why.

Overall, the listeners preferred completely unvoiced speech reconstructed using band-limited white noise over completely unvoiced speech reconstructed using resampled sinusoids with randomized phases. The listeners all agreed that the unvoiced speech reconstructed using sinusoids had a slightly buzzy quality that the unvoiced speech reconstructed using band-limited white noise did not have. However, the same listeners felt speech containing both voiced and unvoiced components sounded more natural when reconstructed using resampled sinusoids with randomized phases for unvoiced harmonics. Speech with mixed excitation sounded like it had a superimposed "whisper" in the reconstructed speech when the unvoiced components were built using band-limited white noise. Part of the latter result may be due to limitations in voicing decisions inherent to MBE analysis.

As a result of this informal testing of the unvoiced reconstruction methods, the following combination of the two unvoiced reconstruction methods was chosen:

1. Completely unvoiced frames are reproduced using band-limited white noise as is done in MBE.
2. Frames with any voiced components are reproduced using a sinusoidal harmonic model with randomized phases for unvoiced sinusoids. Resampling in unvoiced areas is performed so that the spacing between sinusoids in unvoiced areas is less than 100 Hz to ensure the most natural sounding unvoiced components.

Regardless of what components (voiced or unvoiced) make up the frame of speech, a harmonic model is used. A harmonic model is used even in the case of completely unvoiced frames (a minimum pitch is assigned), but this "pitch" is used only to provide dense sampling of the spectrum. In unvoiced cases, the pitch does not represent a true harmonic structure. Therefore, the selection of a harmonic model as the basic structure for synthesis results in the need for four sets of parameters. First, a fundamental frequency is required to serve as the basis for the harmonics. Second, the harmonics must be assigned voicing decisions. Each voiced harmonic must then have a corresponding amplitude and phase. For the synthesis of entirely unvoiced frames, only the corresponding amplitudes of the "harmonics" of a minimum pitch are required for representing the vocal tract response.

Accurate pitch, harmonic amplitude, and voicing analysis are already a large part of IMBE and EMBE analysis. Thus these parts of the analyzer will not be discussed. However, phase analysis is not discussed in IMBE and EMBE literature and is only mentioned briefly in MBE. This is because phase has previously been considered secondary to other parameters and, in the case of IMBE and EMBE, it is generated synthetically during reconstruction—it is not analyzed at all during analysis. Therefore, the next chapter is dedicated to the discussion of the reintroduction of phase as a parameter.

CHAPTER V

THE REINTRODUCTION OF PHASE AS A PARAMETER

Phase Analysis

With IMBE, EMBE, and the mid-rate implementation of MBE, it appears that the use of a synthetic phase model based only on the linear excitation phase with randomized jitter (3.4-6) results in loss of tonal naturalness for voiced areas. This phase model does not seem to accurately represent the relationship between harmonic phases. Since the goal of designing a new synthesis approach is to enhance the overall quality of MBE-based vocoders, especially the tonal quality, incorporating the measured phase back into voiced reconstruction is an intuitive first step if the desire for a moderate bit rate can be temporarily ignored.

In order to use measured phases in reconstruction, the harmonic phases must first be analyzed and coded in the transmitter. In MBE and STC, analysis of spectral parameters is performed in the frequency domain using the short-time Fourier transform (STFT) of the framed speech. Equation (5.1) below denotes the STFT [18].

$$S(\omega)=\frac{1}{N+1}\sum_{n=-N/2}^{N/2}s(n)\exp(-jn\omega) \tag{5.1}$$

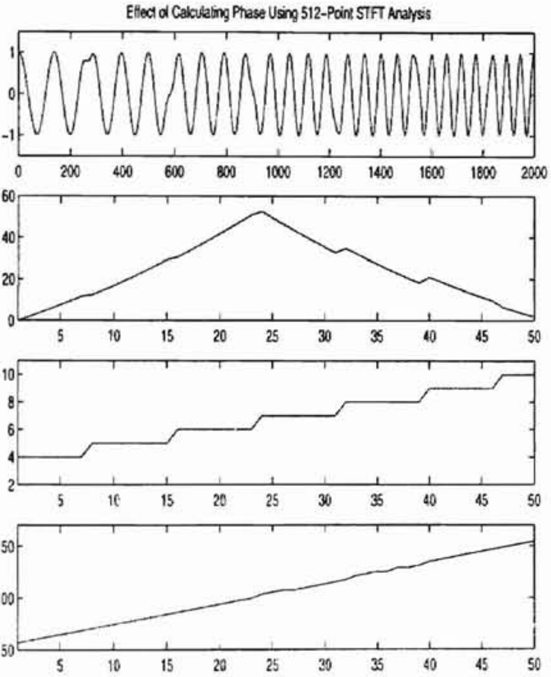
The length (N) of the STFT is an issue critical to parameter estimation. Assuming a fixed frame length, M, then $N \geq M$. Points outside the frame length (from M to N-1) are assigned values of zero (zero padded). Zero padding does not alter the STFT results. Rather, it allows for longer STFT's which improve the frequency resolution. This is very desirable for accurate analysis of frequency domain characteristics. However, due to the computational complexity of the STFT, any increase in length also increases the number of calculations required to perform the analysis. The increased computation makes real-time applications more difficult to realize. A tradeoff must be reached between the feasibility of long STFTs for real applications and the need for increased frequency resolution. Typically the length of the STFT for analysis is limited to about 512 points. This results in a frequency resolution of 15.625 Hz per STFT sample with acceptable computational complexity for real-time applications.

Phases may be calculated from the real and imaginary parts of the STFT using (5.2) [31]. Unfortunately, the fundamental frequency and its harmonics rarely occur at integer indices of the STFT. The phase must generally be determined at arbitrary frequencies which correspond to locations between STFT indices. Assuming that we are limited to a STFT of length 512, there are three methods considered to solve this problem. The first is computing the phase using the closest integer STFT index. The second is interpolating the phase between STFT indices, and the third is interpolating the complex spectrum between STFT indices and calculating the phase from the complex interpolation.

$$\theta(\omega) = ARG[S(\omega)] = \arctan\left[\frac{S_I(\omega)}{S_R(\omega)}\right] \tag{5.2}$$

Method 1: Computing the Phase from the Closest Integer STFT Index

The easiest and most direct solution is to compute the phase from the closest integer STFT index. Seldom is the easiest solution the best, however. When using a 512-point STFT, this method results in audible roughness in the reconstruction every time the closest integer goes from being rounded down to rounded up or vice-versa. Figure 8 shows an example of this roughness using a MATLAB simulation of McAulay and Quatieri's original STC reconstruction method. A single linearly chirped sinusoid was analyzed frame by frame. The frequency was estimated from the magnitude spectrum by cubically interpolating the 512-point STFT magnitude spectrum to simulate a 4096-point magnitude spectrum and picking the peak from the simulated 4096-point spectrum. The phase of the closest integer 512-point STFT index was used. The errors at the index skips are clearly visible in the reconstructed signal. Severe deviations in the waveform have occurred, although the waveform is not actually discontinuous.



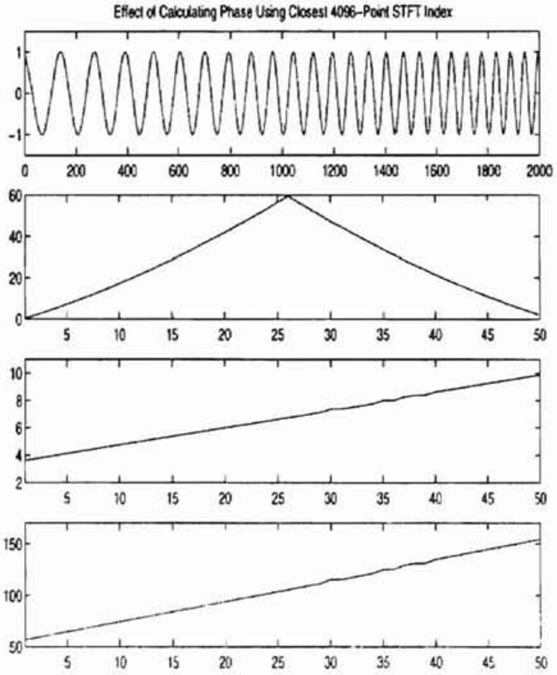


Figure 8. Using 512-point STFT **Figure 9. Using 4096-point STFT**

Effect of calculating phase from closest integer STFT index

a) Reconstructed sinusoid at 8000kHz sampling rate.
b) Unwrapped phase (in radians) versus frame number.
c) 512-point index number selected for phase estimate versus frame number.
d) Frequency (in Hertz) versus frame number. (512-point STFT frequency estimated from peak of 512-point STFT cubically interpolated to simulate 4096-point STFT spectrum.)

Figure 9 shows the same simulation using a 4096-point STFT. The reconstructed chirped sinusoid no longer has the severe deviations that were seen in Figure 8. Both Figure 8 and 9 have approximately equal frequencies, but their phases are different. Therefore, the frequency has not caused the severe deviations—phase has. Although Figure 9 no longer has visible waveform deviations at the index skips, the skips are still audible although much less severe than in the 512-point STFT illustration. This results in what can be described subjectively as "rough"-sounding reconstructed speech.

Method 2: Interpolation of the Phase Between STFT Indices

Obviously, using the closest integer STFT index to compute the phase when using a 512-point STFT is not sufficient to produce high quality results. The next solution is to interpolate the phase between the STFT indices. This may not seem intuitively difficult, but it rapidly becomes an arduous task.

The phases obtained from (5.2) are ambiguous, i.e., the principle phase is obtained rather than the continuous phase, denoted by $\arg[S(\omega)]$. Principle phases are limited to the range $[-\pi, \pi)$. This is referred to as being "wrapped" in the range $[-\pi, \pi)$. Therefore, the phase response will have discontinuous jumps of 2π radians. The continuous phase between consecutive indices may travel outside $[-\pi, \pi)$ on the unit circle, but it is "wrapped" back to this range [31]. Therefore, to properly interpolate between consecutive phases using (5.2), the phase must first be "unwrapped" to simulate continuous phase. Phase unwrapping can be performed using the relationship described by Equation (5.3) [32]. Note that θ_w^n is the wrapped phase at index number n.

$$\theta_{uw}^{n} = \begin{cases} \theta_{w}^{n} + 2\pi & \theta_{w}^{n} - \theta_{w}^{n-1} < -\pi \\ \theta_{w}^{n} - 2\pi & \theta_{w}^{n} - \theta_{w}^{n-1} > \pi \\ \theta_{w}^{n} & \left|\theta_{w}^{n} - \theta_{w}^{n-1}\right| < \pi \end{cases} \tag{5.3}$$

Accurate interpolation between unwrapped phases requires that phase aliasing is not occurring. Phase aliasing occurs when the frequency resolution is greater than half the frequency at which the phase is rotating through 2π radians. For example, if the phase is rotating an average of 2π radians every 40 Hz in the spectrum, the frequency resolution of the spectrum must be at most 20 Hz to prevent phase aliasing. If this requirement is not met, the spectrum is essentially under-sampled, and the "decimated" phases between spectral indices cannot be recovered through interpolation.

The maximum frequency resolution can be calculated using Equations (5.4-5). Equation (5.4) denotes the phase slope, m_ϕ (radians/Hz), where f_r is the STFT frequency resolution (Hz), and n and m are STFT indices. Note that calculating m_ϕ requires knowing the continuous, or unwrapped, phases. Using the phase slope from (5.4), the maximum frequency resolution, f_R (Hz), required to prevent phase aliasing is calculated using Equation (5.5).

$$m_{\phi} = \frac{\arg\left[S\left(\omega_{n}\right)\right] - \arg\left[S\left(\omega_{m}\right)\right]}{\left(f_{r}\right)\left(n - m\right)} \tag{5.4}$$

$$f_{R} = \left(\frac{2\pi}{\left|m_{\phi}\right|}\right)\frac{1}{2} \tag{5.5}$$

As an example of how phase aliasing affects interpolation using unwrapped phases, let's look at two specific examples using Figure 10 as a reference. Figure 10 was generated from the phase calculations of the STFT of a voiced frame of speech using

(5.2). Figure 10a shows principle phases calculated using an 8192-point STFT in gray, and principle phases calculated using a 512-point STFT in black. From the calculated phases of the 8192-point STFT, the phases appear to be nearly linear in nature although there are obvious exceptions (near index 10.5). Therefore, we will linearly interpolate the phase at the 512-point STFT indices 2.5 and 6.5—two areas where the phase appears to be approximately linear. If phase aliasing does not occur, we should obtain values approximately equal to the corresponding 8192-point STFT phases.

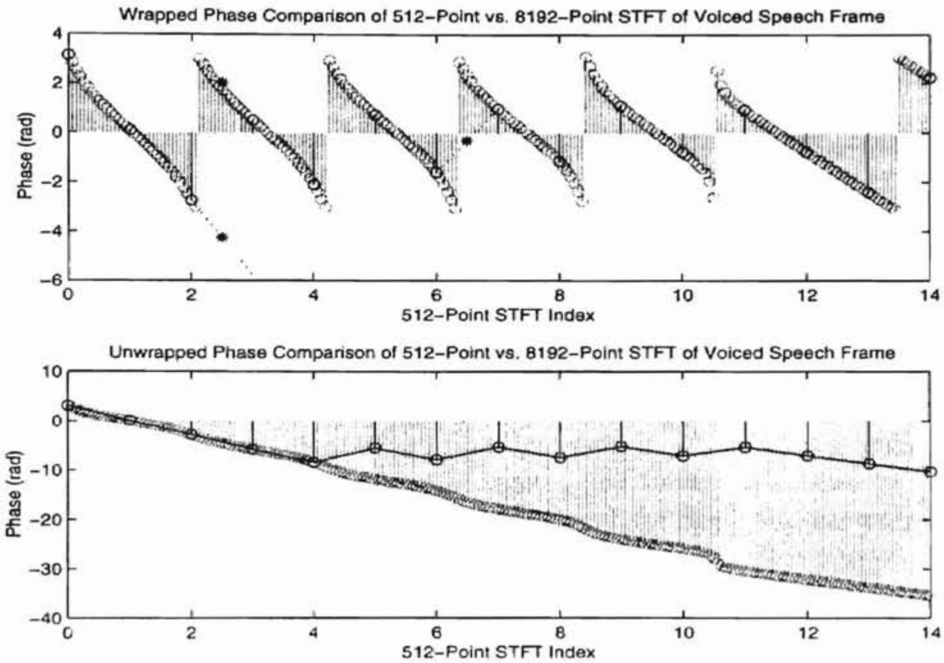


Figure 10. Phase interpolation example using a 512-point STFT

a) Wrapped phase comparison of a 512-point vs. 8192-point STFT of a voiced frame of speech.
b) Unwrapped phase comparison of a 512-point vs. 8192-point STFT of a voiced frame of speech.

The phase at index 2 is -2.74 radians, and the phase at index 3 is 0.51 radians. The difference between the two is 3.25 radians. Using (5.3), the phase at index 3 is unwrapped to -5.77 radians. Linearly interpolating to obtain the phase at index 2.5

results in -4.26 radians. Wrapped back to the range $[-\pi, \pi)$, the interpolated phase at index 2.5 is 2.03 radians. As shown in Figure 10a by the asterisks, the interpolated phase at index 2.5 is approximately equal to the phase calculated from the 8192-point STFT (1.71 radians).

Now let us interpolate the phase at index 6.5. The phase at index 6 is -1.59 radians, and the phase at 7 is 0.98 radians. This is only a difference of 2.57 radians. Using (5.3), the unwrapped phase equals the wrapped phase. Linearly interpolating the phase for index 6.5 results in an interpolated phase of -0.31 radians, as indicated by the asterisk at index 6.5 on Figure 10a. However, the phase calculated from the 8192-point STFT is 2.35 radians. Obvious from the 8192-point STFT, the phase should have unwrapped between these indices, but could not using (5.3).

Upon further investigation using the continuous phases shown in Figure 10b and Equations (5.4) and (5.5) to calculate the maximum frequency resolution, f_R, the reason for the radically different results obtained from interpolating the phases at indices 2.5 and 6.5 becomes clear. For interpolating at index 2.5, the phase slope, m_ϕ, between indices 2 and 3 is -0.19 radians/Hz. This results in a maximum frequency resolution of 16.22 Hz. Therefore the 512-point STFT with a frequency resolution of 15.625 Hz is below the maximum frequency resolution, fulfilling the frequency resolution requirement to prevent phase aliasing.

For interpolating at index 6.5, m_ϕ between indices 6 and 7 is -0.24 radians/Hz. This results in a maximum frequency resolution of 13.23 Hz. Thus the frequency resolution of 15.625 Hz exceeds the maximum frequency resolution required to prevent phase aliasing. Attempts to interpolate between these indices may result in incorrect phase values, as we saw in our example. As shown in the continuous phase plot in Figure 10b, phase aliasing occurs beginning at index 4 and continuing through index 11. Few of the phases in this area are properly unwrapped. An especially severe problem area occurs between indices 10 and 11 where the phase slope increases dramatically.

Unfortunately, phase aliasing as seen in Figure 10 is not a rare occurrence when using a 512-point STFT. Figure 11 shows the calculated phases of a 512-point STFT of a voiced frame of speech versus the calculated phases of an 8192-point STFT taken from the same speech signal.

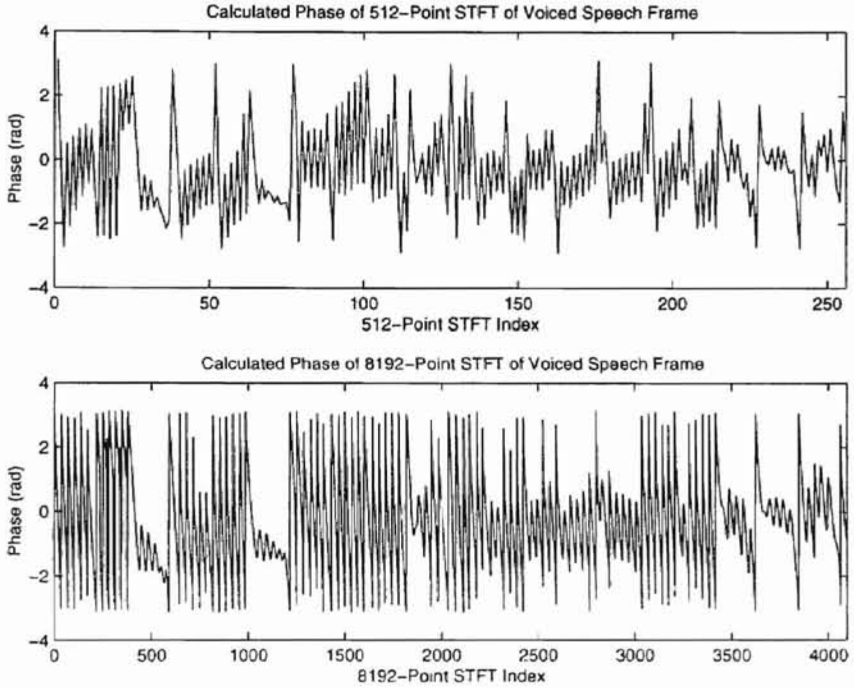


Figure 11. Phase comparison of a 512- vs. 8192-point STFT (voiced frame)

The phases between indices 0 and 1000 on the 8192-point STFT index scale are clearly problem areas for the 512-point STFT. Throughout this area, m_ϕ averages between -0.25 radians/Hz and -0.21 radians/Hz. This restricts the maximum frequency resolution to 12.5 Hz. The 512-point STFT falls far short with a frequency resolution of 15.625Hz.

Multiple voiced speech frames were analyzed for phase aliasing problems, and phase aliasing was found to occur frequently. In most instances, an 8192-point STFT was required to prevent phase aliasing, although phase aliasing still occurred occasionally even with that length STFT. Again, due to the computational restrictions necessary for real time implementation, such lengthy STFT's are not realistic. Therefore, this method, like the first, is not sufficient for accurate analysis of phase.

Method 3: Complex Interpolation of the STFT

The third solution to the phase interpolation problem is to perform a complex cubic interpolation of the STFT and calculate the phases from the interpolated STFT. Cubic splines are computationally expensive, but this factor can be reduced by calculating only the values needed rather than fitting a cubic spline to the entire 512-point STFT. Figure 12 shows the real part of a 512-point STFT of the same voiced frame of speech used in Figure 11 as compared to the real part of an 8192-point STFT.

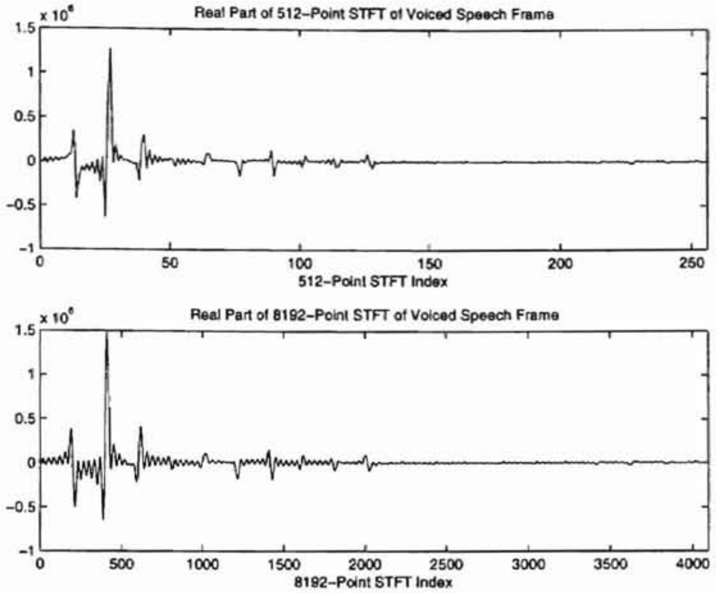


Figure 12. Comparison of real part of 512-vs. 8192-point STFT

The imaginary part of each is similar in structure to the real part. In the complex domain, the sampling is dense enough that all of the peaks in the 8192-point STFT are present in the 512-point STFT. Therefore, the real and imaginary parts of the STFT can be easily interpolated to simulate an 8192-point STFT using a cubic spline, as shown in Figure 13. However, note that the spline fit is not always a good fit. In fact, between the range of 200 and 400 on the 8192-point STFT index, the error is as high as 100%.

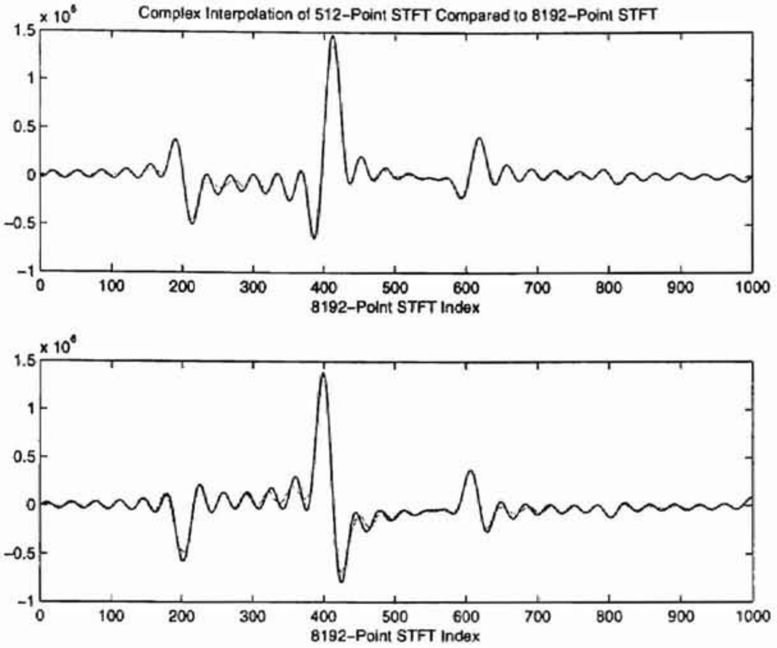


Figure 13. Complex interpolation of 512-point STFT

a) Real part of 8192-point STFT (interpolated 512-point STFT in gray).
b) Imaginary part of 8192-point STFT (interpolated 512-point STFT in gray).

After the real and imaginary parts of the 512-point STFT are interpolated, the phase is calculated from (5.2) using the interpolated complex STFT values. Figure 14 shows the results of complex-interpolation from the voiced frame used in Figure 13 before.

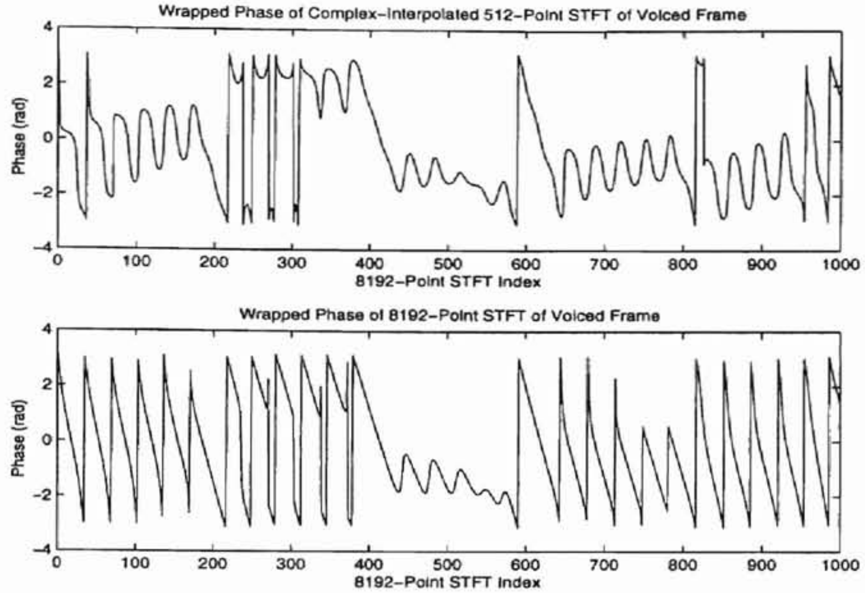


Figure 14. Phase results of complex-interpolated 512-point STFT

Unfortunately, as Figure 14 shows, the differences in the cubic interpolation of the real and imaginary parts of the STFT significantly alter the phase calculations. Complex interpolation of the 512-point STFT still produces incorrect phase estimates at an increased computational cost.

Analysis Conclusion

Because using an STFT with a length greater than 512-points is not realistic for real-time applications, none of these three methods for phase analysis produces results accurate enough for good phase estimation. All require further heuristic smoothing to reduce roughness in the sinusoidal reconstruction. None of these methods permits production of high quality reconstruction of speech due to the roughness in the sinusoidal reconstruction. Even so, if the roughness can be overlooked, the natural tonal quality is retained, indicating that phase is important for tonal naturalness. Unfortunately, to reduce the audible roughness produced by inaccurate phases to an acceptable level, an extremely long STFT (greater than or equal to 8192 points) must be used. Again, this is not realistic due to the significant increase in computation. Thus another method for obtaining phase must be sought.

Synthetically generating phases seems to be the only other solution, since phases cannot be analyzed accurately at reasonable STFT lengths. This brings us back to the original problem of an insufficient synthetic phase model—the excitation phase model with additional noisy jitter added to the phases in higher harmonics does not sufficiently model real phases. Synthetic phases produced from this model result in speech which has varying degrees of synthetic tonal quality. Another model must be used if tonal quality is to be improved. Indeed, another model exists which takes into account not only the excitation phase but the phase due to the vocal tract response. It is the cepstral phase model.

Cepstrum Theory

It has been shown that the phases of voiced speech can be divided into the sum of two separate phases—linear excitation phase, which is the result of vocal fold excitation, and system phases, which are the result of the vocal tract response [18]. The linear excitation phases are calculated using (3.4) and (3.5). For convenience, they are restated here as Equations (5.6) and (5.7).

$$\phi_0(k) = \phi_0(k-1) + \frac{\left(\omega_0^{k-1} + \omega_0^{k}\right)}{2}\left(\frac{S}{f_s}\right) \tag{5.6}$$

$$\phi_l^k = l\phi_0(k) \tag{5.7}$$

The excitation phases are used in modeling the synthetic phases in IMBE and EMBE. This much of the phase model is not new to MBE-based coders. However, the introduction of system phases from the cepstrum is.

The system phases are calculated from the set of cepstral coefficients generated from the harmonic amplitudes [18]. Equation (5.8) denotes the system phase calculations where M is the number of cepstral coefficients and ω is the frequency in radians.

$$\Phi(\omega) = -2\sum_{m=1}^{M} c_m \sin(m\omega) \tag{5.8}$$

It has been shown that M greater than approximately 44 is sufficient to produce good phase estimation [18]. In the simulations that follow, M is chosen to be 50. The final phase is then the sum of the excitation and system phases. This is denoted by Equation (5.9).

$$\theta_l\ (\omega) = \phi_l + \Phi(\omega) \tag{5.9}$$

Two important notes need to be made before the discussion of cepstral phase continues. First, notice from (5.6) that the excitation phase depends on the current and past fundamental frequencies as well as the past excitation phase. Due to this dependence on the past, an incorrect pitch in one frame results in a cascade of errors in current and future excitation phase calculations. Secondly, notice from (5.8) that the system phases, in contrast to the excitation phase, depend only on the current cepstral coefficients which are a function of the current vocal tract response only. There is no dependence on past calculations. This makes the system phases computed from the cepstrum much more dynamic and robust to errors than the excitation phase.

The cepstral coefficients (c_m) are generated from the natural logarithm of the magnitude spectrum. The log is used so that the convolution of the excitation and vocal tract response (5.10) can be separated into an additive function (5.13). This general idea is illustrated in Equations (5.10-5.13), where $\hat{s}(n)$, $\hat{e}(n)$, and $\hat{\theta}(n)$ refer to the inverse STFT of (5.12).

$$s(n) = e(n) * \theta(n) \tag{5.10}$$

$$S(\omega) = E(\omega)\Theta(\omega) \tag{5.11}$$

$$\log|S(\omega)| = \log|E(\omega)| + \log|\Theta(\omega)| \tag{5.12}$$

$$\hat{s}(n) = \hat{e}(n) + \hat{\theta}(n) \tag{5.13}$$

After taking the inverse STFT, denoted by Equation (5.14), the cepstral coefficients produced are linearly separable into an excitation response and a vocal tract response. This is shown by Equations (5.15a) and (5.15b) [33]. Note that P is the pitch period in time-domain samples.

$$s(n) = \frac{1}{2\pi}\int_0^{2\pi} S(\omega)\exp(j\omega n)\,d\omega \tag{5.14}$$

$$c(n) = c_e(n) + c_\theta(n) \tag{5.15a}$$

$$c(n) = \begin{cases} c_e(n) + c_\theta(n) & n = 0 \\ c_\theta(n) & 0 < n < P \\ c_e(n) & n \geq P \end{cases} \tag{5.15b}$$

From (5.15b), it is clear that the coefficients of the excitation response and vocal tract response are easily separated. The cepstral coefficients representing only the vocal tract response are then filtered out and used in (5.8) to calculate the system phase. Figure 15 shows this process. For further detail on cepstral theory and representation see [33].

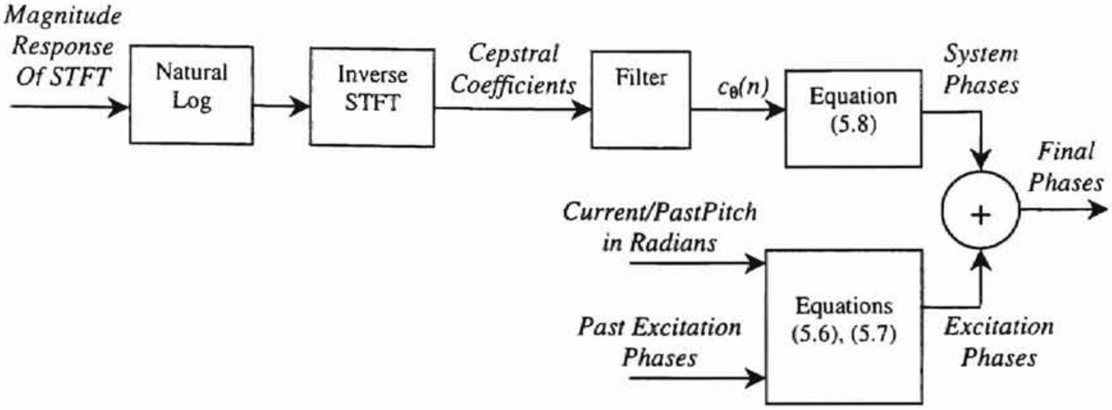


Figure 15. Block diagram of cepstral phase calculations

The cepstral calculation process of Figure 15 can be altered so that only the harmonic amplitudes are needed rather than the whole magnitude response of the SFTF. Because the vocal tract response needs to be isolated anyway, the natural log can be applied to the harmonic amplitudes only. A cubic spline is then fitted to the log of the harmonic amplitudes to recreate the vocal tract response. At this point, the recreated spectrum has no excitation left. The inverse STFT of the spline fit is taken, producing the cepstral coefficients of only the vocal tract response. No filtering is needed to separate the coefficients of the vocal tract response. Figure 16 shows a basic block diagram of the process.

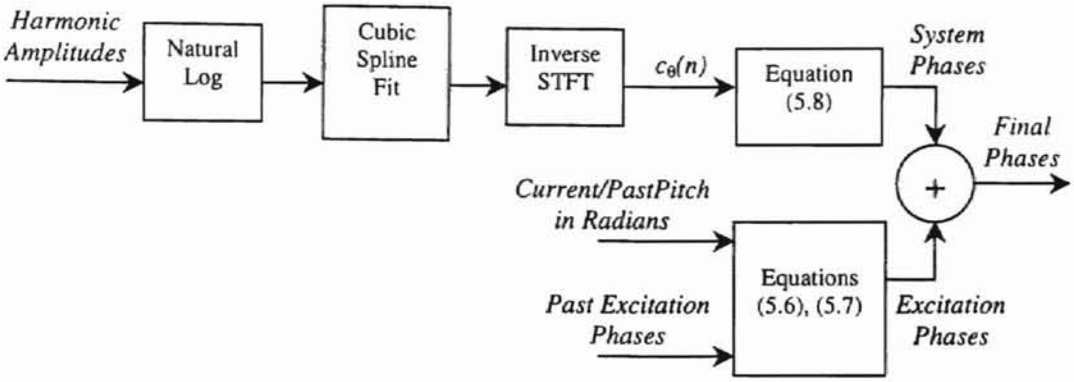


Figure 16. Modified block diagram of cepstral phase calculations

By eliminating the need for the entire spectrum and only requiring the harmonic amplitudes, the cepstral phase model can now be generated entirely in the synthesizer. This eliminates the need to send any additional information in the analyzer than what is already sent by MBE-based coders such as IMBE and EMBE.

Cepstral Phase Model Results

Phases generated using the cepstral model produce very natural-sounding voiced speech which is superior to voiced speech reproduced using any the three previously mentioned methods of estimating phase. The two methods of generating the cepstral coefficients—using the entire spectrum or only the harmonics—produce audibly indistinguishable results. Both methods completely remove the audible roughness produced from using analyzed phases, eliminating the need for heuristic smoothing of the phase. In addition, since the second method for producing the cepstral coefficients requires using only the pitch and harmonic amplitudes, the second method can be used to generate the harmonic phases entirely in the synthesizer.

However, in contrast to the high quality reconstruction of voiced speech, using the cepstral phase model for sinusoidal reconstruction of unvoiced speech components results in very buzzy-sounding speech. By replacing the cepstral phase model with random phases in unvoiced components, as was discussed earlier, unvoiced areas can again be reproduced well using sinusoidal reconstruction [18].

Cepstral Limitations

Before leaving the phase issue, the limitations of the cepstral model need to be addressed. The cepstral phase model produces very high quality reconstructed voiced speech under normal conditions. However, it is not as robust as using measured phases even though measured phases introduce their own difficulties.

The cepstral phase model is adversely affected by pitch errors and incorrect representation of the vocal tract response. Unvoiced areas incorrectly classified as voiced which are located immediately prior to voiced areas, are highly susceptible to the production of incorrect cepstral phases which can cause synthetic tonal qualities over the entire voiced area. This generally occurs when the pitch changes dramatically between frames. Measured phases, since they stand alone and are not derived from other parameters, are not affected by such errors. Fortunately, the problem of extended synthetic tonal qualities can be eliminated by using a specific method of frame smoothing (overlap addition) discussed in Chapter 6.

Such problems which are the result of incorrect parameter estimation are a drawback of any synthetically generated data. If the data used in a calculation is incorrect, the calculated data will not be correct either. Therefore, correct parameter estimation in the analyzer is critical to high quality results when using the cepstral phase model. Unfortunately, harmonic coders such as MBE are particularly susceptible to pitch errors. However, correct parameter estimation is critical to high quality results for any type of speech reconstruction model. Even the quality of the best coders will degrade as the number of parameter estimation errors increases.

Despite the cepstral phase dependence on correct parameter estimation, the phases generated using the cepstral phase model are better than those obtained by any of the other methods discussed. Reconstructed speech based on the cepstral phase model is smoother than any of the speech reconstructed using the methods discussed for measuring phase. Of the two synthetic phase models discussed (the linear excitation model with additional noise jitter and the cepstral phase model), the cepstral phase model produces reconstructed speech with more natural tonal quality. Using cepstral phases also has the additional benefit of eliminating the need to send any additional information in the analyzer since the phases can be generated completely in the synthesizer.

This completes the foundation necessary for the development of an enhanced reconstruction method. Now let us briefly review the decisions made up to this point in our discussion.

1) A harmonic model serves as the basic synthesis structure.

   a) Frames with any voiced components are reproduced using a sinusoidal harmonic model with randomized phases for unvoiced sinusoids. Resampling in unvoiced areas is performed so that the spacing between sinusoids in unvoiced areas is less than 100 Hz.

   b) Completely unvoiced frames are reproduced using band-limited white noise as is done in IMBE and EMBE.

2) Phases are generated synthetically in the synthesizer using the cepstral phase model based on the sum of a linear excitation phase and a system phase.

Now the basic structure is established and all the parameters are available for use in the synthesizer. The development of the enhanced sinusoidal reconstruction method will proceed with the issue of how to connect parameters smoothly from frame to frame.

CHAPTER VI

CONNECTION OF PARAMETERS FROM FRAME TO FRAME

Even if all the parameters are estimated correctly and the basic structure fits the speech perfectly, the reconstructed speech will not be of high quality unless all the parameters can be connected smoothly from frame to frame. Pivotal to all of the smoothing and interpolation is the selection of a frame connection method. The method for connecting frames affects how much smoothing and interpolation must be done to the parameters. Certain methods such as overlap addition inherently smooth the parameters as the frames are connected. Others, such as non-overlapping methods, require explicit smoothing of parameters. Therefore, the type of frame connection must be established before the parameter smoothing discussion can continue with any sense of relevance.

Two basic methods of frames connection are considered. The first is overlap addition. It is by far the simpler of the two and has many benefits related to this simplicity. It is also a method which inherently smoothes the other parameters as the frames are connected, so no further parameter smoothing is required for this method. The other method is a non-overlapping method developed by McAulay and Quatieri for STC. It is much more complex and, for convenience, this method will be called the "STC non-overlapping method." The STC non-overlapping method is quite computational and requires explicit smoothing of every parameter. However, it offers the advantage of precise (explicit) "control" of all aspects of the reconstruction. A necessary discussion of parameter smoothing follows the discussion of the STC non-overlapping method.

Overlap Addition

The process of overlap addition is very straight-forward. Basically, for each harmonic in the current and future frames, the amplitudes, phases, and frequencies are used to construct time-domain sine waves two frames long. All the sinusoids are summed using Equations (6.1) and (6.2), where n is the sample number, S is the number of samples per frame shift, f_s is the sampling frequency, l is the harmonic number, and k is the frame number.

$$\theta_l(n) = \theta_l + \omega \frac{n}{f_s} \qquad n = [-S:(S-1)] \tag{6.1}$$

$$s_{tmp}(n) = \sum_{l=1}^{L} A_l \cos[\theta_l(n)] \qquad n = [-S:(S-1)] \tag{6.2}$$

Note that since the frequency is held constant, phase (the integral of the frequency) is linear (6.1). The sum is weighted by a tapered window. The only restriction on the tapered window is that the overlapped sum of the tapered windows must equal one at every point. The simplest overlapping taper is the fully overlapping triangular window, shown in Figure 17. Note that N is the number of samples generated for each synthesis frame.

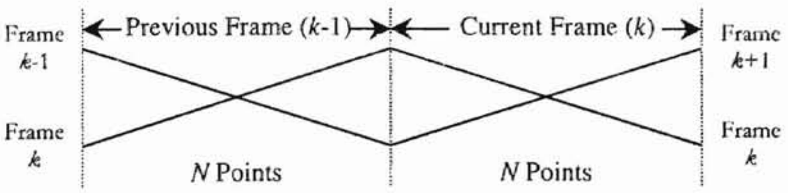


Figure 17. Overlap addition tapering (triangular, fully overlapping)

The second half of the triangular-tapered reconstruction from the current frame is overlapped and added to the first half of the triangular-tapered reconstruction from the future frame to produce a very smooth transition between frames (6.3) [17].

$$s^{k}(n)=s_{tmp}^{k}(n)\frac{S-n}{S}+s_{tmp}^{k+1}(n-S)\frac{n}{S} \qquad n=[0:(S-1)] \tag{6.3}$$

This same method works for unvoiced reconstruction and is in fact always used to connect the unvoiced frames. Two frames of band-limited white noise are constructed in the time domain using the parameters of the current frame. A triangular window is applied, and the unvoiced data is overlapped and added as indicated above. In both the voiced and unvoiced case, the resulting transition is completely smooth.

Overlap addition has numerous advantages, the first of which is simplicity. Only a simple sum of two linear tapers, involving only real operations, is required. The only additional computation is the cost of reconstructing the sinusoids for two frames instead of one. All frames, regardless of whether or not they are voiced, unvoiced, or contain mixed excitation, can be connected this way.

Secondly, overlap addition requires no further parameter smoothing. There is no consideration of which harmonics to connect from frame to frame or how to smooth amplitude, frequency, and phase. In effect, the tapered window takes care of all smoothing.

In addition to the simplicity and the elimination of further parameter smoothing, overlap addition eliminates the extremely synthetic tonal qualities produced by cepstral phases when severe pitch errors occur. This makes the overlap addition method very robust in cases of severe errors in parameter estimation. This is one of its most recommending qualities.

The only limitation of overlap addition is that reconstruction frame lengths must be restricted to small sizes. As long as frames of 12.5ms duration or less [18] are used, reconstructed speech using overlap addition is extremely smooth and of very high quality. However, rapid decay of quality is experienced with the increase of the frame duration above 12.5ms—a limitation at lower bit rates. By 23ms frame duration, overlap addition produces poor quality speech [17]. It was for this reason that the STC non-overlapping method of frame connection was developed.

The STC Non-Overlapping Method

As a contrast to the overlap addition frame connection method, the STC non-overlapping method proposed by McAulay and Quatieri [17] is highly computational and requires further smoothing of all the parameters. However, the STC non-overlapping method is better able to produce quality speech when long frames are used. Again, this is typically beneficial for low rate coders.

The basic concept of the STC non-overlapping frame connection method is that each parameter is interpolated within each frame so that at frame boundaries the parameters match exactly with no discontinuity. No overlapping is performed whatsoever between adjacent frames. This method is very dependent both on the smooth connection of parameters and on deciding which parameters to connect. If the parameters are correctly smoothed, the STC method is known to produce very high quality speech when using a non-harmonic model like the one used in the original STC design by McAulay and Quatieri.

However, the results of using this method with a harmonic model, voicing decisions, and a separate unvoiced reconstruction method are not known. The original STC design did not account for any of these conditions. In addition, the non-overlapping connection method was specifically designed for the reconstruction of long frames—not short ones. Since the enhanced reconstruction method being developed is designed for MBE-based mid-rate speech coding with a harmonic model, voicing decisions, separate unvoiced reconstruction, and typically shorter frames, it is the quality of reconstruction resulting from the implementation of the STC non-overlapping frame connection method under these new conditions that is of interest.

At this point, a comparison between the two methods of frame smoothing would be appropriate. However, to implement the STC non-overlapping frame connection method with high quality results, the issues of parameter connection and smoothing must first be addressed. Because overlap addition does not require any further parameter connection discussion, but the STC non-overlapping method does, the discussion of parameter smoothing which follows relates only to the STC non-overlapping method. The comparison between the two methods will continue at the end of this chapter. With that in mind, we will pursue the issues related to parameter smoothing.

Parameter smoothing raises many questions. Which harmonics in the current frame connect to which harmonics in the previous frame? Should the change in frequency from frame to frame be an issue? What types of interpolation—linear, quadratic, or cubic—are best to smooth the different harmonic parameters?

The questions are numerous and involved. To resolve these, we must begin by deciding how to best connect harmonics from frame to frame. The parameters of harmonics cannot be correctly smoothed from frame to frame until they are correctly matched, so the matching process becomes the first step.

Harmonic Matching

Harmonic matching is really a matter of deciding which set of parameters to connect from frame to frame. There are several factors that must be considered when harmonics are matched. Must the harmonic relationship from frame to frame always be preserved? In other words, will the first harmonic always be connected to the first harmonic, and the second to the second, etc.? Will the change in pitch from frame to frame be allowed to affect the connection of parameter sets? For instance, a pitch change of 15 Hz results in a frequency change of 300 Hz at the twentieth harmonic. Should such frequency changes be considered as a factor for parameter set matching? Along those same lines, is the true speech relationship not based on the matching of harmonics at all, but on the matching of closest frequencies? Indeed, should harmonic numbers be completely disregarded between adjacent frames and the parameter sets be matched completely by which harmonic frequencies are the closest?

All of these ideas must be considered, but they can be broken down into the combination of two basic techniques for matching parameter sets. The first technique is matching harmonics directly one-to-one, and the second is matching closest frequencies without regard to harmonic number.

Method 1: Direct Harmonic Number Matching

Matching parameter sets by harmonic number is the simpler of the two basic techniques. Each harmonic is matched to the corresponding harmonic number in the next frame regardless of pitch change. Thus the first harmonic always matches the first; the second always matches the second, etc. Figure 18 shows an example of matching harmonics directly by harmonic number.

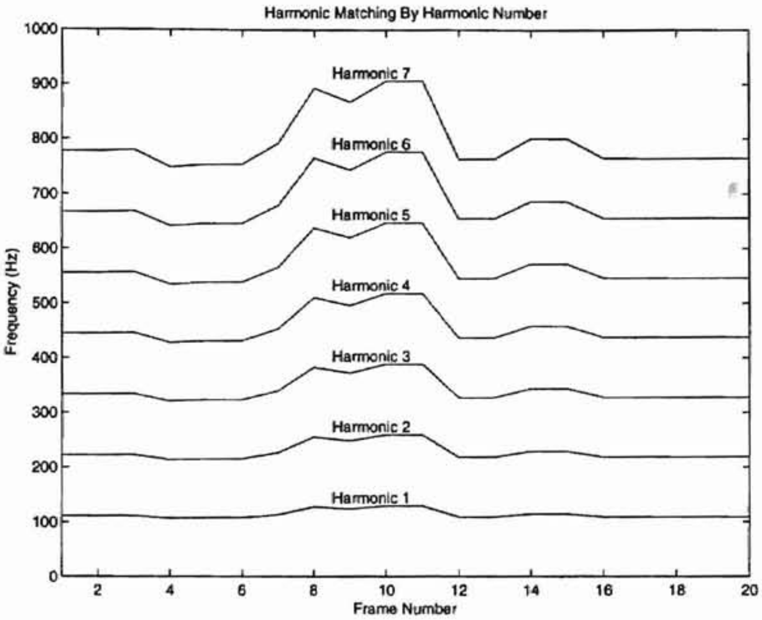


Figure 18. Harmonic matching by harmonic number

As Figure 18 illustrates, small changes in the pitch result in increasingly larger changes in frequency as the harmonic number increases. For example, the frequency change of the first harmonic from frames 7 to 8 is only 14 Hz. However, by harmonic 7, the frequency change is nearly 100 Hz. Again, the frequency change of the first harmonic from frames 11 to 12 is only 18.5 Hz. By harmonic 7, the frequency change is 130 Hz. Fairly insignificant frequency changes in the fundamental frequency result in large frequency sweeps in the upper harmonics. Undesirable frequency sweeps can become audible, especially if pitch errors such as doubling and halving occur. This places more weight on pitch decisions, an area which is already inherently non-robust.

Therefore, it seems logical that direct harmonic number matching without further modification is probably not the best method to use.

In MBE, harmonics are normally matched according to their harmonic numbers, but the fundamental frequency change is considered during harmonic matching. This seems like a logical modification to harmonic matching by harmonic number. If the pitch changes by more than a set threshold, harmonic matching by number is abandoned and the harmonics are treated as if preceded and followed by unvoiced harmonics [13]. This is analogous to the birth and death of a track, which is part of the second method of harmonic matching—matching by closest frequency.

Method 2: Matching by Closest Frequency

The second basic technique is the opposite extreme of direct harmonic matching without regard to frequency change. With matching harmonics by closest frequency, the harmonics with the smallest frequency changes are matched without regard to harmonic number. Unfortunately, this is much more complex than direct harmonic number matching. The actual description of the process is fairly simple although its implementation is much more complex due to necessary iterative comparisons required for confirming the frequency matches.

Because harmonic numbers are no longer used, an alternate method of keeping track of the harmonic connections from frame to frame must be used. McAulay and Quatieri assign parameter sets to "tracks." These tracks are used to follow the connections of the harmonics from frame to frame. To perform the actual matching, McAulay and Quatieri developed a three step algorithm for track matching. [17] Figure 19 illustrates the different cases for each step in the algorithm. Note that ω represents the harmonic frequency in radians, *k* represents the current frame, *k*+1 represents the subsequent frame, *n* represents the harmonic number in the current frame, and *m* represents the harmonic number in the future frame.

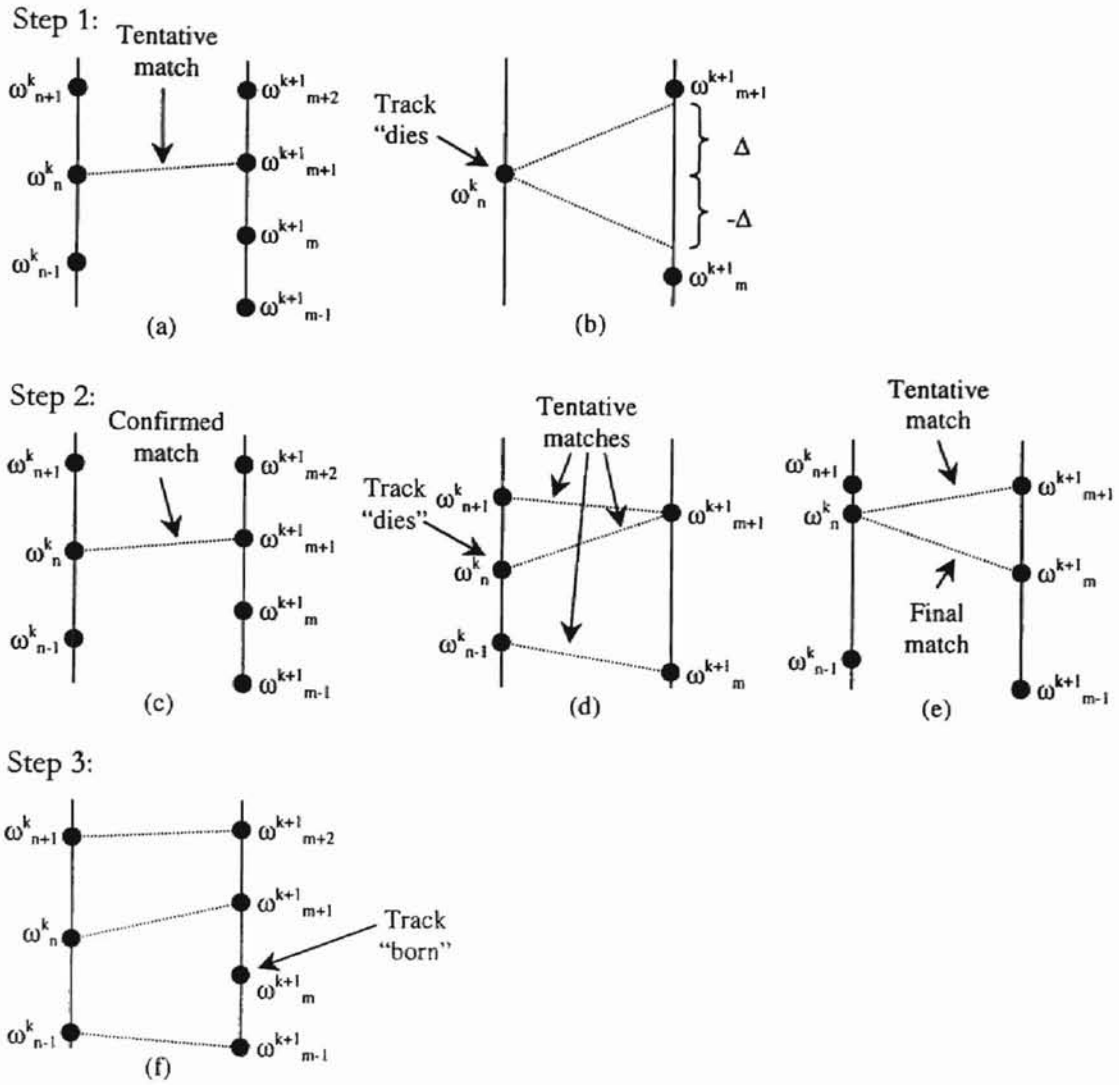


Figure 19. Detailed steps for track matching

In the first step of the algorithm, tentative matches are made between all harmonics in the current and future frames. Each harmonic in the current frame is matched to the harmonic with the closest frequency in the future frame (Figure 19a), unless the frequency difference is greater than a "matching interval" Δ. In that case, the track "dies" (Figure 19b).

Second, the tentative matches are either confirmed (Figure 19c) or denied. For example, if two harmonics in the current frame are matched to the same harmonic in the future frame, the track of the harmonic with the largest difference in frequency in the current frame must "die" (Figure 19d). If the first tentative match is not the best match, then a reassignment to another match may occur (Figure 19e). In the case of using a harmonic model (MBE) rather than peak matching (STC), this particular case can be eliminated. If two harmonics are matched tentatively to the same harmonic in another frame, the closest frequency will be matched and the other must "die" or be "born." A harmonic in the current frame will never be matched to a harmonic in the subsequent frame which is not the initial tentative match.

The last step is to "birth" all unmatched future harmonics into new tracks. This is illustrated in Figure 19f. Figure 20 illustrates the results of all three steps of matching harmonics by frequency. It uses the same pitches from Figure 18 as a comparison. Notice how different the contour for matching the parameter sets by frequency is compared to the contour for matching directly by harmonic number.

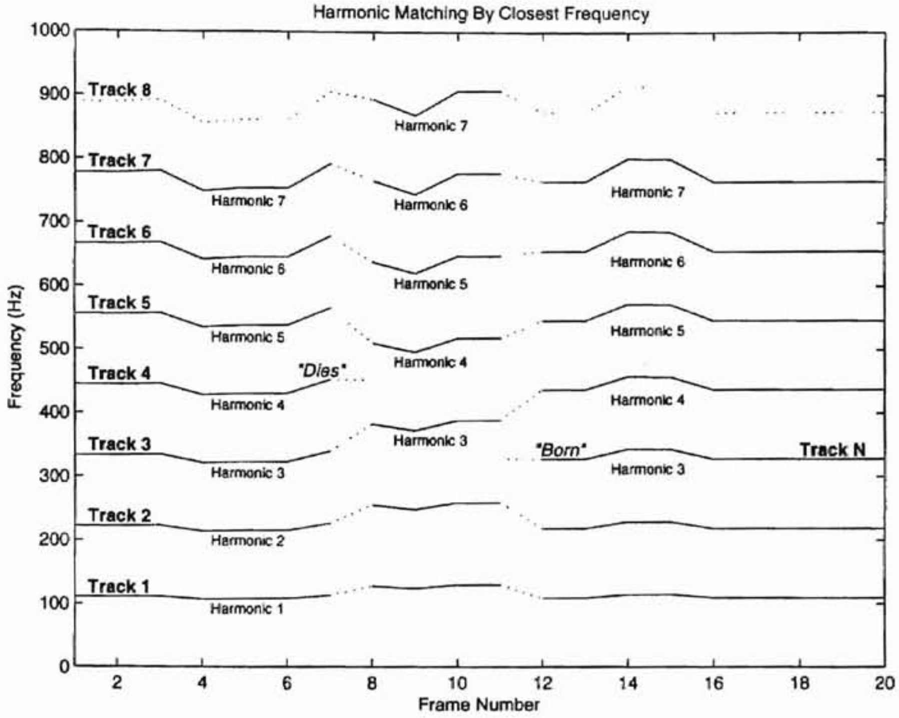


Figure 20. Harmonic matching by closest frequency

Between frames 7 and 8, a pitch change of 14 Hz occurs, resulting in harmonics above harmonic number 3 no longer being matched to the same harmonic number. Harmonic 4 is now connected to harmonic 5 from the previous frame; harmonic 5 is connected to harmonic 6 from the previous frame, etc. Notice that although harmonic 4 is matched to harmonic 5 in frame 8, track 4 has no match in frame 8 and therefore dies in frame 8.

Similarly, between frames 11 and 12 where a pitch change of 18.5 Hz occurs, the harmonics switch tracks again. Note that harmonic 3, although it was assigned to a track in frame 11, was not matched at frame 12 and therefore had to be born in frame 11 into a new track, track N.

Comparison of Harmonic Matching Methods

Both methods of harmonic matching were implemented into an STC-based synthesizer of an EMBE-based coder for comparison. The analysis method was held constant in each case, and the synthesizer was implemented using the STC non-overlapping frame method with full parameter smoothing, which will be discussed later.

After informal listening tests were conducted with three or four volunteers, all agreed that the direct harmonic matching method was not optimal. Though by far the simpler method, matching by harmonic number had the adverse side effect of audibly ramping upper harmonics (producing frequency sweeps) when the pitch change from frame to frame was large. Otherwise, the tonal quality was excellent and very natural. The frequency matching method removed all the frequency sweeps and also produced speech with excellent tonal quality. Unfortunately, matching by closest frequency added significant logical complexity and computation time to correctly match the tracks. The complexity of implementing tracks greatly increased the difficulty in producing error-free track matching.

Considering these results, it seemed that the best solution for harmonic matching might be to modify the matching by harmonic number method so that pitch changes are accommodated as they are in MBE. Using the MBE modification as precedence, testing was performed to match harmonics directly unless the fundamental frequency change was greater than ten percent. In these cases, harmonics were not matched at all, but were "birthed" and immediately "killed," similar to what was done to the tracks in frequency matching when the track had no previous or subsequent track match.

After further informal listening tests, the slightly altered harmonic matching by harmonic number with pitch changes considered appeared to retain the high tonal quality of both harmonic matching by harmonic number and frequency matching. In addition, the audible high frequency sweeps which occurred as a result of harmonic matching by harmonic number were removed while retaining the simplicity of this method. In fact, listeners could not distinguish any audible change in the speech quality between frequency matching and this method. For this reason, the somewhat altered harmonic matching by harmonic number method was chosen as the method to match harmonics from frame to frame.

Now that a good method for matching harmonics from frame to frame has been established, interpolation must be performed to smoothly connect the parameters of each harmonic from frame to frame. Again, the three parameters under consideration are amplitude, frequency, and phase. We will begin with the discussion of amplitude interpolation.

Amplitude Interpolation in Time

Amplitude interpolation is intuitively the most straight forward interpolation of all the parameters. According to McAulay and Quatieri's original STC paper, the obvious solution to the amplitude interpolation problem is to use linear amplitude interpolation between frames [17]. Equation (6.4) is given where n is the sample number, k is the frame number, and S is the number of samples per frame shift. This linear amplitude interpolation technique is used in IMBE and EMBE also.

$$A_l^k(n) = A_l^k + \frac{A_l^{k+1} - A_l^k}{S} n \tag{6.4}$$

However, if the changes in amplitude are very large, linear amplitude interpolation results in audible amplitude discontinuities at frame boundaries. This fact is not addressed in the literature. Figure 21a shows a single harmonic of a real voiced speech segment reconstructed using STC reconstruction with linearly interpolated amplitudes. A discontinuity can be seen at sample number 170 when the amplitude changes suddenly. Although this discontinuity is subtle in the plot, it is easily audible when a single sinusoid is reconstructed.

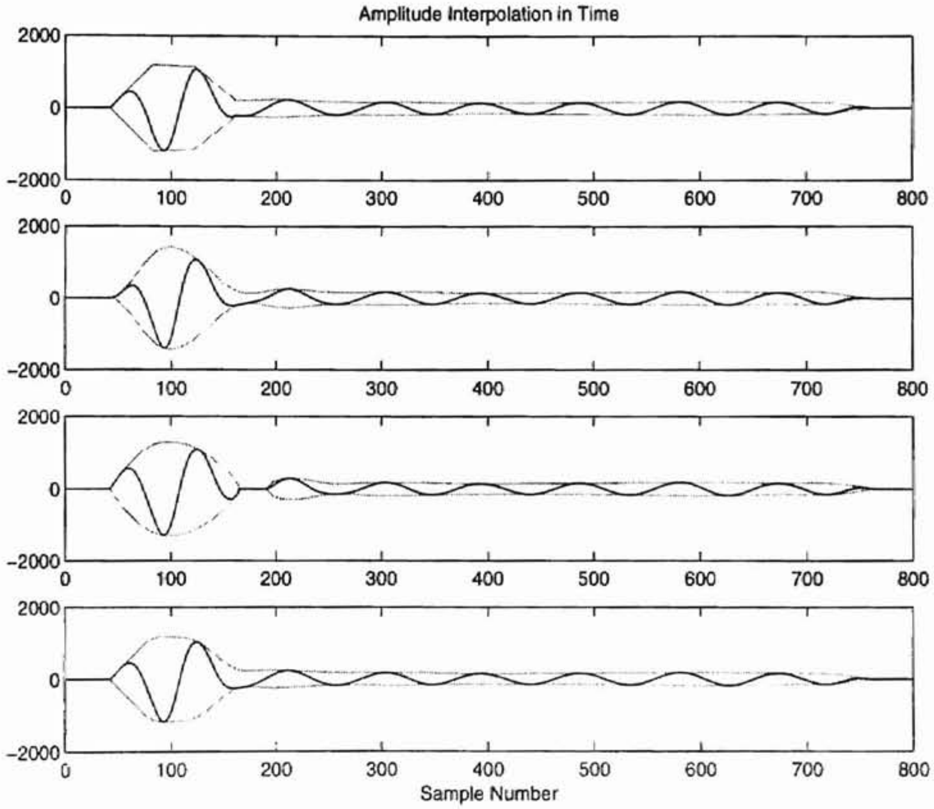


Figure 21. Amplitude interpolation in time

a) Linear amplitude interpolation
b) Cubic amplitude interpolation
c) Squared/square-root cubic amplitude interpolation
d) Linear amplitude interpolation with moving average filter

Because this problem is not addressed in STC or MBE literature, several different possible methods of removing these audible discontinuities are explored. These methods include cubic spline interpolation, squared/square-rooted cubic spline interpolation, and linear interpolation with a moving average filter applied.

The first method used to eliminate the linear interpolation discontinuities is to replace the linear interpolation with cubic amplitude interpolation. Figure 21b illustrates that these discontinuities are completely eliminated when a cubic spline is applied to the same amplitudes.

Unfortunately, cubic amplitude interpolation, although it removes the discontinuities, adds computational complexity and sometimes results in undesirable amplitude undershoots and overshoots at the beginning and ending of voiced areas. Undershoots (negative amplitudes) reverse the phase by 180 degrees, creating discontinuities at the positive and negative transition. Overshoots produce audibly "bursty" and overemphasized voiced onsets and endings.

Neither case is desirable, and both are more likely to occur with increased severity if harmonics are matched by closest frequency. With harmonic one-to-one matching, as an area transitions from unvoiced to voiced and voiced to unvoiced, the harmonic amplitudes will tend to ramp up and down, respectively, over several frames. However, when using closest-frequency matching, a harmonic in the middle of a voiced area may have a very high amplitude but may not be matched to a harmonic in the next frame due to frequency matching. These harmonics must "die" by the next frame. This sudden rapid change from high amplitude to zero amplitude creates severe undershoots in the cubic interpolation, particularly with short frames. Likewise, harmonics which are suddenly "born" in the middle of a voiced area because of frequency matching tend to severely overshoot.

Eliminating these two problems inherent to cubic interpolation is difficult because they tend to occur at the places linear interpolation results in discontinuities. To remove undershoots, the amplitudes returned from the cubic spline can be checked for negative values. If any negative amplitudes are returned, the cubic interpolation may be replaced with linear interpolation. However, this may reintroduce discontinuities into the reconstructed signal.

The negative amplitudes cannot simply be replaced with amplitudes of zero because this reduces the amplitude onset and decay time. The slowly tapered, continuous changes required for smooth speech reconstruction are replaced by extremely abrupt onsets and decays. In fact, this sounds little better than no interpolation. The waveforms become audibly rough.

Overshoots are even more difficult to remove. They can be reduced by fitting the spline to the square of the amplitudes rather than the amplitudes themselves and then taking the square root of the spline fit. This is the second method applied. The overshoots are reduced significantly by using this method, as illustrated in Figure 21c. At sample number 100, the amplitude overshoot of the cubic spline was approximately ten percent. By applying the cubic spline to the square of the amplitudes and then taking the square root of the spline fit, the overshoot is reduced to approximately one percent.

Although applying the cubic spline to the square of the amplitudes and then taking the square root of the spline fit removes the overshoots, undershoots are created which are much worse than the overshoot problems of the normal cubic spline.

Assigning the negative amplitudes (undershoots) amplitudes of zero can result in the complete loss of the harmonic amplitude in the middle of a voiced area. This is clearly seen in Figure 21c from sample numbers 170 to 195. The number of undershoots and their severity are greatly increased, and rise and fall times are more abrupt. The application of a cubic spline to the square of the amplitudes and then taking the square root of the spline fit is thus a very poor solution resulting in worse smoothing than the original linear amplitude interpolation.

The final method used to resolve the amplitude interpolation problem involves filtering the linear interpolation. A moving average filter is applied to the linearly interpolated amplitudes. The moving average filter replaces the value at the center of the filter with the average of all the values in the filter. Equation (6.5) denotes a simple moving average filter of length *N*. *N* is normally chosen to be an odd number.

$$A_l(n)=\frac{1}{N}\sum_{m=n-(N-1)/2}^{n+(N-1)/2}A_l(m) \tag{6.5}$$

The filter is applied to every interpolated harmonic amplitude. It has no effect on linear areas, smoothing off only the sharp corners produced at frame boundaries. A moving average filter also cannot create any overshoots or undershoots. Although it adds computational complexity to linear interpolation, it removes the discontinuities and produces no overshoots or undershoots. Figure 21d illustrates linear amplitude interpolation with a 21-point moving average filter applied. Notice that the linear areas are well preserved, and only the sharp transitions are smoothed. In addition, the discontinuity at sample number 170 is completely removed.

Because of the undesirable overshoots and undershoots of cubic interpolation, the moving average filter was selected as the best amplitude interpolation method. It retains the simplicity of linear interpolation with added computation and removes the amplitude discontinuities that were a problem, without introducing undesirable overshoots and undershoots. Further, if implemented recursively, the linear moving average filter adds only a little overhead to the interpolation process.

Frequency/Phase Interpolation in Time

Now that the method of amplitude interpolation has been decided, the remaining parameters, phase and frequency, will be discussed. These two parameters are probably the most important in all the reconstruction done. Because frequency is the derivative of phase, frequency and phase are inseparable. Any frequency interpolation in time will determine the phase interpolation in time and vice versa.

To smoothly connect the frequency and phase parameters in the STC non-overlapping frame connection method, McAulay and Quatieri use a sophisticated cubic phase interpolation (quadratic frequency) technique rather than the more traditional linear frequency interpolation which results in quadratic phase interpolation. As mentioned in Chapter 3, this extra degree of freedom in the frequency and phase results in the loss of a unique solution, making multiple frequency/phase paths legitimate solutions.

To overcome this, McAulay and Quatieri developed a method for unwrapping the phase so that the slope (frequency) of the phase trajectory is maximally smooth. Equations (6.6), (6.7), (6.8), and (6.9) denote the calculations necessary to obtain the interpolated phase. All of these calculations must be made for each harmonic for each frame. The variable M denotes the minimum number of 2π radian wraps that the phase goes through from the current frame to the next frame so that the phase path is maximally smooth. For detail on the derivations, see [17].

$$x = \frac{1}{2\pi}\left[\left(\theta^k + \omega^k T - \theta^{k+1}\right) + \left(\omega^{k+1} - \omega^k\right)\frac{T}{2}\right] \tag{6.6}$$

$$M = round(x) \tag{6.7}$$

$$\begin{bmatrix} \alpha(M) \\ \beta(M) \end{bmatrix} = \begin{bmatrix} \frac{3}{T^2} & \frac{-1}{T} \\ \frac{-2}{T^3} & \frac{1}{T^2} \end{bmatrix} \begin{bmatrix} \theta^{k+1} - \theta^k - \omega^k T + 2\pi M \\ \omega^{k+1} - \omega^k \end{bmatrix} \tag{6.8}$$

$$\theta^k(t) = \theta^k + \omega^k t + \alpha(M)t^2 + \beta(M)t^3 \tag{6.9}$$

This leaves the problem of how to generate a phase value for the birth and death of tracks. McAulay and Quatieri calculate the birth and death phases using linear excitation phase. This is accomplished by using Equations (6.10) and (6.11) for the birth and death of a harmonic, respectively. In both cases, the frequencies for the current and subsequent frames are equal ($\omega^k = \omega^{k+1}$).

$$\theta^k = \theta^{k+1} - \omega^{k+1} t \tag{6.10}$$

$$\theta^{k+1} = \theta^k + \omega^k t \tag{6.11}$$

Taking the birth and death cases to the extreme, suppose that for every frame, we birth all the current harmonics and kill all the previous ones using Equations (6.10-1). The phases are therefore all linear, the harmonics are not matched, and the amplitudes are not smoothed. This should sound familiar. In fact, this is the overlap addition method presented earlier. We have returned full circle and now have all the necessary parameter connections and smoothing required to compare the two frame connection methods.

Before returning to the comparison of the overlap addition and the STC non-overlapping frame methods begun at the first of the chapter, a brief review of the STC non-overlapping frame method will help tie all the pieces together. The following four properties of the STC non-overlapping method developed in this chapter are significant:

1. Frames are connected smoothly by the smooth interpolation of all parameters. No overlapping of frames is performed.
2. Harmonics are connected by harmonic number except when the pitch change is greater than ten percent. In that case, the harmonics are not matched at all. They are "birthed" in the previous frame and "die" in the next frame.
3. Amplitudes are interpolated using linear interpolation with a moving average filter. This removes the discontinuities associated with linear interpolation without introducing overshoot and undershoot problems.
4. The phase is interpolated cubicly, resulting in quadratic frequency interpolation. The phase trajectory is chosen such that the phase is maximally smooth.

Overlap Addition vs. STC Non-Overlapping Method

In order to compare the quality of reconstruction resulting from the use of the two different methods of reconstruction, informal tests were again performed. Using the parameter connection and smoothing described above for the STC non-overlapping method, the STC non-overlapping method was implemented in the synthesizer of a fully functional EMBE-based coder. The overlap addition method was implemented in a similar fashion. Speech files were processed through the coders using the same analysis but different reconstruction methods. Several people were asked to listen to the reconstructed synthetic speech and give their opinion as to which version of reconstructed speech they preferred.

Overall, listeners could not distinguish any quality difference between the two methods. In fact, both methods sounded so similar that the listeners had difficulty distinguishing any differences.

Because the results of both methods were so similar, a comparison of the requirements for each of the two methods is useful. Overlap addition requires no parameter connection or smoothing. The phase is interpolated linearly, so no higher order calculations are required. The STC non-overlapping method requires smoothing of all parameters. In addition, correct parameter smoothing is critical to the smoothness of the reconstructed speech. Amplitude interpolation requires applying a moving average filter to the linearly interpolated values. Phases are interpolated cubically. All of this amounts to a significant increase in computation and logical complexity when compared to the overlap addition method. When all of this is considered along with the fact that the reconstruction results produced by the two methods are audibly indistinguishable from one another, then obviously the better choice is overlap addition.

This decision completes the development of the enhanced sinusoidal model for reconstruction. The following is a brief summary of all the final decisions:

1) A sinusoidal harmonic model is used for voiced and mixed excitation frames. Unvoiced areas are resampled at less than 100 Hz spacing.
2) Unvoiced frames are built using band-limited white noise.
3) Phases are generated synthetically in the synthesizer using a cepstral phase model.

4) The overlap addition method is used to smoothly connect frames. No additional parameter connection or smoothing is required.

This reconstruction method has been completely implemented in the EMBE 8.0kbps vocoder. The next chapter will conclude with the results of this implementation.

CHAPTER VII

RESULTS AND CONCLUSION

The previous chapters have presented several different approaches to improve the quality of reconstruction through the development of a new enhanced sinusoidal reconstruction method for mid-rate MBE-based coders. As discussed earlier, the goal of the development of this new reconstruction method was to improve the tonal quality of MBE-based coders by reintroducing phase as a parameter and then developing a reconstruction method which utilizes phase and produces synthetic speech with more natural quality. This chapter will serve to draw all the pieces together by summarizing the design of the enhanced reconstruction procedure and reiterating the options, the decisions made, and the deciding factors that led to the final enhanced sinusoidal reconstruction method. To conclude, the quality of the fully-implemented enhanced reconstruction method developed in this thesis will be discussed.

Basic Synthesis Structure

We discussed two basic sinusoidal synthesis structures which are available—harmonic, as is typically used in MBE, and non-harmonic, as is used in the original STC. The harmonic model assumes a fundamental frequency and harmonic relationships to that frequency. A sum of harmonically related sinusoids is used to reconstruct the voiced components. A non-harmonic model does not assume a fundamental frequency; instead all components are reconstructed using a sum of non-harmonically related sinusoids with frequencies obtained by locating spectral peaks.

The non-harmonic model requires sending spectral peak locations in addition to amplitudes. This significantly increases the bit rate and makes a non-harmonic model impractical for mid-rate speech coding. Therefore, a harmonic model was selected.

Because a harmonic model does not represent non-harmonically related noisy speech components such as unvoiced areas well, an alternate method for reconstructing these components is normally used. Two methods were discussed in Chapter 4. The first used sinusoids with randomized phases in unvoiced areas. The limitation is that the sinusoids must be spaced no further than 100 Hz apart in frequency to adequately represent noise. The second method used band-limited white noise, as is done in MBE coders such as IMBE and EMBE.

Informal testing found that combining both methods for representing unvoiced speech produced high quality results. Completely unvoiced speech sounded more natural when reconstructed using band-limited white noise, and unvoiced components of speech with mixed excitation sounded more natural when reconstructed using resampled sinusoids with randomized phases.

The Reintroduction of Phase as a Parameter

Since MBE literature already covers in detail the analysis of pitch, voicing decisions, and harmonic amplitudes, analysis of these parameters was not discussed. It was simply assumed that suitable methods existed to estimate accurately these parameters. However, phase as a parameter is not discussed in normal MBE literature. Because phase would be an integral part of the new reconstruction method, phase estimation was considered.

To analyze phase, the most obvious solution is to use the STFT (5.1), since it is already used to calculate spectral amplitudes. Due to constraints on computational complexity, STFT calculations were restricted to a length of 512 points. While sufficient for analysis of the magnitude spectrum, this length proves to be insufficient for phase analysis. However, three methods were discussed for phase estimation. The first, picking the phase calculated at the closest STFT index, resulted in poor quality reconstruction. The second method, interpolating the phase between STFT indices, also did not result in high quality reconstruction. To properly interpolate between indices, the STFT frequency resolution must be less than or equal to half the frequency at which the phase is rotating through 2π radians. If this requirement is not met, the spectrum is undersampled, and the decimated phases between spectral indices cannot be recovered through interpolation. This is referred to as phase aliasing. In the example of the voiced frame given in Chapter 5, the maximum frequency resolution necessary to prevent phase aliasing was found to be 12.5 Hz. The 512-point STFT, which has a frequency resolution of 15.625 Hz, was not sufficient for correct interpolation in such cases which, unfortunately, occur frequently. The third method, complex interpolation of the STFT, also did not result in high quality reconstruction. Inaccurate fit of the cubic spline to the real and imaginary parts of the STFT during complex interpolation resulted in large phase errors.

As a result of the inability to obtain accurate measured phases, the generation of synthetic phases using an alternate phase model, the cepstral model, was pursued. The cepstral phase model is based on the assumption that phase can be divided into two components—linear excitation phases (5.6-7) and system phases (5.8). The system phase is based on the cepstral coefficients. To generate the cepstral coefficients, a cubic spline is fitted to the natural log of the harmonic amplitudes so as to remove the excitation. The inverse STFT is computed, producing the cepstral coefficients. Since phase generated from the cepstral model is based on only the harmonic amplitudes (system phase) and the pitch (excitation phase), and these parameters are already sent by the analyzer, the cepstral phases can be generated completely in the synthesizer.

The cepstral phases proved to produce high quality, smooth reconstruction. Their only limitation is that they are highly dependent on pitch. Severe pitch errors can sometimes produce adverse effects such as brief bursts of synthetic tonal quality. However, this effect can be eliminated by using an appropriate frame connection method.

Connection of Parameters from Frame to Frame

Two frame connection methods known to produce high quality reconstruction were discussed. Overlap addition, the first method, is very simple. It reconstructs two frames of speech using linear excitation phase interpolation, and then applies a triangular window to the speech. The left half of the window is added to the right half of the triangular-tapered previous frame, and the right half is added to the left half of the triangular-tapered next frame, hence the name "overlap addition." No further interpolation is required. The second method, referred to here as the STC non-overlapping method, connects frames smoothly without overlapping and adding. To do this, each of the parameters must be smoothly connected and interpolated at frame boundaries.

Harmonic matching was the first parameter connection/smoothing topic for the STC non-overlapping method. Two methods for connecting harmonics were presented. First, harmonics may be matched directly by their harmonic number. For example, the first harmonic is always connected to the first harmonic, the second harmonic is always connected to the second harmonic, etc. This method resulted in speech with natural tonal quality, but audible high frequency sweeps when the pitch changed significantly. The second method, frequency matching, connects the harmonics with the closest frequencies without regard to harmonic number. This method resulted in speech with excellent tonal quality, but the algorithm for frequency matching is computational and complicated. The final method, picked as the best solution, was to match harmonics by harmonic number unless the fundamental frequency changes by greater than ten percent between frames. In that case, the harmonics are not matched. Instead, they are "birthed" in the previous frame and "die" in the next frame. This retained the natural tonal quality, removed the audible high frequency sweeps, and kept the computation and logical complexity to a moderately low level.

The second parameter connection/smoothing topic for the STC non-overlapping method was amplitude interpolation in time. Although linear interpolation is normally used, it was found to produce amplitude discontinuities when large changes in amplitude occurred. Three alternate methods were tested to try to eliminate this problem. The first, cubic amplitude interpolation, completely removed the discontinuities but resulted in overshoots and undershoots of ten percent or more. The second method, squaring the amplitudes, applying the cubic spline and then taking the square root of the spline, reduced the overshoot problem to about one percent, but created severe undershoots. The third method, linear interpolation with a moving average filter applied, proved to be the best solution. The discontinuities resulting from linear interpolation were removed, and no undesirable overshoots or undershoots were created.

The final topic of parameter connection/smoothing for the STC non-overlapping method was frequency/phase interpolation in time. To smoothly connect each harmonic in terms of the phase and frequency, STC cubic phase interpolation was used. Phase is interpolated cubicly so that the frequency trajectory has the smallest variation. The resulting frequency trajectory is quadratic.

After the parameter connection/smoothing issues were finalized for the STC non-overlapping method, a comparison was made between the overlap addition frame connection method and the STC non-overlapping method. Both produced equivalent high quality, smooth speech with very natural tonal quality. The deciding factor was then the computational difference between the two methods. Overlap addition requires no parameter smoothing, and the phase is interpolated linearly. However, up to two times as many frames (depending on the type of window taper) must be reconstructed because the frames are overlapped. The STC non-overlapping method requires only the reconstruction of the exact number of frames needed. However, further harmonic connection decisions as well as the smoothing of all parameters are required. A moving average filter must be applied to the linearly interpolated amplitudes, and the phase is interpolated cubicly. All of this together is very computationally expensive. As a result, the overlap addition method was selected for the final design.

A brief summary of the final enhanced sinusoidal reconstruction method follows:

1) A harmonic model serves as the basic synthesis structure.
    a) Frames with any voiced components are reproduced using a sinusoidal harmonic model with randomized phases for unvoiced sinusoids. Resampling in unvoiced areas is performed so that the spacing between sinusoids in unvoiced areas is less than 100 Hz.
    b) Completely unvoiced frames are reproduced using band-limited white noise as is done in MBE.
2) Phases are generated synthetically in the synthesizer using the cepstral phase model based on the sum of linear excitation phases and system phases.
3) The overlap addition method with full overlap and triangular tapering is used to smoothly connect frames. No additional parameter connection or smoothing is required.

Speech reconstructed using this enhanced sinusoidal reconstruction method is high quality, has very natural tonal quality, and is comparable to other reconstruction methods developed for mid-rate speech coding.

The reconstruction described in this paper has been fully implemented in a speech coder under development at Oklahoma State University, referred to as EMBE 8.0kbps. Informal testing results show that this coder is comparable to other mid-rate coders such as the 8.0kbps VSELP developed by Motorola [11]. However, the EMBE 8.0kbps coder is still being enhanced, and full parameter quantization is not yet completed. Therefore, definitive testing of this reconstruction method is incomplete. Final enhancements and quantization should be finished within the next few months. Once quantization is complete, objective testing will be performed so that appropriate comparisons can be made with other mid-rate coders. The objective testing will include Diagnostic Acceptability Measure (DAM) tests, which indicate the quality and naturalness of the reconstructed speech, and Diagnostic Rhyme Test (DRT) testing, which measures the intelligibility of the reconstructed speech.

Future Research

There are several issues not addressed in the reconstruction discussion of this paper. These include the application of a post-filter and the use of other spectral enhancements such as spectral warping and pre-emphasis and de-emphasis.

Probably the most important topic not discussed is post-filtering. Applying a post-filter to the reconstructed speech is necessary to enhance the intelligibility and overall quality of the synthetic speech. Post-filters typically lower the formant valleys to reduce the amount of coding noise. The application of a post-filter removes much of the "muffled" quality often inherent in synthetically generated speech and improves the objective test results.

Other spectral enhancements may also be beneficial. Enhancements such as pre-emphasis and de-emphasis may further reduce the noise floor and require less post-filtering than is otherwise needed. In addition, spectral warping may be beneficial for reducing the bit rate while maintaining the accuracy necessary for high quality reconstruction.

A word of caution is included regarding spectral alterations. Because the phases are generated from the cepstral phase model, alteration of the spectrum may alter the cepstral phases, resulting in possible degradation of tonal naturalness. The effects of such alteration are not known and should be considered when evaluating the results of additional spectral enhancements.

REFERENCES

[1] K. Teague, B. Leach and W. Andrews, "Development of a High-Quality MBE Based Vocoder for Implementation at 2400 bps," *Proceedings of IEEE Wichita Conference on Communications, Networking and Signal Processing*, pp. 129-33, April 1994.

[2] W. Andrews, *Design of a High Quality 2400 Bit Per Second Enhanced Multiband Excitation Vocoder*, Oklahoma State University, Stillwater, OK, 1994, M.S. Thesis.

[3] K. Teague, W. Andrews and B. Walls, "Harmonic Speech Coding at 2,400 bps", *10th Annual Mid-America Symposium on Emerging Computer Technologies*, 1996.

[4] B. Walls, *Enhanced Spectral Modeling for Sinusoidal Speech Coders*, Oklahoma State University, Stillwater, OK, 1996, M.S. Thesis.

[5] K. Teague, W. Andrews and B. Walls, "Enhanced Modeling of Discrete Spectral Amplitudes," *IEEE Workshop on Speech Coding for Telecommunications Proceedings*, pp. 13-4, Pocono Manor, PA, Sept 7-10 1997.

[6] K. Teague, W. Andrews, "Enhanced Spectral Modeling for MBE Speech Coders," to appear in *31st Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Nov 2-5, 1997.

[7] E. Daniel and T. Parker, "A High Quality 8kbps Enhanced Multiband Excitation Speech Coder," *8th Annual Graduate Student Research Symposium*, Oklahoma State University, March 1997.

[8] CCITT Yellow Book. *Recommendation G.721*, Vol. 3.

[9] J. C. Campbell, T. E. Tremain and V. C. Welch, "The Federal Standard 1016 4800 bps CELP Voice Coder," *Digital Signal Processing*, Vol. 1, 1991.

[10] R. Fenichel, "Federal Standard 1016, Telecommunications: Analog to Digital Conversion of Radio Voice by 4,800 bit/second Code Excited Linear Prediction (CELP)," *National Communications System, Office of Technology and Standards*, Washington, DC, 14 Feb 1991.

[11] I. Gerson and M. Jasiuk, "Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 8 kbits/s," *Proceedings of ICASSP-90*, pp. 461-4, New Mexico, April 1990.

[12] D. Griffin, *Multiband Coder*, MIT, Cambridge, MA, 1987, Ph.D. Dissertation.

[13] D. Griffin and J. Lim, "Multiband Excitation Vocoder," *IEEE ASSP*, pp. 1223-35, Vol. 36, No. 8, August 1988.

[14] J. Hardwick and J. Lim, "A 4800 bps Multiband Excitation Speech Coder," *Porceedings of ICASSP-88*, pp. 374-7, April 1988.

[15] Inmarsat Satellite Communications Services, "Inmarsat-M System Definition, Issue 3.0-Module 1: System Description," Nov 1991.

[16] APCO, "NASTD Federal Project 25 Vocoder: Version 1.0," Dec 1992.

[17] R. McAulay and T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE ASSP*, pp. 744-54, Vol. ASSP-34, No. 4, 1986.

[18] R. McAulay and T. Quatieri, "Sinusoidal Coding," in *Speech Coding and Synthesis* (W. B. Kleijn and K. Paliwal, eds.), pp. 123-70, Amsterdam, The Netherlands, Elsevier Science Publishers B. V., 1995.

[19] J. Deller, J. Proakis and J. Hansen, *Discrete-Time Processing of Speech Signals*, pp.434-58, Macmillan Publishing Company, New York, 1993.

[20] A. McCree and J. De Martin, "A 1.6 Kb/s MELP Coder for Wireless Communications," *IEEE Workshop on Speech Coding for Telecommunications Proceedings*, pp. 23-4, Pocono Manor, PA, Sept 7-10 1997.

[21] D. Hermes, "Pitch Analysis," in *Visual Representations of Speech Signals* (M. Cooke, S. Beet, and M. Crawford, eds.), pp. 3-25, John Wiley & Sons Ltd., 1993.

[22] J. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," *IEEE Transactions on Audio and Electroacoustics*, pp. 367-77, Vol. AU-20, No. 5, December 1972.

[23] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of IEEE*, pp. 561-580, Vol. 63, No. 4, April 1975.

[24] D. Griffin and J. Lim, "A New Pitch Detection Algorithm," in *Digital Signal Processing-84* (V. Capellini and A. Constantinides, eds.), pp.395-9, North Holland, Elsevier Science Publishers B. V., 1984.

[25] Y. Medan, E. Yair, and D. Chazan, "Super Resolution Pitch Determination of Speech Signals," *IEEE Transactions on Signal Processing*, pp. 40-8, Vol. 39, No. 1, January 1991.

[26] D. Griffin and J. Lim, "A New Model-Based Speech Analysis/Synthesis System," *IEEE Conference on Acoustics, Speech, and Signal Processing*, pp.513-6, Tampa, FL, March 1985.

[27] L. Van Immerseel and J. Martens, "Pitch and Voiced/Unvoiced Determination with an Auditory Model," *J. Acoustics Soc. Am.* 91 (6), June 1992.

[28] R. McAulay and T. Quatieri, "Pitch Estimation and Voicing Detection Based on a Sinusoidal Model," *Proceedings of ICASSP-90*, pp. 249-52, Albuquerque, NM, March 1990.

[29] J. Makhoul and J. Wolf, "Linear Prediction and Spectral Analysis of Speech," BBN Inc., Rep. 2304, 1972.

[30] Y. Medan and E. Yair, "Pitch Synchronous Spectral Analysis Scheme for Voiced Speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 1321-8, Vol. 37, No. 9, September 1989.

[31] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*, p.215-7, Prentice Hall, New Jersey, 1989.

[32] "Unwrap," *MATLAB Reference Guide*, p.534, The MathWorks, Inc., Natick, MA, August 1992.

[33] J. Deller, J. Proakis and J. Hansen, *Discrete-Time Processing of Speech Signals*, pp.352-401, Macmillan Publishing Company, New York, 1993.

2

VITA

Tabitha Joy Parker

Candidate for the Degree of

Master of Science

Thesis: ENHANCED SINUSOIDAL RECONSTRUCTION FOR A HIGH QUALITY, MID-RATE MBE SPEECH CODER

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Pine Bluff, Arkansas, December 1, 1972, the daughter of William and Carolyn Parker.

Education: Graduated from White Hall High School, White Hall, Arkansas in May 1991; received Bachelor of Science degree in Computer Engineering Technology from the University of Arkansas at Little Rock, Little Rock, Arkansas in May 1995. Completed the requirements for the Master of Science degree with a major in Electrical Engineering at Oklahoma State University, Stillwater, Oklahoma in May 1998.

Experience: Graduate Teaching Assistant, Oklahoma State University, August 1995 to December 1996. Hardware Design Engineer Intern, DCA Inc., Cushing, OK, Summer 1996. Graduate Research Assistant, Oklahoma State University, October 1996 to present.

Professional Memberships: IEEE, Golden Key National Honor Society, Phi Kappa Phi, Tau Alpha Pi.