Time-Scale Modification of Speech: A Time-Frequency Approach

Benoit Sylvestre

Department of Electrical Engineering
McGill University, Montreal
April 1991

A Thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Engineering.

Abstract

Time-scale modification (TSM) is a process whereby signals are compressed or expanded in time in a manner which preserves their original frequency characteristics. This work explores TSM algorithms for sampled speech. A known approach [2] which is based on the *short-time Fourier transform* (STFT) is first reviewed, then modified to provide high-quality TSM of speech signals at a lower computational cost. The proposed algorithm resembles the *sinusoidal speech model* (SSM) based approach [9], yet incorporates new phase compensatory measures to prevent excessive structural deterioration of the time-scaled signal. In addition, a novel incremental scheme for modifying polar parameters results in substantial computational savings.

Sommaire

La *modification d'échelle de temps* (MET) est un procédé qui effectue la compression et l'expansion temporelle de signaux sans déformer leurs caractéristiques fréquentielles. Cet ouvrage traite d'algorithmes MET pour la parole échantillonée. Une méthode connue [2] dont la théorie est fondée sur l'analyse de Fourier de courte durée est d'abord revisée, puis ensuite modifiée afin d'en améliorer l'efficacité. L'algorithme proposé est semblable à un procédé MET courant qui se sert d'un modèle sinusoïdal de la parole [9]. Cependant, le nouvel algorithme compense la phase des composantes fréquentielles servant à la synthèse selon une méthode inédite afin de préserver les signaux modifiés de déformations structurales excessives. En outre, une nouvelle technique différentielle conçue pour la modification de paramètres polaires économise considérablement le nombre de calculs.

Acknowledgements

Many thanks to Dr. P. Kabal, whose helpful comments, suggestions and technical advice resulted in significant improvements in the final version of this document. I am indebted to the Natural Sciences and Engineering Research Council (NSERC) of Canada as well as Bell-Northern Research (BNR), whose generous financial support made this project possible. I also extend my gratitude towards the people at the Institut national de la recherche scientifique (INRS) en télécommunications, students and staff alike, for their help regarding the use of various research tools.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This work explores a specific class of digital signal processing algorithms whereby speech signals are compressed or expanded along their time-axis in a manner which preserves their original frequency characteristics. A *time-scale modification* (TSM) system is capable of varying the playback rate of a digital audio recording without the frequency distortion which would result from linearly scaling its time-axis. Thus a listener perceives changes in the apparent *rate of articulation* of time-scaled speech, but not in the speaker-dependent features such as pitch and timbre. In the case of time-scaled music, only changes relating to *tempo* are perceived while the tone colors of the instruments remain the same.

Since audio information is conveyed by a combination of temporal and spectral features, the problem of modifying time independently of frequency is indeed challenging, as the two dimensions cannot be easily decoupled: linearly scaling the time-axis by β corresponds to linearly scaling the frequency-axis by $1/\beta$ and vice versa. Such is the nature of the fundamental paradox which must be addressed by TSM algorithms.

Remarkably few papers have been published on the subject, a fact which does not reflect the true potential of TSM technology. While the cost of digital audio is dropping rapidly, *time* is becoming an increasingly expensive resource. Profession-

als and consumers alike may eventually expect TSM features to be incorporated into every conceivable digital audio reproduction device ranging from compact disc (CD) players to reading machines for the blind. TSM could even be used for audio data compaction or for more esoteric applications, such as synchronizing soundtracks to films and videos. The technology has yet to be perfected and, not unlike speech coding, challenges our understanding of the structural properties of sound. The lack of consensus in the literature is ample proof that some important questions remain unanswered.

Early efforts ([1], 1954) achieved rate-changed speech by selectively inserting or deleting periodic waveform segments in or from the speech signal. Only voiced portions could be effectively time-scaled. Moreover, the output was susceptible to discontinuities and other artifacts due to the sensitivity of the method to pitch variations in the source signal. Current integrated circuit technology now permits more sophisticated solutions to be considered.

Portnoff's 1981 paper [2] represents a significant milestone in TSM research. Using short-time Fourier analysis principles, the author developed a speech model in which time could be manipulated independently of frequency. Portnoff hinted at the fact that the proposed TSM scheme could not "preserve the structure of [...] an arbitrary signal", yet no further mention of this limitation was made in the simulation portion of the paper. Either Portnoff's rhetoric discouraged further investigation in this direction or the paper was considered final and authoritative, for no other related contribution could be found in the literature spanning the 1981–1984 period.

A later paper, authored by Griffin and Lim ([3], 1984), presented a radically different approach to the problem. Their method consisted of estimating the desired rate-changed speech by minimizing, in the mean-square sense, the euclidean distance between the short-time Fourier transform of the original and the rate-

changed sequences. The so-called ***least-squares error estimation*** (LSEE) algorithm was expensive in that it iteratively searched for the optimal fit for the short-time Fourier transform of the original sequence on the new time-scale. Each iteration required a forward *and* inverse discrete Fourier transform (DFT) computation. In addition, the rate of convergence depended largely on the accuracy of the initial estimate of the rate-changed sequence, thus on the statistics of the original. Griffin and Lim stated that the resulting time-scaled speech "appear[ed] to be superior" in quality to that generated by Portnoff's method. The nature of the subjective performance improvements was not specified in the simulation portion of the paper. Nevertheless, the Griffin and Lim contribution prompted further research efforts in this direction.

Roucos and Wilgus ([4], 1985) accelerated the rate of convergence of the LSEE algorithm by improving the initial estimate of the rate-changed speech. Until then only noise and crude LPC estimates had been used for initializing the algorithm. The new approach, dubbed the *synchronized overlap-add* (SOLA) algorithm, consisted of calculating an initial estimate which best matched the original speech data in the vicinity of the sample instant of interest.

The LSEE concept was more recently adapted by Abe *et al.* ([5], 1989) for simultaneous modification of pitch and duration of speech. Hardam ([6], 1990) claimed to having improved the subjective performance of the SOLA algorithm while reducing its complexity. Asi and Saleh ([7], 1988) applied the LSEE concept with an interesting twist: they showed that the rate-changed speech could be related to the original speech by a linear, periodically time-varying filter. The greatest virtue of this method was its unsurpassed simplicity. Its performance, however, was observed to be quite sensitive to pitch variations. Asi and Saleh later adapted their filtering theory to allow simultaneous scaling of time and frequency ([8], 1990).

One Portnoff-like approach can be noted in the wake of the apparent LSEE trend. McAulay and Quatieri ([9], 1986) demonstrated how time and frequency transformations could be applied to their particular *sinusoidal speech model* (SSM) [10]. The TSM principle they used in conjunction with their SSM was essentially the same as that proposed by Portnoff because both approaches have common origins. It would be fair to say that the potential of SSM technology in the area of low-bit rate speech coding as well as speech modification has somewhat rekindled general interest in time-frequency representations.

Despite the relative sparseness of TSM literature, one is confronted with an otherwise rich assortment of theoretical views and algorithm complexities. There are, on one hand, techniques which make virtually no assumption about the structure of the waveform to be processed. Their overall appeal lies in the fact that many common speech modification objectives can be achieved by way of simple, robust algorithms, albeit with often expensive high-speed hardware. Then there are those which isolate by way of rigorous or heuristic models the desired features of a waveform for subsequent modification. The latter methods may sometimes reduce hardware speed requirements, but only at the expense of more complex software and hardware configurations. The performance of such systems usually depends much on source model accuracy. Fortunately, the modeling of speech is in general successful due to the primarily resonant nature of the vocal tract. By comparison, the modeling of arbitrary sources such as music is a much more arduous task and will often resort to the axioms of speech production. In spite of their shortcomings, source models can be finely tuned to the perceptual characteristics of the human auditory system to improve performance. It has become more widely accepted that perception science is playing an increasingly important role in the field of audio coding [11].

The subjective performance of LSEE algorithms is undoubtedly limited by the

fact that they are insensitive to the structure of the processed signal. Furthermore, the performance measures which drive them are completely arbitrary. The LSEE option, though popular, provides no assurance that spreading sequence estimation error uniformly over time and frequency is appropriate for all types of waveforms. The very existence of noise shaping audio coders [12] would indicate otherwise.

No strong statement can be made in favor of the time-frequency approach either because the structural impact of TSM on arbitrary waveforms has never been assessed. Portnoff anticipated some sort of structural distortion but produced no experimental data in this regard. Structural distortion was observed by McAulay and Quatieri who attributed the fault to speech parameter estimation errors, not to the TSM procedure itself.

Assuming some degree of structural distortion is unavoidable, can the performance of time-frequency based TSM algorithms still rival that of their iterative counterparts at a lesser computational cost? This question warrants further investigation, given the potential benefits of time-frequency representations and the relative computational simplicity of SSM-based speech modification algorithms.

We propose, therefore, to first study Portnoff's TSM method more closely and then determine the exact nature of the aforementioned waveform deterioration. It will be shown that the accumulation of phase error is chiefly responsible for deteriorating rate-changed speech over time. Novel methods which restrict the distortion within acceptable perceptual bounds will be proposed along with several important simplifications to Portnoff's original design. *Phase modulation* will be the preferred method in the final design. The resulting time-frequency TSM algorithm, which will combine a new *incremental* parameter modification scheme with the advantages of polar synthesis [10], will be more robust than its predecessors in the way just described, while affording high-quality rate-changed speech at a computational cost comparable, if not less, to that of the SSM-based version.

The development is organized as follows: Chapter 2 formulates the TSM problem in mathematical terms and identifies the underlying causes of structural deterioration in rate-changed speech; Chapter 3 discusses several design options for a practical TSM system and outlines the final implementation; Chapter 4 reports the results obtained for a computer simulation; Chapter 5 summarizes the key results and suggests other research avenues.

Chapter 2

Theory of Time-Scale Modification

In this chapter we establish the mathematical foundation for a short-time Fourier analysis based TSM system. Following a statement on the definitions and the notation conventions to be used throughout and a brief review of linear time-scaling, we develop a time-frequency model for speech and derive from it an expression for the desired rate-changed speech. The *short-time Fourier transform* (STFT) is then considered as a means for estimating the parameters required for the synthesis of rate-changed speech. In closing, it is shown that some degree of distortion is unavoidable under the proposed TSM approach.

2.1 Definitions

Discrete-Time Sequences and Transforms

The discrete-time sequence or signal $x(n)$ (where n is integer-valued) represents the samples of a continuous-time, bandlimited waveform $x(t)$ with the sampling interval normalized to unity. The Fourier transform of $x(n)$, denoted by $X(\omega)$ where ω is continuous, is defined as [13]

$$X(\omega) = \sum_{n=-\infty}^{+\infty} x(n)e^{-j\omega n}. \tag{2.1}$$

The function $X(\omega)$ is periodic in ω with period 2π. The original sequence is recoverable via the inversion formula [13]

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} X(\omega)e^{j\omega n}\, d\omega. \tag{2.2}$$

If $t(n,m)$ represents the samples of a time-varying system, where the system response is viewed as a function of m at time n, it will be convenient to define its *time-varying* Fourier transform as

$$T(n,\omega) = \sum_{m=-\infty}^{+\infty} t(n,m)e^{-j\omega m} \tag{2.3}$$

and the corresponding inverse formula as

$$t(n,m) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} T(n,\omega)e^{j\omega m}\, d\omega. \tag{2.4}$$

In many instances, the time-varying Fourier transform will also be regarded as a sequence in n, with ω treated as a parameter or an index which further distinguishes the sequence. Thus the frequency dimension corresponding to the sequence $T(n,\omega)$ is *independent* of ω. In general, the temporal features or time-varying parameters of a sequence $f(n,\omega,\ldots)$ appear as a function of n, whereas its spectral features or frequency-varying parameters, as a function of ω.

Time-Scaling

The time-scaling factor of a sequence is represented by some rational number β. The range $\beta > 1$ corresponds to time-scale *compression*, and the range $0 < \beta < 1$, to time-scale *expansion*. The time-scaled or *rate-changed* version of a sequence $f(n,\ldots)$ will usually be written as $f^{\beta}(n,\ldots)$. This notation should not to be confused with $f(\beta n,\ldots)$, which denotes the *linearly* time-scaled version of $f(n,\ldots)$. In general, a rate-changed sequence will be obtained through non-linear means.

2.2 Linear Time-Scaling

One of the goals of this section is to illustrate the interdependence of time and frequency by means of a continuous-time and two discrete-time examples. In so doing, we review the fundamental results of classical decimation and interpolation theory which will subsequently prove useful.

Continuous-Time

Linearly scaling the time-axis of a continuous-time waveform $x(t)$ by β corresponds to linearly scaling its frequency-axis by $1/\beta$ for any $\beta \neq 0$. This is apparent from the continuous-time inverse Fourier expression

$$\begin{aligned} x(\beta t) &= \int_{-\infty}^{+\infty} X(f) e^{j2\pi f \beta t}\, df \\ &= \int_{-\infty}^{+\infty} X(f/\beta) e^{j2\pi f t}\, df/\beta. \end{aligned} \tag{2.5}$$

The discrete-time result, though restricted by sampling rate considerations, is conceptually the same. For illustration, we examine the Fourier representations for two special cases

Discrete-Time

Consider an integer time-scale compression factor β_c satisfying the condition $\beta_c > 1$. Let the compressed sequence $x_c(n) = x(\beta_c n)$ be obtained by discarding $\beta_c - 1$ samples from $x(n)$ at intervals of β_c samples. The corresponding Fourier transform is [14]

$$X_c(\omega) = \frac{1}{\beta_c} \sum_{i=0}^{\beta_c - 1} X\left(\frac{\omega - 2\pi i}{\beta_c}\right). \tag{2.6}$$

The function $X_c(\omega)$ is formed by overlapping shifted and rescaled copies of the original Fourier transform $X(\omega)$.

If the resulting sampling rate of $x_c(n)$ is greater than the Nyquist rate of $x(n)$, the original sequence $x(n)$ is in principle recoverable from $x_c(n)$. The sampling rate of a sequence can be reduced by a factor β without aliasing provided the original sampling rate is at least β times greater than the Nyquist rate of that sequence. Stated in terms of the (normalized) bandwidth ω_x of $x(n)$, the equivalent condition is

$$\omega_x < \pi/\beta. \tag{2.7}$$

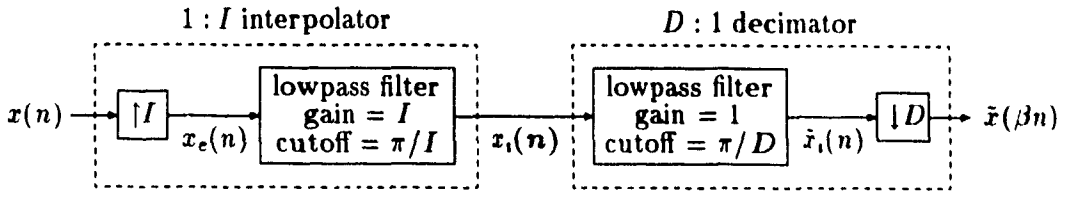
An anti-aliasing filter will normally be applied to $x(n)$ before time compression. The operation which consists of reducing the sampling rate of a sequence with pre-filtering is referred to as downsampling or *decimation*. We now address another important case.

Consider a time-scale expansion factor β_e satisfying the condition $0 < \beta_e < 1$ and whose reciprocal is an integer. Let the expanded sequence $x_e(n) = x(\beta_e n)$ be obtained by inserting $1/\beta_e - 1$ zeros between the samples of $x(n)$. The corresponding Fourier transform is [14]

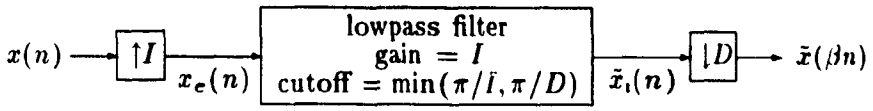
$$X_e(\omega) = X(\omega/\beta_e). \tag{2.8}$$

In this case, the function $X_e(\omega)$ is obtained directly by compressing the frequency-axis by $1/\beta_e$.

Aliasing is of no concern here because the sampling rate is being increased. However, the sequence $x_e(n)$ is of no practical use in its present form, due to its zero samples. Non-zero samples are substituted in their place by removing from $X_e(\omega)$ all frequency scaled images of $X(\omega)$ except at integer multiples of 2π.



(a)



(b)

Figure 2.1: a) System for changing the sampling rate by a non-integer factor $\beta = D/I$. b) Simplified system in which the decimation and the interpolation filters are combined. (After Oppenheim and Schafer [14].)

The operation which consists of increasing the sampling rate of a sequence with post-filtering is referred to as upsampling or *interpolation.*

The above Fourier representations hold only for integer β_c and $1/\beta_e$ factors. However, the decimation and interpolation procedures need not be restricted to integer β and $1/\beta$ factors.

Figure 2.1 depicts a system which combines the above decimation and interpolation techniques to modify the sampling rate of a sequence by the effective ratio D/I. By choosing integers such that D/I is arbitrarily close to β, we may approximate virtually any sampling rate conversion factor. The system configuration shown in Figure 2.1 is the best arrangement for preserving the bandwidth structure of $x(n)$ while allowing both the pre- and the post-filters to be combined. The linearly time-scaled sequence $x(\beta n)$ can be obtained directly from $x(n)$ via the convolution

$$x(\beta n) = \sum_{r=-\infty}^{+\infty} f(nD - rI)x(r), \tag{2.9}$$

where $f(n)$ is a $1 : I$ interpolating filter. The linear time-scaling operation specified by (2.9) will be referred to as "bandlimited interpolation".

Other filtering methods exist for implementing linear time-scaling systems for arbitrary β factors [15].

2.3 Speech Model

A speech signal is characterized by a sequence of air pressure waves modulated by vocal tract movements [16]. Speech is often viewed in many coding, enhancement and modification applications as the response of a linear time-varying filter to an excitation source. The filter approximates the time-varying spectral characteristics of the vocal tract whereas the excitation source may be a periodic signal, resulting in *voiced* speech, or a noisy and aperiodic one to produce *unvoiced* speech. The major spectral difference between both speech classes is that unvoiced speech has no underlying harmonic structure as does voiced speech. Voicing transitions are not covered in the subsequent analysis; the impact of this omission will be examined in Section 2.7.

Figure 2.2 illustrates the model which will serve as the basis for the mathematical representation of speech developed in this section. For voiced speech, the excitation source $v(n)$ consists of a train of unit-samples where the unit-sample spacing corresponds to the local pitch period $P(n)$ of the actual speech. Viewed as a sequence, $P(n)$ is *slowly* time-varying in the sense we shall describe shortly. The excitation source $u(n)$ for unvoiced speech is assumed to be white noise, that is, a stationary random sequence having a flat spectrum. The spectral characteristics of the glottal source and the lossy acoustic behavior of the vocal tract are represented

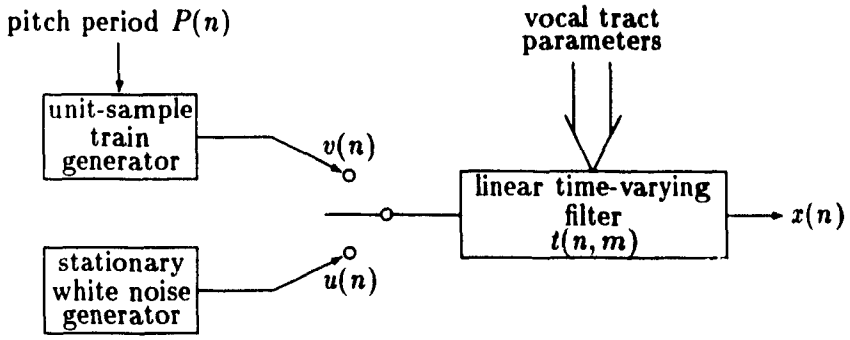


Figure 2.2: Terminal-analog model of the vocal system. (After Schafer and Rabiner [18].)

by the time-varying, finite-length filter $t(n,m)$. We recall that the filter response is viewed as a function of m at time n.

For many speech sounds, it is reasonable to assume that the speech parameters remain fixed over 10–20ms intervals [17]. Changes in vocal tract geometry and articulator movement account for the time-varying nature of $t(n,m)$. These variations, as those in $P(n)$, are assumed to be slower than the decay rate of the impulse response of the vocal tract [19]. Consequently, the filter is said to be *slowly* time-varying, or nearly fixed for the duration of its memory M_v, and is referred to as a "quasi-stationary" system. Likewise, the time-varying Fourier transform of $t(n,m)$, which is defined as in (2.3), and the pitch period $P(n)$ are assumed to be nearly fixed over M_v. Since $P(n)$ is slowly time-varying, the impulse train $v(n)$ is said to be "quasi-periodic".

We now address the problem of formulating a mathematical representation for each speech class. The remainder of this section is based upon the development found in [19].

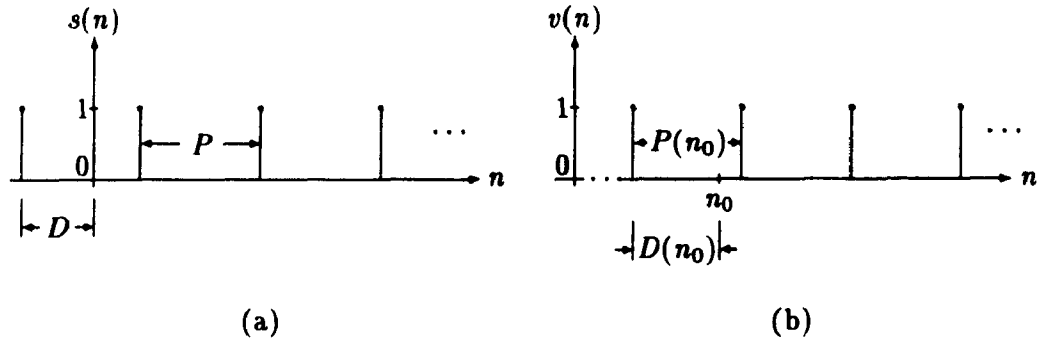


Figure 2.3: a) Periodic unit-sample train. b) Quasi-periodic unit-sample train. (After Portnoff [19].)

2.3.1 Voiced Speech

It will be shown that a voiced speech signal $x(n)$ can be expressed in terms of a linear combination of time-varying complex exponentials. Specifically, the excitation source $v(n)$ will be modeled as a sum of exponentials and the final expression for $x(n)$ will be obtained through superposition, since the speech output is assumed to be the response of a linear filter.

We begin by formulating a representation for the periodic unit-sample train shown in Figure 2.3a). The goal is to establish the mathematical form of $v(n)$. Let

$$s(n) = \sum_{r=-\infty}^{+\infty} \delta(n + D - rP), \tag{2.10}$$

where the integer P denotes the period and the integer D, the distance of the first unit-sample appearing to the left of the time origin. The unit-sample is defined as

$$\delta(n - n_0) = \begin{cases} 1 & \text{for } n = n_0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

The sequence $s(n)$ is said to be periodic in n with period P. The Fourier series representation for (2.10) is [20]

$$s(n) = \frac{1}{P} \sum_{k=0}^{P-1} \exp\left[j2\pi k(n + D)/P\right]. \tag{2.12}$$

The harmonic representation for the quasi-periodic unit-sample sequence $v(n)$ shown in Figure 2.3b) can be approximated by a similar expression.

The term "quasi-periodic" implies that $v(n)$ has a locally periodic behavior. The sequence $v(n_0 + \ell)$ is assumed to be periodic in ℓ, an integer variable, with period $P(n_0)$ in the vicinity of n_0 and can be locally represented as [19]

$$v(n_0 + \ell) \approx \sum_{r=-\infty}^{+\infty} \delta\Big(\ell + D(n_0) - rP(n_0)\Big). \tag{2.13}$$

If n_0 is viewed as a variable time origin, $D(n_0)$ and $P(n_0)$ are the respective time-varying analogs of D and P.

These quantities are assumed to be nearly fixed over the observation interval spanned by ℓ, which includes the duration of the vocal tract filter memory M_v. We therefore postulate

$$D(n_0 + \ell) \approx D(n_0) \tag{2.14}$$

$$P(n_0 + \ell) \approx P(n_0) \tag{2.15}$$

over the range

$$|\ell| < M_v/2. \tag{2.16}$$

This "local stationarity" assumption will often be exploited in the sequel.

The local harmonic representation for (2.13), given the above conditions on $D(n_0)$ and $P(n_0)$, follows directly from (2.12), i.e.

$$\begin{aligned} v(n_0 + \ell) &\approx \frac{1}{P(n_0)} \sum_{k=0}^{P(n_0)-1} \exp\Big[j2\pi k\Big(\ell + D(n_0)\Big)/P(n_0)\Big] \\ &= \frac{1}{P(n_0)} \sum_{k=0}^{P(n_0)-1} \exp\Big[jk\Big(\phi(n_0) + \Omega(n_0)\ell + \phi_0\Big)\Big], \end{aligned} \tag{2.17}$$

where

$$\Omega(n_0) = 2\pi/P(n_0) \tag{2.18}$$

$$\phi_0 = \Omega(0)D(0) \tag{2.19}$$

$$\phi(n_0) = \Omega(n_0)D(n_0) + 2\pi I(n_0) - \phi_0. \tag{2.20}$$

The quantity $\phi(n_0)$ is referred to as the ***instantaneous phase*** of the excitation source whereas $\Omega(n_0)$ represents the local pitch or ***fundamental frequency.*** The integer number $I(n_0)$ guarantees the uniqueness of the exponential phase argument over time as in (2.12) and has the initial condition $I(0) = 0$. The constant ϕ_0 will allow the phase offset at the time origin to be preserved under a rate-change modification by ensuring $\phi(0) = 0$.

Due to its dependence on $P(n_0)$, the fundamental frequency is also slowly time-varying, i.e.

$$\Omega(n_0 + \ell) \approx \Omega(n_0) \qquad \text{for } |\ell| < M_v/2. \tag{2.21}$$

The exponential phase argument of (2.17) may be regarded as the instantaneous phase corresponding to the specific time instant $n_0 + \ell$. Consequently, the local representation for $\phi(n_0)$ is

$$\phi(n_0 + \ell) \approx \phi(n_0) + \Omega(n_0)\ell \qquad \text{for } |\ell| < M_v/2. \tag{2.22}$$

Setting $\ell = -1$ leads to a convenient first backward difference approximation of the fundamental frequency,

$$\Omega(n_0) \approx \phi(n_0) - \phi(n_0 - 1). \tag{2.23}$$

The recursive structure of $\phi(n_0)$ suggests that the instantaneous phase can be constructed by summing the fundamental frequency values over all time. We therefore define

$$\phi(n) = \sum_{r=1}^{n} \Omega(r), \tag{2.24}$$

with the initial condition $\phi(0) = 0$. The instantaneous phase $\phi(n)$ will be referred to as an *unwrapped* phase sequence because it is uniquely defined over time.

Letting $n = n_0 + \ell$ and applying equations (2.15) and (2.22) to the local harmonic representation given by (2.17), we obtain a final model for the voiced speech excitation,

$$v(n) = \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} \exp\left[jk\Big(\phi(n) + \phi_0\Big)\right]. \tag{2.25}$$

Applying this excitation to the vocal tract filter $t(n,m)$ generates the voiced speech signal $x(n)$ by way of the superposition sum

$$x(n) = \sum_{m=-\infty}^{+\infty} t(n,m)v(n-m). \tag{2.26}$$

Substituting the harmonic representation for $v(n-m)$ into this expression (with $\ell = -m$), and using the local stationarity assumption given by (2.15), we obtain

$$x(n) \approx \sum_{m=-\infty}^{+\infty} t(n,m) \left\{ \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} \exp\left[jk\Big(\phi(n) - \Omega(n)m + \phi_0\Big)\right] \right\}. \tag{2.27}$$

Interchanging the order of summation and substituting (2.3) into the above expression yields

$$x(n) \approx \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} T\Big(n, k\Omega(n)\Big) \exp\left[jk\Big(\phi(n) + \phi_0\Big)\right]. \tag{2.28}$$

A more concise form for modeling voiced speech as a linear combination of harmonically related complex exponentials is [19]

$$x(n) = \sum_{k=0}^{P(n)-1} c_k(n) \exp\left[jk\phi(n)\right], \tag{2.29}$$

where

$$c_k(n) = \frac{1}{P(n)}\, T\big(n, k\Omega(n)\big) \exp\left[jk\phi_0\right] \tag{2.30}$$

$$\phi(n) = \begin{cases} \sum_{r=1}^{n} \Omega(r) & \text{for } n > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.31}$$

Because of their dependence on slowly time-varying quantities, the complex harmonic amplitudes $c_k(n)$ are also slowly time-varying and so may be regarded as narrowband sequences. Their bandwidths are much less than the fundamental frequency $\Omega(n)$, a property which will prove useful in a subsequent STFT analysis.

2.3.2 Unvoiced Speech

Since unvoiced speech has no underlying harmonic structure, speech events in this case are best characterized in terms of their second order statistics. A relationship between the time-varying power spectrum of the unvoiced speech signal $x(n)$ and the time-varying parameters of the vocal tract filter $t(n, m)$ will be established.

The excitation source $u(n)$ of Figure 2.2 is assumed to be a zero-mean, white noise sequence, meaning that the random variables $u(n_1)$ and $u(n_2)$ are uncorrelated for every $n_1 \neq n_2$ [21]. The auto-correlation sequence for $u(n)$ is

$$\begin{aligned} R_u(\ell) &= E\left[u(n+\ell)u^*(n)\right] \\ &= \sigma_u^2\, \delta(\ell), \end{aligned} \tag{2.32}$$

where the integer variable ℓ denotes a local time interval, $E[\cdot]$ is the expected value operator, * denotes complex conjugate and σ_u^2 is the variance of $u(n)$.

As in the voiced speech case, the unvoiced speech signal $x(n)$ can be evaluated from the superposition sum

$$x(n) = \sum_{m=-\infty}^{+\infty} t(n, m)u(n-m). \tag{2.33}$$

The resulting sequence has zero mean and is non-stationary, yet will be referred to as a "quasi-stationary" random process to reflect the fact that it is the output of a quasi-stationary linear system.

The time-varying auto-correlation sequence for $x(n)$ is given by

$$\begin{aligned} R_x(n,\ell) &= E\left[x(n+\ell)x^*(n)\right] \\ &= E\left[\sum_{q=-\infty}^{+\infty} t(n+\ell,q)u(n+\ell-q)\sum_{m=-\infty}^{+\infty} t^*(n,m)u^*(n-m)\right]. \end{aligned} \tag{2.34}$$

Since the $E[\cdot]$ operator is distributive and the variations in $t(n,m)$ are presumed negligible over the correlation interval ℓ, that is,

$$t(n+\ell,m) \approx t(n,m) \qquad \text{for } |\ell| < M_v/2, \tag{2.35}$$

(2.34) becomes

$$\begin{aligned} R_x(n,\ell) &\approx \sum_{q=-\infty}^{+\infty}\sum_{m=-\infty}^{+\infty} t(n,q)t^*(n,m)E\left[u(n+\ell-q)u^*(n-m)\right] \\ &= \sum_{q=-\infty}^{+\infty}\sum_{m=-\infty}^{+\infty} t(n,q)t^*(n,m)\,\sigma_u^2\delta(\ell-q+m) \\ &= \sum_{m=-\infty}^{+\infty} \sigma_u^2\, t(n,m+\ell)t^*(n,m). \end{aligned} \tag{2.36}$$

The last expression is a convolution with respect to ℓ with n treated as a parameter. Applying the linear system convolution property [22] to (2.36) yields

$$R_x(n,\ell) \approx \frac{1}{2\pi}\int_{-\pi}^{+\pi} \sigma_u^2\,|T(n,\omega)|^2\, e^{j\omega\ell}\, d\omega. \tag{2.37}$$

The time-varying power spectrum of $x(n)$ may be defined as

$$S_x(n,\omega) = \sigma_u^2\,|T(n,\omega)|^2\,. \tag{2.38}$$

Unvoiced speech will therefore be modeled as a quasi-stationary zero-mean random process which is characterized by its second moment $R_x(n,\ell)$ or, equivalently, by its time-varying power spectrum $S_x(n,\omega)$. Both sequences are assumed to be slowly time-varying.

2.4 Representation of Rate-Changed Speech

We now investigate how the respective models obtained for voiced and unvoiced speech can be modified to represent the desired rate-changed speech $x^{\beta}(n)$. In each case, $x^{\beta}(n)$ can be modeled as the output of a rate-changed, time-varying vocal tract filter $t^{\beta}(n,m)$ driven by the appropriate rate-changed excitation source, either $v^{\beta}(n)$ or $u^{\beta}(n)$.

The rate-changed version of a sequence can be obtained by linearly scaling its n-axis only in cases where the modification does not affect the spectral features of the resulting speech. For example, the following definition for the rate-changed vocal tract filter

$$t^{\beta}(n,m) = t(\beta n, m) \tag{2.39}$$

affects only the temporal features of $t(n,m)$, not the frequency response of the filter at time βn. The corresponding rate-changed Fourier transform is

$$T^{\beta}(n,\omega) = T(\beta n, \omega). \tag{2.40}$$

The rate-changed versions of the local pitch period $P(n)$ and the fundamental frequency $\Omega(n)$ are also obtained in this way. The effect of linearly time-scaling an excitation source, however, varies according to the structure under consideration.

The remainder of this section is based upon the development found in [2].

2.4.1 Voiced Speech

The model for the rate-changed voiced speech excitation $v^{\beta}(n)$ is expected to parallel (2.25), namely

$$v^{\beta}(n) = \frac{1}{P(\beta n)} \sum_{k=0}^{P(\beta n)-1} \exp\left[jk\left(\phi^{\beta}(n) + \phi_0\right)\right]. \tag{2.41}$$

The quantity $\phi^{\beta}(n)$ is the unknown rate-changed instantaneous phase sequence.

Scaling both n and ℓ by β and letting $\ell = -1$ in the local phase representation given by (2.22), we obtain

$$\Omega(\beta n_0) \approx \frac{1}{\beta}\left[\phi(\beta n_0) - \phi\big(\beta(n_0 - 1)\big)\right]. \tag{2.42}$$

The quantity $\Omega(\beta n_0)$ is recognized as the rate-changed version of $\Omega(n_0)$. In accordance with the instantaneous phase definition given by (2.24), we write

$$\begin{aligned}\phi^{\beta}(n) &= \sum_{r=1}^{n} \Omega^{\beta}(r) \\ &\approx \sum_{r=1}^{n} \frac{1}{\beta}\left[\phi(\beta r) - \phi\big(\beta(r-1)\big)\right] \\ &= \frac{\phi(\beta n)}{\beta},\end{aligned} \tag{2.43}$$

with the initial condition $\phi^{\beta}(0) = 0$. An estimate of the rate-changed instantaneous phase is obtained by resampling $\phi(n)$ and multiplying its values by $1/\beta$ to preserve the local pitch. Substituting this result into (2.41) yields the rate-changed excitation source

$$v^{\beta}(n) \approx \frac{1}{P(\beta n)} \sum_{k=0}^{P(\beta n)-1} \exp\left[jk\big(\phi(\beta n)/\beta + \phi_0\big)\right]. \tag{2.44}$$

By substituting (2.44) and (2.39) into the superposition sum given by (2.26), we obtain an expression for the rate-changed voiced speech,

$$x^{\beta}(n) \approx \sum_{k=0}^{P(\beta n)-1} c_k(\beta n) \exp\left[jk\phi(\beta n)/\beta\right], \tag{2.45}$$

where

$$c_k(\beta n) = \frac{1}{P(\beta n)}\, T\big(\beta n, k\Omega(\beta n)\big) \exp\left[jk\phi_0\right]. \tag{2.46}$$

The only non-linear transformation involved in obtaining rate-changed voiced speech is the one affecting the instantaneous phase $\phi(n)$ of the excitation source.

In general, the factor $1/\beta$ will not be an integer. It is therefore important that $\phi(n)$ be accurately represented over time, as extraneous multiples of 2π added to or removed from $\phi(n)$, though invisible on the original time-scale, will normally affect (2.45).

2.4.2 Unvoiced Speech

We see from (2.38) that the spectral features of the time-varying power spectrum $S_x(n,\omega)$ are not altered by a linear time-scaling operation. Thus rate-changed unvoiced speech may be characterized by the following time-varying power spectrum

$$\begin{aligned} S_x^{\beta}(n,\omega) &= \sigma_u^2 |T(\beta n,\omega)|^2 \\ &= S_x(\beta n,\omega). \end{aligned} \tag{2.47}$$

The corresponding time-varying auto-correlation is

$$\begin{aligned} R_x^{\beta}(n,\ell) &\approx \frac{1}{2\pi}\int_{-\pi}^{+\pi} S_x(\beta n,\omega)e^{j\omega\ell}\,d\omega \\ &\approx R_x(\beta n,\ell) \end{aligned} \tag{2.48}$$

over the correlation interval $|\ell| < M_v/2$. Thus the rate-changed excitation source

$$u^{\beta}(n) = u(\beta n) \tag{2.49}$$

preserves the local statistics of the original unvoiced speech, while the rate-changed filter $t^{\beta}(n,m)$ scales its time-varying parameters. In contrast to the voiced speech case, only linear transformations are required to generate rate-changed unvoiced speech.

2.5 Short-Time Fourier Analysis of Speech

Now that the respective mathematical models for voiced and unvoiced speech have been established, the problem of estimating their parameters for subsequent

modification remains. Due to the time-varying nature of the spectral quantities involved, the STFT should prove adequate for accomplishing this objective, even if major speech parameters such as pitch and vocal tract frequency response cannot be exactly determined.

We begin by defining the STFT and then examine an interpretation of the STFT which will serve the ensuing analysis of each speech class.

2.5.1 STFT Definition and Interpretation

A special case of the time-varying Fourier transform, the STFT of a discrete-time sequence $x(n)$ is defined as [23]

$$X(n,\omega) = \sum_{m=-\infty}^{+\infty} h(n-m)x(m)e^{-j\omega m}, \tag{2.50}$$

where $h(n)$ is a window sequence of finite length M, commonly referred to as an *analysis* filter. Like the Fourier transform, the STFT is periodic in ω with period 2π. Given the Fourier transform pair

$$h(n-m)x(m) \stackrel{\mathcal{F}}{\Longleftrightarrow} X(n,\omega), \tag{2.51}$$

the inverse STFT definition follows directly from (2.2) with $m = n$,

$$x(n) = \frac{1}{2\pi h(0)} \int_{-\pi}^{+\pi} X(n,\omega)e^{j\omega n}\, d\omega. \tag{2.52}$$

The condition $h(0) \neq 0$ must be satisfied for $x(n)$ to be recoverable. All samples of $x(n)$ which are multiplied by non-zero $h(n)$ samples can be recovered the same way. A more general inverse STFT formula incorporates a *synthesis* filter $f(n)$ which shapes the real and imaginary parts of $X(n,\omega)$ prior to synthesis,

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \sum_{r=-\infty}^{+\infty} f(n-r)X(r,\omega)e^{j\omega n}\, d\omega. \tag{2.53}$$

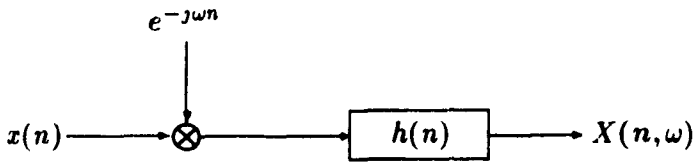


Figure 2.4: Short-time Fourier transform viewed as the output of a demodulator followed by an analysis filter. (After Portnoff [19].)

The constant $1/h(0)$ has been absorbed into $f(n)$.

The most useful STFT interpretation for our impending analysis views $X(n,\omega)$ as a time-varying parameter signal (with ω treated as a parameter or sequence index) of the form [24]

$$Q(n) = \sum_{m=-\infty}^{+\infty} Z[x(m)]h(n-m), \tag{2.54}$$

where $Z[\cdot]$ is some transformation designed to extract a specific signal property from $x(n)$. When convolved with the analysis filter $h(n)$, $Z[x(n)]$ yields a parameter signal whose spectral characteristics depend on $h(n)$ and the transformation itself.

In STFT analysis, $Z[\cdot]$ corresponds to a demodulator which shifts the spectral energy of $x(n)$ downwards by ω radians. If $h(n)$ is a narrowband lowpass filter, $X(n,\omega)$ is also a narrowband lowpass signal whose temporal features vary according to those of $x(n)$ in the neighborhood of ω. The system configuration corresponding to the interpretation just described is illustrated in Figure 2.4. The output satisfies

$$X(n,\omega) = h(n) *_n x(n)e^{-j\omega n}, \tag{2.55}$$

where $*_n$ denotes a convolution with respect to n.

If $x(n)$ were applied to a bank of N such systems operating in parallel with the analysis frequencies set at $\omega_k = 2\pi k/N$, the collective output would consist of N narrowband lowpass sequences $X(n,\omega_k)$ which embody the temporal features of $x(n)$ in the neighborhood of harmonically related frequencies.

2.5.2 Voiced Speech

By substituting (2.29) into the STFT formula, we obtain an expression for the STFT of a voiced speech signal $x(n)$ in terms of its harmonic representation,

$$X(n,\omega) = \sum_{m=-\infty}^{+\infty} \sum_{k=0}^{P(m)-1} h(n-m)c_k(m)\exp\left[jk\phi(m)\right] e^{-j\omega m}. \tag{2.56}$$

We assume that the duration of the analysis filter $h(n)$ is sufficiently short (i.e. no greater than 20ms [17]) so that the local pitch period $P(n)$ and the impulse response of the vocal tract appear nearly fixed over the analysis interval. The local approximations

$$P(m) \approx P(n) \tag{2.57}$$

$$c_k(m) \approx c_k(n) \tag{2.58}$$

$$\phi(m) \approx \phi(n) + \Omega(n)(m-n) \tag{2.59}$$

can be used to simplify (2.56) to give [19]

$$\begin{aligned} X(n,\omega) &\approx \sum_{m=-\infty}^{+\infty} \sum_{k=0}^{P(n)-1} h(n-m)c_k(n) \\ &\qquad \times \exp\left[jk\Big(\phi(n) + \Omega(n)(m-n)\Big)\right] e^{-j\omega m} \\ &= \sum_{k=0}^{P(n)-1} c_k(n) \left\{ \sum_{m=-\infty}^{+\infty} h(n-m)\exp\left[-j\Big(\omega - k\Omega(n)\Big)m\right] \right\} \\ &\qquad \times \exp\left[jk\Big(\phi(n) - \Omega(n)n\Big)\right] \\ &= \sum_{k=0}^{P(n)-1} c_k(n) H\Big(k\Omega(n) - \omega\Big) \exp\left[j\Big(k\phi(n) - \omega n\Big)\right]. \end{aligned} \tag{2.60}$$

If the bandwidth of $H(\omega)$ is less than *half* the local pitch $\Omega(n)$, then $X(n,\omega)$ may be viewed as the superposition of $P(n)$ non-overlapping weighted images of $H(\omega)$ shifted at regular frequency intervals. A bandpass representation of voiced speech follows from (2.60), with

$$X(n,\omega) = H\big(k\Omega(n) - \omega\big) \exp\Big[j\big(k\phi(n) - \omega n\big)\Big] \tag{2.61}$$

over the frequency band defined by

$$|\omega - k\Omega(n)| < \omega_h, \tag{2.62}$$

and $X(n,\omega) = 0$ elsewhere. The constant ω_h denotes the cutoff frequency of $H(\omega)$.

If the speech parameters are to appear nearly fixed over the analysis interval, the bandwidth of $H(\omega)$ should be wide enough to pass any of the harmonic amplitudes with negligible distortion.

We now investigate the magnitude and phase structure of $X(n,\omega)$ in order to estimate the underlying speech parameters. The STFT magnitude and *unwrapped* phase sequences are defined as

$$M(n,\omega) = |X(n,\omega)| \tag{2.63}$$

$$\theta(n,\omega) = \arg X(n,\omega) + 2\pi I(n,\omega), \tag{2.64}$$

where $I(n,\omega)$ is an integer which guarantees the uniqueness of $\theta(n,\omega)$. The arg[·] operator is defined as

$$\arg X(n,\omega) = \arctan\left(\frac{\text{Imag}\{X(n,\omega)\}}{\text{Real}\{X(n,\omega)\}}\right). \tag{2.65}$$

and generates a phase value in the $-\pi$ to π range.

According to (2.61), the magnitude of the vocal tract filter within a frequency band can be obtained directly from $M(n,\omega)$, albeit with a gain factor, i.e.

$$M(n,\omega) = \begin{cases} \left|c_k(n)H\left(k\Omega(n) - \omega\right)\right| & \text{for } \left|\omega - k\Omega(n)\right| < \omega_h \\ 0 & \text{otherwise.} \end{cases} \tag{2.66}$$

Unfortunately, $\theta(n,\omega)$ combines the phase contribution of the vocal tract filter $t(n,m)$, the analysis filter $h(n)$ and the excitation source $v(n)$. A procedure for estimating the individual contributions of these quantities from $\theta(n,\omega)$ appears important, considering the nature of the rate-change modifications specified earlier for our speech model. For this purpose, we define two distinct unwrapped phase components, $\alpha(n,\omega)$ and $\nu(n,\omega)$, such that

$$\theta(n,\omega) = \alpha(n,\omega) + \nu(n,\omega) \tag{2.67}$$

$$X(n,\omega) = M(n,\omega)\exp\left[j\left(\alpha(n,\omega) + \nu(n,\omega)\right)\right], \tag{2.68}$$

where

$$\alpha(n,\omega) = \arg c_k(n) + \arg H\left(k\Omega(n) - \omega\right) \tag{2.69}$$

$$\nu(n,\omega) = k\phi(n) - \omega n \tag{2.70}$$

over the frequency band $|\omega - k\Omega(n)| < \omega_h$.

The term $\alpha(n,\omega)$ contributes a slowly time-varying phase because the vocal tract filter and the fundamental frequency $\Omega(n)$ are nearly fixed for the duration of the analysis interval. The quantity $\alpha(n,\omega)$ will be referred to as the *phase modulation* component.

The term $\nu(n,\omega)$, by comparison, can vary more quickly in n because the corresponding instantaneous frequency can be as high as ω_h, which is bounded by half the pitch of the voiced speech signal. The quantity $\nu(n,\omega)$ will be referred to as the *frequency modulation* (FM) component. Given its dependence on the instantaneous phase of the excitation source, $\nu(n,\omega)$ may be expressed as in (2.24),

$$\nu(n,\omega) = \begin{cases} \sum_{r=1}^{n} \Omega(r,\omega) & \text{for } n > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2.71}$$

over the frequency band $|\omega - k\Omega(n)| < \omega_h$. The quantity $\Omega(n,\omega)$ is the STFT instantaneous frequency defined as [19]

$$\Omega(n,\omega) = k\Omega(n) - \omega. \tag{2.72}$$

We note that $\Omega(n,\omega)$ is also a slowly time-varying sequence due to its dependence on $\Omega(n)$, and as such satisfies the local phase representation

$$\nu(n+\ell,\omega) \approx \nu(n,\omega) + \Omega(n,\omega)\ell \qquad \text{for } |\ell| < M_v/2. \tag{2.73}$$

Setting $|\ell| = 1$ gives

$$\Omega(n,\omega) \approx \nu(n,\omega) - \nu(n-1,\omega) \tag{2.74}$$

and

$$\Omega(n,\omega) \approx \nu(n+1,\omega) - \nu(n,\omega). \tag{2.75}$$

Since the variations in $\alpha(n,\omega)$ are negligible compared to those in $\nu(n,\omega)$, the instantaneous frequency of the STFT can be approximated by

$$\begin{aligned}\Omega(n,\omega) &\approx \theta(n,\omega) - \theta(n-1,\omega) \\ &\approx \nabla_n^b \theta(n,\omega)\end{aligned} \tag{2.76}$$

or

$$\begin{aligned}\Omega(n,\omega) &\approx \theta(n+1,\omega) - \theta(n,\omega) \\ &\approx \nabla_n^f \theta(n,\omega),\end{aligned} \tag{2.77}$$

where $\nabla_n^b[\cdot]$ and $\nabla_n^f[\cdot]$ are the first backward and forward difference operators respectively. A reasonable estimate for $\Omega(n,\omega)$ is the average of the first backward and forward differences of the unwrapped STFT phase $\theta(n,\omega)$. An estimate for the FM component can therefore be evaluated from the sum [2]

$$\hat{\nu}(n,\omega) = \begin{cases} \sum_{r=1}^{n} \frac{1}{2}(\nabla_r^b + \nabla_r^f)\theta(r,\omega) & \text{for } n > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2.78}$$

over the frequency band $|\omega - k\Omega(n)| < \omega_h$. We now establish the criterion for choosing the unwrapping integer $I(n,\omega)$.

Since the STFT instantaneous frequency $\Omega(n,\omega)$ is bounded by the cutoff frequency of the analysis filter, $\theta(n,\omega)$ should in principle satisfy

$$|\nabla_n^b \theta(n,\omega)| < \omega_h \tag{2.79}$$

or, equivalently,

$$\left|\arg X(n,\omega) + 2\pi I(n,\omega) - \theta(n-1,\omega)\right| < \omega_h. \tag{2.80}$$

Because the phase difference in (2.80) is modified in steps of 2π, the minimum value range for $|\nabla_n^b \theta(n,\omega)|$ is 0 to π. Since proper resolution of the fine harmonic structure of voiced speech spectra requires that ω_h be much less than π, there may be no value of $I(n,\omega)$ satisfying (2.80). Therefore, the most stringent unwrapping criterion that we may postulate while guaranteeing a solution for $I(n,\omega)$ is

$$|\nabla_n^b \theta(n,\omega)| \leq \pi. \tag{2.81}$$

The unwrapped STFT phase sequence $\theta(n,\omega)$ can be calculated by adding or removing multiples of 2π to $\arg X(n,\omega)$ until (2.81) is satisfied.

While $\theta(n,\omega)$ may be smooth in that it contains jumps of no more than π, jumps in the $\pi/2$ to π range can still occur due to sign changes in the real and imaginary components of $X(n,\omega)$. A smoother unwrapped phase sequence, therefore, is precisely the $\arg X(n,\omega)$ sequence with *all* jumps of integer multiples of π removed [2],

$$\theta_\pi(n,\omega) = \arg X(n,\omega) + \pi I(n,\omega). \tag{2.82}$$

Since the minimum value range for $|\nabla_n^b \theta_\pi(n,\omega)|$ is 0 to $\pi/2$, the unwrapping criterion for $\theta_\pi(n,\omega)$ is

$$|\nabla_n^b \theta_\pi(n,\omega)| \leq \pi/2. \tag{2.83}$$

The phase sequence $\theta_\pi(n,\omega)$ can be calculated by adding or removing multiples of π to $\arg X(n,\omega)$ until (2.83) is satisfied. We stress that $\theta_\pi(n,\omega)$ does *not* accurately represent the STFT phase because integer multiples of π are not necessarily invisible. Yet this phase sequence is better suited for FM component estimation because its unwrapping criterion is more stringent than the one specified in (2.81). Substituting $\theta(n,\omega)$ by $\theta_\pi(n,\omega)$ in (2.78) should improve the FM component estimate.

To summarize, we have shown that the parameters of the harmonic representation of voiced speech can be approximated by STFT-based estimates under certain assumptions regarding the design of the analysis filter:

1. The length of $h(n)$ should be sufficiently short so that major speech parameters such as pitch and vocal tract frequency response appear nearly fixed for the duration of the analysis interval.

2. The bandwidth of $H(\omega)$ should be sufficiently wide to pass each of the time-varying harmonic amplitudes with negligible distortion, yet narrow enough to pass at most one such component. The cutoff frequency ω_h should be less than half the source pitch.

2.5.3 Unvoiced Speech

The STFT of an unvoiced speech signal is a stochastic process. Using the interpretation of the STFT as the output of a demodulator followed by a lowpass filter,

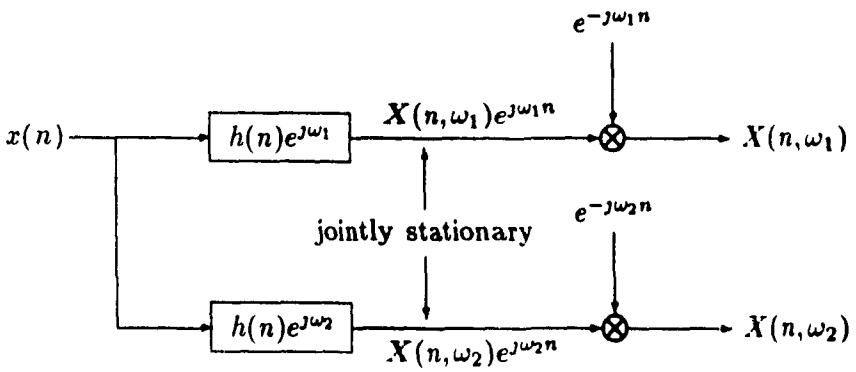


Figure 2.5: Short-time Fourier transform viewed as the output of a bandpass analysis filter followed by a demodulator.

this section treats the STFT $X(n,\omega)$ as a set of random sequences indexed by ω and sketches a second order statistical representation of the STFT.

We begin by assuming the signal $x(n)$ is stationary. It can be shown that the corresponding STFT is also stationary for a given ω. However, the sequences $X(n,\omega_1)$ and $X(n,\omega_2)$ are in general not jointly stationary for $\omega_1 \neq \omega_2$ [19], meaning that the joint statistics between the two sequences vary over time. In order to simplify the development, it would be desirable to study jointly stationary STFT sequences. To this end, we consider an alternate interpretation of the STFT [2, 25],

$$\begin{aligned} X(n,\omega) &= \sum_{m=-\infty}^{+\infty} h(n-m)x(m)e^{-j\omega m} \\ &= \left\{ \sum_{m=-\infty}^{+\infty} h(n-m)x(m)e^{j\omega(n-m)} \right\} e^{-j\omega n} \\ &= \left\{ x(n) *_n h(n)e^{j\omega n} \right\} e^{-j\omega n}. \end{aligned} \tag{2.84}$$

The STFT can be viewed as the demodulated output of a bandpass filter $h(n)e^{j\omega n}$. Since the responses of two linear time-invariant (LTI) systems to the same stationary input are jointly stationary, then the sequences $X(n,\omega_1)e^{j\omega_1 n}$ and $X(n,\omega_2)e^{j\omega_2 n}$ are also jointly stationary (Figure 2.5).

The cross-correlation sequence for these modified STFT sequences is

$$R_Y(\ell) = E\left[Y(n+\ell,\omega_1)Y^*(n,\omega_2)\right], \tag{2.85}$$

where

$$Y(n,\omega_i) = X(n,\omega_i)e^{j\omega_i n}. \tag{2.86}$$

If we substitute $X(n,\omega_i)$ by the STFT formula in the above expression and evaluate (2.85) while relaxing the stationarity assumption on $x(n)$ such that $x(n)$ becomes quasi-stationary in the sense previously defined, we obtain a time-varying cross-correlation sequence in terms of the original STFT sequences, $X(n,\omega_1)$ and $X(n,\omega_2)$ [19]. The time-varying *auto*-correlation sequence for $X(n,\omega)$ is

$$R_X(n,\ell,\omega) \approx \frac{1}{2\pi}\int_{-\pi}^{+\pi} S_x(n,\omega+\varphi)|H(\varphi)|^2 e^{j\varphi\ell}\, d\varphi, \tag{2.87}$$

where $S_x(n,\omega+\varphi)$ is the time-varying power spectrum of the unvoiced speech signal $x(n)$. The function $H(\omega)$ is the Fourier transform of the analysis filter $h(n)$. The variable ω is treated as a parameter and φ is the frequency variable corresponding to ℓ. From (2.87) we deduce that the time-varying power spectrum for $X(n,\omega)$ is a lowpass version of the original power spectrum in the vicinity of ω at time n, namely

$$S_X(n,\varphi,\omega) = S_x(n,\omega+\varphi)|H(\varphi)|^2. \tag{2.88}$$

Contrary to our earlier treatment of rate-changed unvoiced speech, the rate-changed version of (2.87) will not be obtained simply through linear time-scaling as in (2.48), because the evaluation of the modified STFT $X^\beta(n,\omega)$ involves a non-linear transformation. However, the analytical approach described above will later serve in determining the time-varying power spectrum of a signal synthesized from $X^\beta(n,\omega)$.

If the bandwidth of $H(\omega)$ is kept as narrow as in the analysis of voiced speech, $X(n,\omega)$ is a narrowband lowpass sequence. The STFT therefore varies sufficiently slowly that the estimators used to calculate $\alpha(n,\omega)$ and $\nu(n,\omega)$ in the voiced speech case are believed to be adequate for the analysis of unvoiced speech [19].

2.6 Synthesis of Rate-Changed Speech

The results of the preceding sections provide the necessary framework for synthesizing rate-changed speech from appropriately modified STFT-based estimates of the original speech parameters. A general synthesis equation for rate-changed voiced *and* unvoiced speech will be postulated, thereby avoiding the practical problem of distinguishing each speech class.

2.6.1 Voiced Speech

The rate-changed harmonic representation obtained for voiced speech, as given by (2.45), suggests that the corresponding STFT modification should consist of a linear time-scaling operation and a non-linear phase modification. It will be shown that the modified STFT from which the rate-changed speech signal $x^\beta(n)$ can be synthesized is given by [2]

$$X^\beta(n,\omega) = X(\beta n,\omega)\exp\left[j\left(\frac{1}{\beta}-1\right)\nu(\beta n,\omega)\right] \tag{2.89}$$

$$= M(\beta n,\omega)\exp\left[j\Big(\alpha(\beta n,\omega)+\nu(\beta n,\omega)/\beta\Big)\right]. \tag{2.90}$$

The magnitude, phase modulation and frequency modulation quantities were previously defined as

$$M(n,\omega) = \Big|c_k(n)H\Big(k\Omega(n)-\omega\Big)\Big|$$

$$\alpha(n,\omega) = \arg c_k(n) + \arg H\Big(k\Omega(n)-\omega\Big)$$

$$\nu(n,\omega) = k\phi(n) - \omega n$$

over the non-overlapping frequency bands delimited by $|\omega - k\Omega(n)| < \omega_h$.

Applying the general STFT synthesis formula (2.53) to the rate-changed version of the superposition sum given by (2.60), namely

$$X^{\beta}(n,\omega) \approx \sum_{k=0}^{P(\beta n)-1} c_k(\beta n) H\Big(k\Omega(\beta n) - \omega\Big) \exp\Big[j\Big(k\phi(\beta n)/\beta - \omega n\Big)\Big], \quad (2.91)$$

we obtain [2]

$$\begin{aligned} x^{\beta}(n) &\approx \frac{1}{2\pi}\int_{-\pi}^{+\pi} \sum_{r=-\infty}^{+\infty} f(n-r) \Bigg\{ \sum_{k=0}^{P(\beta r)-1} c_k(\beta r) \\ &\qquad \times H\Big(k\Omega(\beta r) - \omega\Big) \exp\Big[j\Big(k\phi(\beta r)/\beta - \omega r\Big)\Big] \Bigg\} e^{j\omega n}\, d\omega \\ &= \sum_{r=-\infty}^{+\infty} \sum_{k=0}^{P(\beta r)-1} f(n-r) c_k(\beta r) \exp\left[jk\phi(\beta r)/\beta\right] \\ &\qquad \times \left\{ \frac{1}{2\pi}\int_{-\pi}^{+\pi} H\Big(k\Omega(\beta r) - \omega\Big) \exp\left[j\omega(n-r)\right]\, d\omega \right\} \\ &= \sum_{r=-\infty}^{+\infty} \sum_{k=0}^{P(\beta r)-1} f(n-r) h(r-n) c_k(\beta r) \\ &\qquad \times \exp\Big[jk\Big(\phi(\beta r)/\beta + \Omega(\beta r)(n-r)\Big)\Big]. \end{aligned} \quad (2.92)$$

Exploiting the local representation for $\phi(n)$ given by (2.59) with both m and n scaled by β, (2.92) becomes

$$x^{\beta}(n) \approx \sum_{r=-\infty}^{+\infty} \sum_{k=0}^{P(\beta r)-1} f(n-r) h(r-n) c_k(\beta r) \exp\left[jk\phi(\beta n)/\beta\right]. \quad (2.93)$$

Assuming the variations in the rate-changed pitch period sequence $P(\beta n)$ can be neglected over any sample interval less than the length of $f(n)h(-n)$, i.e.

$$P(\beta r) \approx P(\beta n) \qquad \text{for } f(n-r)h(r-n) \neq 0, \quad (2.94)$$

(2.93) can be rewritten as

$$x^{\beta}(n) \approx \sum_{k=0}^{P(\beta n)-1} \Bigg(\sum_{r=-\infty}^{+\infty} f(n-r) h(r-n) c_k(\beta r) \Bigg) \exp\left[jk\phi(\beta n)/\beta\right]. \quad (2.95)$$

The summation over r represents the convolution of the rate-changed harmonic amplitude $c_k(\beta n)$ with the composite filter $f(n)h(-n)$. We recall that the speech parameters are nearly fixed for the duration of the analysis filter $h(n)$. Since the effective length of the composite filter cannot exceed that of $h(n)$, the approximations used to derive (2.95) are justified. Equivalently, the effective bandwidth of the composite filter is wide enough to pass the rate-changed harmonic amplitudes $c_k(\beta n)$ with negligible distortion. Therefore, (2.95) reduces to

$$x^{\beta}(n) \approx \sum_{k=0}^{P(\beta n)-1} c_k(\beta n) \exp\left[jk\phi(\beta n)/\beta\right], \tag{2.96}$$

which is the same expression for rate-changed voiced speech as (2.45).

2.6.2 Unvoiced Speech

It was argued in [2] that the STFT modification used to synthesize rate-changed voiced speech can also be used to approximate the desired rate-changed unvoiced speech. Based on the second order representation for the STFT outlined earlier, Portnoff showed that the time-varying power spectrum of the signal $y(n)$ synthesized from (2.90) is approximately the same as that of the ideal rate-changed signal $x^{\beta}(n)$.

The time-varying power spectrum of the rate-changed unvoiced speech estimate $y(n)$ was evaluated as [2]

$$S_y(n,\omega) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} S_x(\beta n, \omega+\varphi) G_{\beta}(\varphi)\, d\varphi, \tag{2.97}$$

where $S_x(n,\omega)$ is the time-varying power spectrum of the original unvoiced speech signal $x(n)$. The function $G_{\beta}(\omega)$ represents the composite effect of the analysis and synthesis filters as a function of the time-scaling factor β. Hence, $S_y(n,\omega)$ is a smoothed version of the desired rate-changed time-varying power spectrum

$S_x(\beta n, \omega)$, and the degree of smearing depends on β. Several simplifying assumptions were made in the course of the development:

1. The underlying speech process is Gaussian.
2. The unvoiced speech spectrum is sufficiently smooth that the spectral resolution afforded by the analysis filter $h(n)$ is adequate for calculating the second moments of the STFT random sequences.
3. The effect of the phase modulation component $\alpha(n, \omega)$ is negligible.

Portnoff claimed that in practice the amount of spectral smearing resulting from the synthesis of rate-changed unvoiced speech using (2.90) is "acceptable" [2], an observation which confirms the inherent smoothness of unvoiced speech spectra [26].

2.7 Distortion in Rate-Changed Speech

Using a structural interpretation of the STFT, we shall show that some degree of distortion in rate-changed speech is unavoidable under the proposed TSM method. It will be argued that the accumulation of phase error, for which the principal causes will be examined in order of ascending importance, gradually deteriorates the structure and perceptual quality of rate-changed signals in general.

Waveform Events and Structure

A waveform *event* is defined as a set of consistent temporal and spectral features which characterize a waveform over a finite time interval. The magnitude and phase quantities of the STFT of a signal $x(n)$, expressed as a function of both time and frequency, offer a convenient representation for these features. We assume that

the temporal boundaries of an event are marked by abrupt changes in the phase and magnitude characteristics of the underlying waveform.

The signal $x(n)$ may be segmented, according to some temporal and spectral feature similarity measure, into a string of contiguous sub-waveforms $x_i(m)$ representing individual events of length L_i. The i-th event is defined as

$$x_i(m) = \begin{cases} x\left(m + \sum_{r=0}^{i-1} L_r\right) & \text{for } 0 \le m < L_i \\ 0 & \text{otherwise.} \end{cases} \tag{2.98}$$

The original signal $x(n)$ is recovered by concatenating the $x_i(m)$ as follows,

$$x(n) = \sum_{i=0}^{\infty} x_i\left(n - \sum_{r=0}^{i-1} L_r\right). \tag{2.99}$$

The STFT of the i-th event will be written as $X_i(m,\omega)$, which is just $X(n,\omega)$ expressed as a function of L_i and m. Each event may be represented by the inverse STFT formula

$$x_i(m) = \int_{-\pi}^{+\pi} M_i(m,\omega) \exp\left[j\theta_i(m,\omega)\right] e^{j\omega m}\, d\omega \qquad 0 \le m < L_i, \tag{2.100}$$

where the quantities $M_i(m,\omega)$ and $\theta_i(m,\omega)$ represent the STFT magnitude and unwrapped phase data collected for the i-th event. The proportionality constant, $1/2\pi h(0)$, has been absorbed into $M_i(m,\omega)$ for conciseness.

As the sampling frequency f_s is increased, the difference between the last sample of $x_i(m)$ and the first sample of $x_{i+1}(m)$ decreases. Likewise, the difference between the last STFT of the i-th event and the first STFT of the $(i+1)$-th event decreases. Events which satisfy the condition

$$\lim_{f_s \to \infty} \left\{ x_{i+1}(0) - x_i(L_i - 1) \right\} = 0 \tag{2.101}$$

or

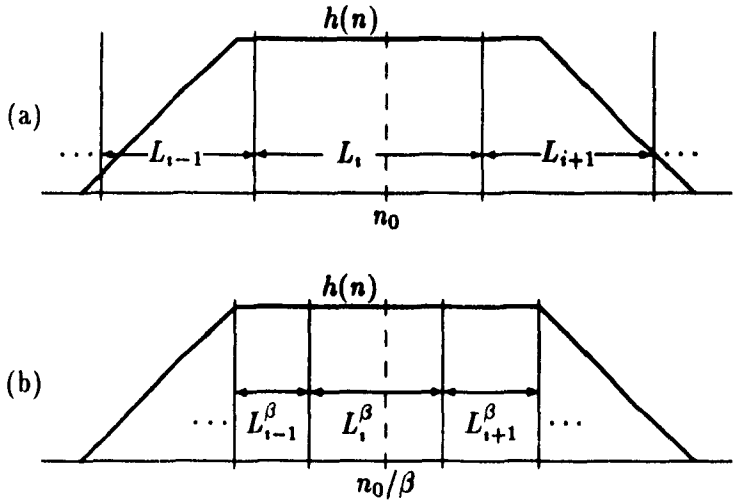


Figure 2.6: a) Waveform events falling within scope of analysis filter $h(n)$ at time n_0 on original time-scale. The i-th event is weighted the most. b) Theoretical waveform event relationship after a rate-change modification. All three events now happen to be weighted evenly. Note that $L_i^{\beta} = \lfloor L_i/\beta \rfloor$.

$$\lim_{f_s \to \infty} \left\{ X_{i+1}(0,\omega) - X_i(L_i - 1, \omega) \right\} = 0 \tag{2.102}$$

will be referred to as "phase-continuous".

The waveform *structure* of $x_i(m)$ depends on the phase and magnitude relationship of the frequency components over time. That relationship is not unique in the context of short-time Fourier analysis. Specifically, the STFT phase and magnitude values calculated for a particular event $x_i(m)$ are influenced by the characteristics of neighboring events falling within the scope of the analysis filter $h(n)$ (Figure 2.6). For every set of magnitude values calculated from a STFT frame whose scope includes the event of interest, $x_i(m)$, and irrelevant events, $x_j(m)$ where $j \neq i$, there is a corresponding set of phase values allowing $x_i(m)$ to be recovered exactly from the synthesis equation given by (2.100). If one quantity is modified without a corresponding modification in the other, the original event $x_i(m)$ cannot, in general, be recovered exactly.

For example, consider the modified event

$$y_i(m) = \int_{-\pi}^{+\pi} M_i(m,\omega) \exp\left[j\left(\theta_i(m,\omega) + \epsilon_i(m,\omega)\right)\right] e^{j\omega m}\, d\omega, \qquad (2.103)$$

where $\epsilon_i(m,\omega)$ is a phase modulation component. We may postulate that the waveform structure of $y_i(m)$ differs from that of $x_i(m)$ unless $\epsilon_i(m,\omega)$ is a constant[1]. Up to what point $\epsilon_i(m,\omega)$ may deviate from a constant such that $y_i(m)$ remains *perceptually* the same as $x_i(m)$ is a difficult question best answered by perception science. However, it is clear that waveform structure is linked in some way to the perceptual character of an event and that excessive structural deterioration of an event must lead to a corresponding deterioration in its perceived quality.

FM Component Estimation

It is impossible to determine the exact value of the FM component $\nu(n,\omega)$ from the unwrapped STFT phase $\theta(n,\omega)$ without *a priori* knowledge of the phase modulation component $\alpha(n,\omega)$. Therefore, the phase modification term of (2.89) cannot be computed exactly. However, FM component estimation errors are not likely to be severe over quasi-stationary and quasi-periodic portions of $x(n)$ because the phase modulation term $\alpha(n,\omega)$ accounts for only a small fraction of the overall STFT phase in such cases.

Local Phase Representation

In practice, $\nu(n,\omega)$ will often deviate from the local phase representation given by (2.73) within the scope of a STFT frame. Consequently, the simple phase modification specified in (2.90) merely ensures that the average slope of $\nu^{\beta}(n,\omega)$ matches that of $\nu(n,\omega)$. We must conclude that the phase modification term

[1]The special case where $\epsilon_i(m,\omega) = \pm\omega n_0$ such that n_0 corresponds to a sample offset in the time domain is of no interest.

of (2.89) itself is flawed, and that some degree of phase error in each frequency component would be unavoidable even if the FM component were known exactly.

Event Boundaries

It is not uncommon for one or more waveform event boundaries to fall within the scope of a STFT frame in the course of ordinary speech processing (Figure 2.6). Since they are marked, according to our previous definition, by abrupt changes in phase and magnitude characteristics, event boundaries violate the fragile quasi-stationarity and quasi-periodicity assumptions of the speech model. Transients can be expected to cause more acute phase errors for the following reasons:

1. Sudden variations in the phase modulation component $\alpha(n,\omega)$ may compete with those in $\nu(n,\omega)$, leading to poorer FM component estimates.

2. The local phase representation given by (2.73) may be completely invalid in the vicinity of transients, leading to poorer phase modification.

Depending on the nature of the transients, the phase disturbances may be selective, thereby increasing the tendency for the phases of adjacent frequency components to become poorly synchronized over time.

Memory

A closer examination of the phase modification term of (2.89) reveals the foremost weakness of the TSM method proposed by Portnoff. The FM component $\nu(n,\omega)$ given by (2.78) has *infinite memory* due to its recursive structure. This fact implies that each one of its phase values depends on the location of the time origin and that estimation errors accumulate indefinitely. The same applies to its rate-changed version, $\nu^{\beta}(n,\omega)$.

Impact of Phase Error

We now relate the accumulation of phase error to the waveform structure concept discussed earlier.

We see from (2.89) that Portnoff's TSM method is an example of an application where the phases of the frequency components of $x_i^\beta(m)$ are modified independently from their magnitude, i.e.

$$x_i^\beta(m) = \int_{-\pi}^{+\pi} M_i(\beta m, \omega) \exp\left[j\Big(\theta_i(\beta m, \omega) + \epsilon_i(m, \omega)\Big)\right] e^{j\omega m}\, d\omega \qquad (2.104)$$
$$0 \le m < \lfloor L_i/\beta \rfloor,$$

where

$$\epsilon_i(m, \omega) = \left(\frac{1}{\beta} - 1\right) \nu_i(\beta m, \omega). \qquad (2.105)$$

The $\lfloor \cdot \rfloor$ operator rounds its parameter to the nearest integer below. For convenience, the length of the i-th rate-changed event will be denoted as L_i^β. The structure of the original event $x_i(m)$ cannot be preserved under a rate-change modification because $\epsilon_i(m, \omega)$ is not a constant. Since the phase and magnitude features of $x_i^\beta(m)$ are consistent for the duration of the event, the resulting structural deterioration could in fact be perceptually benign.

Suppose now that a rate-changed speech signal $x^\beta(n)$ is to be constructed by concatenating its rate-changed events, $x_i^\beta(m)$. The individual rate-changed events are not necessarily phase-continuous in the sense previously defined because the average frequency of each component is preserved while the event durations L_i are altered by a factor of $1/\beta$. Therefore, in general,

$$\lim_{f_s \to \infty} \left\{ X_{i+1}^\beta(0, \omega) - X_i^\beta(L_i^\beta - 1, \omega) \right\} \neq 0. \qquad (2.106)$$

An example of rate-changed events which are not phase-continuous is shown in Figure 2.7.

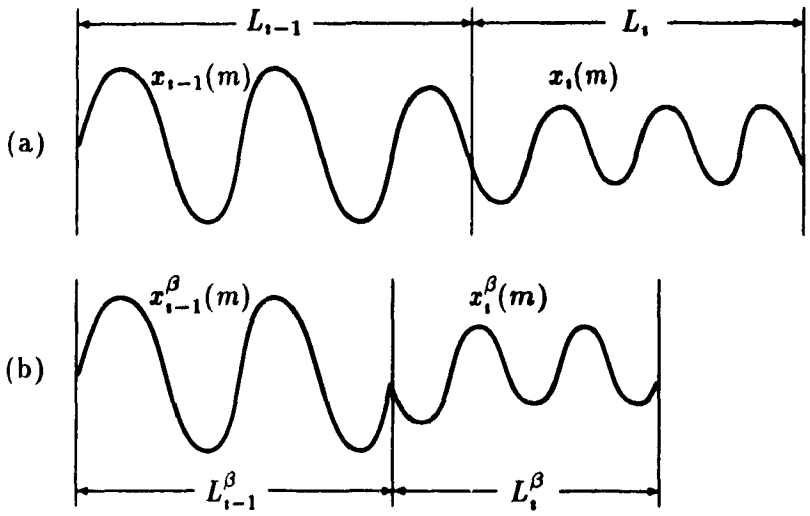


Figure 2.7: Rate-change modification by segmentation and concatenation. a) Original waveform with boundaries for two events. b) Concatenation of individually rate-changed events. Note that for each event the average frequency and initial phase offset have been preserved under the rate-change modification.

There are many ways to modify the magnitude and phase of $X^{\beta}_i(m,\omega)$ to ensure phase continuity. However, we know that the rate-changed STFT magnitude $M_i(\beta m,\omega)$ is already continuous in some sense across event boundaries. Moreover, we are interested only in the structural impact of modifying the phase of an STFT independently of its magnitude.

To this end, we define a phase-modulated version of $x^{\beta}_i(m)$,

$$y^{\beta}_i(m) = \int_{-\pi}^{+\pi} M_i(\beta m,\omega) \times \exp\left[j\Big(\theta_i(\beta m,\omega) + \epsilon_i(m,\omega) + \mu_i(\omega)\Big)\right] e^{j\omega m}\, d\omega. \tag{2.107}$$

The rate-changed signal $x^{\beta}(n)$ is obtained by concatenating the $y^{\beta}_i(m)$ as in (2.99),

$$x^{\beta}(n) = \sum_{i=0}^{\infty} y^{\beta}_i\left(n - \sum_{r=0}^{i-1} L^{\beta}_r\right). \tag{2.108}$$

The phase modulation term $\mu_i(\omega)$ forces the starting phase of a frequency component ω in the i-th rate-changed event to coincide (in the modulo-2π sense) with

the ending phase of the corresponding frequency component in the $(i+1)$-th rate-changed event in the limit of large f_s, i.e.

$$\lim_{f_s\to\infty}\left\{Y_{i+1}^{\beta}(0,\omega) - Y_i^{\beta}(L_i^{\beta}-1,\omega)\right\} = 0. \tag{2.109}$$

The STFT $X_{i+1}^{\beta}(0,\omega)$ is "out of phase" with $X_i^{\beta}(L_i^{\beta}-1,\omega)$ by $\epsilon_i(L_i^{\beta}-1,\omega)$ in the limit of large f_s. In order to make the $y_i^{\beta}(m)$ phase-continuous, the phase offsets induced by previous rate-changed events accumulate in $\mu_i(\omega)$ as this example demonstrates,

$$\begin{aligned}
\mu_0(\omega) &= 0 \\
\mu_1(\omega) &= \lim_{f_s\to\infty}\epsilon_0(L_0^{\beta}-1,\omega) \\
\mu_2(\omega) &= \mu_1(\omega) + \lim_{f_s\to\infty}\epsilon_1(L_1^{\beta}-1,\omega) \\
&\vdots \\
\mu_i(\omega) &= \sum_{r=0}^{i-1}\lim_{f_s\to\infty}\epsilon_r(L_r^{\beta}-1,\omega).
\end{aligned} \tag{2.110}$$

The phase disturbances which accumulate in $\mu_i(\omega)$ eventually alter the original structure and perceptual quality of the $x_i^{\beta}(m)$ in a significant way. In general, therefore, some of degree of structural distortion in rate-changed signals is unavoidable under Portnoff's TSM method.

Chapter 3

Design of a Time-Scale Modification System

The concepts developed in Chapter 2 may now applied to the design of a TSM system for speech. The major design difficulty stems from the STFT itself: in order for the TSM system to be realizable on a digital processor, the frequency-axis of the STFT must be represented by a finite number of samples. Moreover, STFT computations are costly both in terms of real-time and memory. Consequently, allowances for decimating the STFT in time should also be made. Temporal and spectral sampling are design issues which will compromise our former theoretical expectations to some degree, while introducing an extra level of complexity beyond the original requirement of linear time-scaling.

In this chapter we formulate our earlier STFT-based speech parameter estimates in terms of the downsampled *discrete* STFT and establish the theoretical limits on the temporal and spectral sampling intervals. The main components of a time-frequency based TSM system are discussed in sequence. The design includes a waveform structure compensation stage for preventing excessive deterioration of rate-changed speech signals over time. A variety of design options are presented and compared on the basis of complexity and audio quality. The final design is that of a practical TSM system capable of producing high-quality rate-changed

speech under varying conditions at a reasonable computational cost.

3.1 Analysis

DSTFT Definition

The short-time Fourier analyzer computes samples of the STFT of the original speech according to the formula

$$X(sR, k\Omega_N) = \sum_{m=-\infty}^{+\infty} h(sR - m)x(m)\exp[-j\Omega_N m], \tag{3.1}$$

where $\Omega_N = 2\pi/N$ and R represent the spectral and temporal sampling intervals respectively. The sequence $h(n)$ is an analysis window of length M and will be regarded as a *finite impulse response* (FIR) filter of bandwidth ω_h. The downsampled discrete STFT, henceforth abbreviated as DSTFT, can be efficiently computed via the *Fast Fourier Transform* (FFT) class of algorithms, especially if N is a power of 2.

Temporal and Spectral Sampling

The parameters R and N must be carefully selected so that the result of processing the modified DSTFT instead of the actual STFT is approximately the same. The optimal choice is unclear due to the non-linear nature of the TSM processing. Nevertheless, in the absence of any parameter modification ($\beta = 1$), the original STFT $X(n, \omega)$ should be recoverable from $X(sR, k\Omega_N)$. The sequence $X(n, \omega)$ should be sampled "often enough" in the spectral and temporal directions to prevent aliasing in n and its corresponding frequency dimension. Since the bandwidth of $X(n, \omega)$ is bounded by ω_h, the admissible range for the temporal sampling interval R is given by (2.7), i.e.

$$R \leq \pi/\omega_h. \tag{3.2}$$

The admissible range for the number of frequency samples N is [28]

$$N \geq M. \tag{3.3}$$

Thus the temporal and spectral Nyquist intervals are π/ω_h and $2\pi/M$, respectively.

Analysis Filter

While it would be desirable to keep the filter length M as short as possible to reduce storage requirements and computational load, the bandwidth of $H(\omega)$, which is inversely proportional to M, must achieve several aims.

In particular, $H(\omega)$ should be narrow enough to allow proper resolution of voiced speech spectra and adequate estimation of unvoiced speech spectra. Yet the bandwidth of the analysis filter should be broad enough to pass the temporal features of speech with minimum distortion, such that the underlying speech parameters appear nearly fixed for the duration of $h(n)$. The short-time Fourier analysis of voiced speech was conducted earlier under the highly idealized assumption that $H(\omega)$ is zero outside its passband, a design goal which is impossible to achieve with a practical filter. Thus we seek a lowpass FIR filter of length M to satisfy the conflicting requirements stated above, one having a relatively flat and narrow passband as well as sharp attenuation characteristics in the stopband region.

Among the most commonly used analysis windows listed in Table 3.1, the Hamming window offers a good compromise and other notable advantages such as linear phase, ease of computation and non-zero samples over its entire duration. The peak amplitude of the sidelobes in the stopband region is about -41dB. The bandwidth of a Hamming window (one half the width of its main lobe) is

Window Type	*Width of Mainlobe*	*Sidelobe Amplitude* (dB)
Rectangular	$4\pi/(M+1)$	-13
Bartlett	$8\pi/M$	-25
Hanning	$8\pi/M$	-31
Hamming	$8\pi/M$	-41
Blackman	$12\pi/M$	-57

Table 3.1: Specifications for common analysis windows. (After Oppenheim and Schafer [27].)

approximately $\pm 4\pi/M$. Substituting ω_h by this value in (3.2) yields the theoretical upper bound on the temporal sampling interval R in terms of M

$$R \leq M/4. \tag{3.4}$$

According to (2.60), the short-time Fourier spectrum of voiced speech consists of weighted images of $H(\omega)$ centered about harmonics of the fundamental frequency $\Omega(n)$. It was pointed out that the representation holds provided the bandwidth of the analysis filter is less than one half the source pitch. Expressing this restriction in terms of the cutoff frequency of the Hamming window we obtain a lower bound for M,

$$\frac{4\pi}{M} < \frac{\Omega_{min}}{2} \tag{3.5}$$

or

$$M > 4P_{max}, \tag{3.6}$$

where P_{max} is the pitch period corresponding to the lowest expected source pitch Ω_{min}. Male voices can be as low as 60Hz, implying that the analysis interval should be greater than 66ms.

However, speech parameters cannot be considered to be "nearly fixed" for periods greater than 20ms; some stop bursts may be as short as 5-10ms [29].

Since our speech model and all ensuing results rely on the key assumption of "local stationarity", we choose to limit the length of the analysis window to 20ms at the expense of poorer frequency resolution for low-pitched speech signals—this is a commonly accepted compromise in many speech analysis applications [29]. Expressed as a function of the sampling frequency f_s, the upper bound on M is therefore

$$M \leq 20 \times 10^{-3} f_s. \tag{3.7}$$

The pitch range for obtaining a "clean" STFT bandpass representation of voiced speech signals subject to the constraint specified by (3.7) follows directly from (3.6),

$$f_v > 200\text{Hz}. \tag{3.8}$$

For a fixed analysis window duration (in ms), increasing the sampling frequency f_s neither improves the ability of the Fourier analyzer to discriminate the harmonics of a voiced speech signal nor reduces the frequency spacing (in Hz) between the DSTFT harmonics. In short, the frequency resolution of the Fourier analyzer is independent of f_s if the duration of the analysis window is fixed.

3.2 Synthesis

3.2.1 DSTFT Synthesis

The rate-changed speech signal $x^\beta(n)$ may be synthesized from the modified DSTFT $X^\beta(sR, k\Omega_N)$ according to the general formula

$$x^\beta(n) = \frac{1}{N} \sum_{k=0}^{N-1} \sum_{s=-\infty}^{+\infty} f(n - sR) X^\beta(sR, k\Omega_N) \exp\left[j\Omega_N kn\right], \tag{3.9}$$

Portnoff used this formula for generating rate-changed speech [2]. The sequence $f(n)$ is a FIR filter which performs $1 : R$ bandlimited interpolation on both the real and imaginary parts of $X^{\beta}(sR, k\Omega_N)$. The proportionality constant $1/h(0)$ has been absorbed in $f(n)$. The constants Ω_N and R are the spectral and temporal sampling intervals previously defined.

Portnoff claimed that since "the non-linear rate-change modification does not preserve the structure of the STFT of an arbitrary signal, the synthesized signal, in general, depends on the design of $f(n)$" [2]. If it was meant that the structure of a rate-changed signal can be preserved solely through proper interpolation of DSTFT coefficients, we must differ with that statement on the basis of the arguments presented in Section 2.7. Structural deterioration occurs even in the absence of temporal and spectral sampling.

Portnoff recommended the algorithm proposed by Oetken *et al.* [30] for designing the optimal $f(n)$. The sequence $X^{\beta}(n, k\Omega_N)$ is in principle recoverable from its samples provided the real and imaginary parts of $X^{\beta}(sR, k\Omega_N)$ are bandlimited. This issue will be discussed further in Section 3.4. However, since $f(n)$ is non-ideal, interpolation errors are unavoidable.

Small errors in the complex data stream could in fact amount to large phase errors. The non-linear arg[·] operator defined by (2.65) is sensitive, especially in the asymptotic regions of the arctan(·) function. The fact that the real and imaginary parts of $X^{\beta}(sR, k\Omega_N)$ are processed as *independent* bandlimited sequences in (3.9) raises some question. The perceptual impact of the phase disturbances caused by DSTFT interpolation errors is not well understood.

In contrast, DSTFT magnitude errors are proportional to the magnitude of the interpolation errors in the complex data stream.

The synthesis equation given by (3.9) is, in some respect, inefficient. We see from (2.89) that the STFT rate-change modification requires the computation of

one polar parameter, namely the phase modification term. A polar to rectangular coordinate conversion is needed for each DSTFT harmonic in order to generate the real and imaginary parts of $X^{\beta}(sR, k\Omega_N)$ When normally implemented, the conversion requires four multiplications and two additions per DSTFT coefficient, as this example demonstrates,

$$\begin{aligned} z &= (a + jb)e^{j\theta} \\ &= (a\cos\theta - b\sin\theta) + j(a\sin\theta + b\cos\theta). \end{aligned} \tag{3.10}$$

3.2.2 Polar Synthesis

We propose to interpolate the magnitude and phase of $X^{\beta}(sR, k\Omega_N)$ instead of its real and imaginary parts. The magnitude and phase sequences derived from an STFT are not necessarily bandlimited due to the non-linear transformations involved. Consequently, they cannot be recovered exactly from their samples. The rate-changed signal synthesized from the polar representation of $X^{\beta}(sR, k\Omega_N)$ will not, in general, be an exact representation of the inverse DSTFT corresponding to $X^{\beta}(sR, k\Omega_N)$ for $R \neq 1$.

However, the computational requirements for interpolating DSTFT polar parameters are believed to be modest for achieving high-quality synthesis. The magnitude sequence $M(sR, k\Omega_N)$ is slowly time-varying, due to its dependence, as (2.66) indicates, on parameters which are assumed to be nearly fixed over the duration of the analysis interval. Furthermore, the unwrapped phase sequence $\theta(sR, k\Omega_N)$ is also smooth in some sense because the phase unwrapping process removes most of the discontinuities from the $\arg X(sR, k\Omega_N)$ stream. McAulay and Quatieri used linear and cubic polynomials for magnitude and phase interpolation respectively in their SSM system [9, 10].

The proposed synthesis equation for the rate-changed speech *estimate* is

$$\hat{x}^{\beta}(n) = \frac{1}{Nh(0)} \sum_{k=0}^{N-1} \hat{M}^{\beta}(n, k\Omega_N) \exp[j\hat{\theta}^{\beta}(n, k\Omega_N)] e^{j\Omega_N kn}, \tag{3.11}$$

where

$$\hat{M}^{\beta}(n, k\Omega_N) = f_M\left[n, R; \hat{M}^{\beta}(sR, k\Omega_N)\right] \tag{3.12}$$

$$\hat{\theta}^{\beta}(n, k\Omega_N) = f_{\theta}\left[n, R; \hat{\theta}^{\beta}(sR, k\Omega_N)\right] \tag{3.13}$$

for $-\infty < s < \infty$. The $f_M[\cdot]$ and $f_{\theta}[\cdot]$ operators are $1 : R$ interpolating functions. Since not all such functions can be expressed in terms of a linear convolution, we have adopted a more general notation. The polar quantities $\hat{M}^{\beta}(sR, k\Omega_N)$ and $\hat{\theta}^{\beta}(sR, k\Omega_N)$ are written as estimates because they also rely on some form of non-ideal interpolation.

Since the rate-changed speech signal is always real, there are only $N/2$ distinct harmonics and (3.11) reduces to

$$\hat{x}^{\beta}(n) = \frac{2}{Nh(0)} \sum_{k=0}^{N/2-1} \hat{M}^{\beta}(n, k\Omega_N) \cos[\hat{\theta}^{\beta}(n, k\Omega_N) + \Omega_N kn]. \tag{3.14}$$

The result is similar to the SSM synthesis equation used in [9, 10]. The main difference is that the sinusoidal components of (3.14) employ fixed rather than time-varying base frequencies. The polar synthesis equation requires only $N/2$ multiplications[1] per output sample whereas the implementation given by (3.9) requires a minimum of $6N$ (real) multiplications[2].

Informal listening tests indicate that, in the absence of any parameter modification ($\beta = 1$), the speech signal synthesized from (3.14) is virtually indistinguishable from the original when the DSTFT polar parameters are linearly interpolated. This

[1] The apparent multiplication involved in calculating the linear phase term $\Omega_N kn$ can easily be avoided.

[2] $4 \times N/2$ real multiplications for the coordinate conversion and $4N \log_r N$ real multiplications for the actual synthesis, assuming a radix-r FFT algorithm is used.

would suggest that little improvement in subjective performance is to be gained by using Portnoff's bandlimited interpolation approach.

3.3 Phase Unwrapping and Estimation

3.3.1 Phase Unwrapping

We now establish the unwrapping criterion for $\theta(sR, k\Omega_N)$ and the range of R over which it holds. The results are then extended to $\theta_\pi(sR, k\Omega_N)$. We recall that $\theta(n,\omega)$ and $\theta_\pi(n,\omega)$ differ only by an integer multiple of π.

The sampled transform implementation of the unwrapped STFT phase is

$$\theta(sR, k\Omega_N) = \arg X(sR, k\Omega_N) + 2\pi I(sR, k\Omega_N). \tag{3.15}$$

The most stringent phase unwrapping criterion that we may postulate for the unwrapped DSTFT phase is

$$|\nabla_s^b \theta(sR, k\Omega_N)| \leq \pi. \tag{3.16}$$

The expression follows from the same arguments we used to derive (2.81), the unwrapping criterion for $\theta(n,\omega)$. Thus $\theta(sR, k\Omega_N)$ is obtained by adding or removing integer multiples of 2π until (3.16) is satisfied.

Since the instantaneous frequency of the STFT, $\Omega(n,\omega)$, is slowly time-varying, then so must be $\nabla_n^b \theta(n,\omega)$ by virtue of (2.76). Consequently,

$$\nabla_s^b \theta(sR, k\Omega_N) \approx R\Big\{\nabla_n^b \theta(n, k\Omega_N)|_{n=sR}\Big\}. \tag{3.17}$$

Multiplying both sides of (2.79) by R yields the STFT instantaneous frequency bound for the sampled transform implementation

$$|\nabla_s^b \theta(sR, k\Omega_N)| < R\omega_h. \tag{3.18}$$

In order to satisfy (3.16), R must be chosen such that

$$R\omega_h \leq \pi. \tag{3.19}$$

The temporal sampling frequency $2\pi/R$ should therefore be no less than $2\omega_h$, the Nyquist rate for sampling $X(n,\omega)$ in the temporal direction. If the analysis filter $h(n)$ is a Hamming window of length M, the cutoff frequency ω_h is approximately $4\pi/M$. Hence, $R \leq M/4$, which is the same bound we derived earlier for sampling the STFT.

Repeating the above analysis for $\theta_\pi(sR, k\Omega_N)$, which is given by

$$\theta_\pi(sR, k\Omega_N) = \arg X(sR, k\Omega_N) + \pi I(sR, k\Omega_N), \tag{3.20}$$

we find that

$$|\nabla_s^b \theta_\pi(sR, k\Omega_N)| < R\omega_h \leq \pi/2. \tag{3.21}$$

In this case, the temporal sampling frequency $2\pi/R$ should be no less than *twice* the Nyquist rate for sampling $X(n,\omega)$ in the temporal direction. The unwrapping criterion for $\theta_\pi(sR, k\Omega_N)$ holds only for $R \leq M/8$, assuming the analysis filter is a Hamming window of length M.

3.3.2 FM Component Estimation

The FM component estimator $\hat{\nu}(n,\omega)$ for the sampled transform implementation follows directly from (2.78),

$$\hat{\nu}(sR, k\Omega_N) = \begin{cases} \sum_{r=1}^{s} \frac{1}{2}(\nabla_r^b + \nabla_r^f)\psi(rR, k\Omega_N) & \text{for } s > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.22}$$

over the frequency band

$$k|\Omega_N - \Omega(sR)| < \omega_h. \tag{3.23}$$

The unwrapped phase sequence $\psi(sR, k\Omega_N)$ is either $\theta(sR, k\Omega_N)$ or $\theta_\pi(sR, k\Omega_N)$, depending on the desired accuracy. The quantity $\theta_\pi(n,\omega)$ was found to be more suitable for FM component estimation.

If the FM component estimator is computed from $\theta(sR, k\Omega_N)$, the variations in the phase modulation component $\alpha(sR, k\Omega_N)$ are implicitly neglected, i.e.

$$\begin{aligned}\nabla_s^b \hat{\alpha}(sR, k\Omega_N) &= \nabla_s^b\big[\theta(sR, k\Omega_N) - \hat{\nu}(sR, k\Omega_N)\big] \\ &\approx 0.\end{aligned} \tag{3.24}$$

This approach relies more on the fundamental assumption that the speech parameters are nearly fixed over the duration of the analysis interval. The resulting FM component estimate will tend to be coarse, leading, in principle, to more rapid structural deterioration of the rate-changed speech. If the distortion can be tolerated, it would be more economical to compute $\hat{\nu}(sR, k\Omega_N)$ from $\theta(sR, k\Omega_N)$ since the polar synthesis method already uses the unwrapped DSTFT phase.

3.4 Parameter Modification

The STFT rate-change modification, according to (2.89), consists of two basic procedures: linear time-scaling and non-linear phase modification. Each procedure will be treated individually. Then a novel approach specific to the polar TSM synthesis method will be presented.

3.4.1 Linear Time-Scaling

An integral part of the TSM method proposed by Portnoff consists of linearly time-scaling the magnitude and phase of the original STFT.

Since the $X(sR, k\Omega_N)$ sequences (indexed by k) are bandlimited, one way to achieve this objective is to apply the linear time-scaling formula given by (2.9) to the real and imaginary parts of the DSTFT as follows,

$$X(\beta sR, k\Omega_N) = \sum_{r=-\infty}^{+\infty} f(sD - rI)X(rR, k\Omega_N), \tag{3.25}$$

where $\beta = D/I$. The interpolating FIR filter $f(n)$ may again be designed by Oetken's technique [30]. The linearly time-scaled sequences $\theta(\beta sR, k\Omega_N)$ and $\theta_\pi(\beta sR, k\Omega_N)$ are obtained by unwrapping the phase argument of $X(\beta sR, k\Omega_N)$ as described in Section 3.3.

An alternate method which produces perceptually acceptable results consists of linearly time-scaling the polar representation of $X(sR, k\Omega_N)$, i.e.

$$\hat{M}(\beta sR, k\Omega_N) = f_M\big[s, \beta; M(rR, k\Omega_N)\big] \tag{3.26}$$

$$\hat{\theta}(\beta sR, k\Omega_N) = f_\theta\big[s, \beta; \theta(rR, k\Omega_N)\big] \tag{3.27}$$

for $-\infty < r < \infty$. The $f_M[\cdot]$ and $f_\theta[\cdot]$ operators are the interpolating functions previously defined. The same $f_\theta[\cdot]$ interpolator can be used to calculate $\hat{\theta}_\pi(\beta sR, k\Omega_N)$.

Simplicity is the key advantage of the polar approach. As we indicated earlier, $f_M[\cdot]$ need only perform linear interpolation because $M(sR, k\Omega_N)$ is slowly time-varying. The $f_\theta[\cdot]$ interpolator may also be simple in design, such as a first or third order polynomial [10], since the unwrapped DSTFT phase is believed to be sufficiently smooth.

However, the polar approach is less efficient than (3.25) due to the computations involved in obtaining the magnitude sequence $M(sR, k\Omega_N)$[3]. The computational

[3] 2 multiplications and 1 square root operation per magnitude value

requirements of the unwrapped DSTFT phase $\theta(sR, k\Omega_N)$ are irrelevant because both synthesis methods described in Section 3.2 ultimately require some form of phase unwrapping.

3.4.2 Phase Modification

The phase modification step consists of dividing the the linearly time-scaled FM component $\nu(\beta n, \omega)$ by β. This is achieved, as (2.89) indicates, by multiplying the linearly time-scaled STFT $X(\beta n, \omega)$ by a non-linear phase modification term. The sampled transform implementation for this term is

$$\exp\left[j\left(\frac{1}{\beta}-1\right)\hat{\nu}(\beta sR, k\Omega_N)\right]. \tag{3.28}$$

The implementation for the FM component estimator $\hat{\nu}(sR, k\Omega_N)$ was discussed in Section 3.3. Calculating the FM component estimator from $\hat{\theta}(\beta sR, k\Omega_N)$ or $\hat{\theta}_\pi(\beta sR, k\Omega_N)$ implicitly generates $\hat{\nu}(\beta sR, k\Omega_N)$.

If the rate-changed signal is synthesized from the DSTFT synthesis equation given by (3.9), the order in which the linear time-scaling and the phase modification procedures are performed becomes important to prevent frequency aliasing of the original $X(sR, k\Omega_N)$ sequence [2]. The bandwidth of the modified sequence

$$\begin{aligned} Y(n,\omega) &= X(n,\omega)\exp\left[j\left(\frac{1}{\beta}-1\right)\nu(n,\omega)\right] \\ &= M(n,\omega)\exp\left[j\Big(\alpha(n,\omega)+\nu(n,\omega)/\beta\Big)\right] \end{aligned} \tag{3.29}$$

is about $1/\beta$ times the bandwidth of $X(n,\omega)$. In contrast, the bandwidth of $X(\beta n, \omega)$ is β times that of the original. When both scaling methods are combined to produce the rate-changed STFT as in (2.89), the effective bandwidth is approximately the same as that of $X(n,\omega)$. Consequently, for time-scale compression ($\beta > 1$), the phase modification should be implemented ***before*** the linear

time-scaling operation. Conversely, for time-scale expansion ($0 < \beta < 1$), the phase modification should be implemented *after* the linear time-scaling operation.

The implementation order is of no concern to the polar TSM synthesis method because the phase and magnitude sequences are not treated as bandlimited sequences.

3.4.3 A Novel Incremental Approach

The relative smoothness the STFT magnitude and unwrapped phase sequences can be exploited to simplify the parameter modification procedure for the polar synthesis method.

We assume that the FM component estimator $\hat{\nu}(n,\omega)$ given by (3.22) is calculated from the unwrapped STFT phase $\theta(n,\omega)$ and that

$$\nabla_n^b \theta(n,\omega) \approx \nabla_n^f \theta(n,\omega). \tag{3.30}$$

The FM component estimator reduces to

$$\hat{\nu}(n,\omega) = \begin{cases} \theta(n,\omega) - \theta(0,\omega) & \text{for } n > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.31}$$

Consequently, all variations in the phase modulation component $\alpha(n,\omega)$ are ignored, i.e.

$$\begin{aligned} \hat{\alpha}(n,\omega) &= \theta(n,\omega) - \hat{\nu}(n,\omega) \\ &= \theta(0,\omega). \end{aligned} \tag{3.32}$$

The estimate for the rate-changed unwrapped STFT phase follows from (2.90)

$$\begin{aligned} \hat{\theta}^{\beta}(n,\omega) &= \hat{\alpha}(\beta n,\omega) + \hat{\nu}(\beta n,\omega)/\beta \\ &= \theta(0,\omega) + \big(\theta(\beta n,\omega) - \theta(0,\omega)\big)/\beta. \end{aligned} \tag{3.33}$$

The initial STFT phase offsets are preserved under the rate-change modification as desired.

It is implicitly assumed by (3.31) that $\theta(n,\omega)$ satisfies the local phase representation given by (2.73) for the FM component. Scaling both n and ℓ by β and setting $\ell = -1$ in (2.73) yields an alternate representation for the rate-changed STFT instantaneous frequency,

$$\Omega(\beta n,\omega) \approx \frac{1}{\beta}\nabla_n^b \theta(\beta n,\omega). \tag{3.34}$$

Since $\Omega(n,\omega)$ is slowly time-varying, we introduce the following approximation,

$$\Omega(\beta n,\omega) \approx \Omega(\lfloor \beta n \rfloor,\omega). \tag{3.35}$$

By substituting n with $\lfloor \beta n \rfloor$ in (2.76) and comparing the result with (3.34), we obtain the local approximation

$$\frac{1}{\beta}\nabla_n^b \theta(\beta n,\omega) \approx \nabla_{\lfloor \beta n \rfloor}^b \theta(\lfloor \beta n \rfloor,\omega). \tag{3.36}$$

The slope of the linearly time-scaled unwrapped STFT phase $\theta(\beta n,\omega)$ is therefore approximately β times the slope of $\theta(n,\omega)$ in the vicinity of the time-scaled instant βn.

Keeping in line with our usual recursive definition of unwrapped phase quantities, we propose that the phase estimate for $X^\beta(n,\omega)$ be

$$\hat{\theta}^\beta(n,\omega) = \begin{cases} \theta(0,\omega) + \sum_{r=1}^{n} \nabla_{\lfloor \beta r \rfloor}^b \theta(\lfloor \beta r \rfloor,\omega) & \text{for } n > 0 \\ \theta(0,\omega) & \text{for } n = 0. \end{cases} \tag{3.37}$$

Likewise, we propose that the magnitude estimate for $X^\beta(n,\omega)$ be

$$\hat{M}^\beta(n,\omega) = M(\lfloor \beta n \rfloor,\omega). \tag{3.38}$$

Index on New Time-Scale	Index on Original Time-Scale	
n	$\lfloor 1.25n \rfloor$	$\lfloor 0.8n \rfloor$
0	0	0
1	1	0
2	2	1
3	3	2
4	5	3
5	6	4
6	7	4
7	8	5
8	10	6
9	11	7
...	...	...

Table 3.2: Relationship between the sample indices on the original and new time-scales for the incremental approach. For time-scale compression ($\beta = 1.25$), one sample interval is *deleted* every four samples. For time-scale expansion ($\beta = 0.8$), one sample interval is *repeated* every four samples.

Equations (3.37) and (3.38) define a TSM system where no explicit linear time-scaling nor multiplications by $1/\beta$ are required. Time-scale compression ($\beta > 1$), is effectively achieved by periodically *deleting* sample intervals from the original signal $x(n)$, whereas time-scale expansion ($0 < \beta < 1$) consists of periodically *repeating* sample intervals. The examples of Table 3.2 illustrate this point.

The rate-changed unwrapped STFT phase sequence $\hat{\theta}^{\beta}(n,\omega)$ retains essentially the same smoothness properties as the original due to its incremental structure. However, $\hat{M}^{\beta}(n,\omega)$ is in general discontinuous. This is of no great concern because the original STFT magnitude sequence is slowly time-varying.

We note that *variable* TSM is easily implemented by letting the time-scale factor β vary as a function of time in (3.37) and (3.38).

3.5 Waveform Structure Compensation

From the discussion in Section 2.7, it is clear that the performance of a Portnoff-like TSM system could be improved if the impact of infinite memory were reduced. This would help restrict the accumulation of phase error which deteriorates the structure of rate-changed signals. Two new approaches to the problem will be examined.

3.5.1 Waveform Interpolation

We assume that the speech signal $x(n)$ can be segmented into a string of events $x_i(m)$ defined as in Section 2.7. The goal is to construct the rate-changed signal $x^\beta(n)$ by concatenating its individually rate-changed events $x_i^\beta(m)$. It was shown that these are not, in general, phase continuous at their boundaries. The phase quantity $\mu_i(\omega)$ of (2.107) was introduced to correct the problem, but was found to deteriorate the structure of the rate-changed events over time.

Waveform interpolation is suggested as an alternative for removing the discontinuity created as a result of concatenating rate-changed events. While it does not eliminate the accumulation of phase error over a single event, this approach does prevent the phase error from accumulating indefinitely.

The basic idea is illustrated in Figure 3.1. Two rate-changed events are concatenated by overlapping and interpolating the last κ samples of $x_{i-1}^\beta(m)$ with the first κ samples of $x_i^\beta(m)$. The effective length of each rate-changed event is therefore shortened by κ samples. Sample loss can be avoided simply by defining original events which overlap each other by $\lfloor \beta\kappa \rfloor + 1$ samples in the first place, i.e.

$$y_i(m) = \begin{cases} x\left(m + \sum_{r=0}^{i-1} L_r\right) & \text{for } 0 \le m < L_i + \lfloor \beta\kappa \rfloor + 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.39}$$

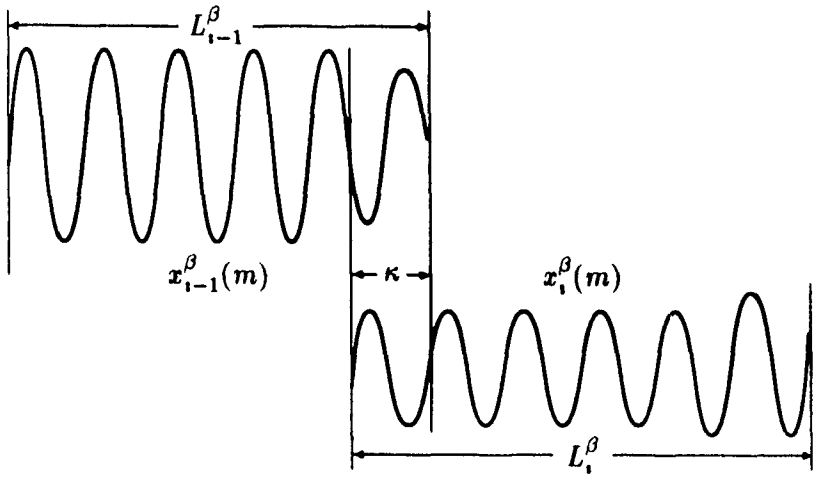


Figure 3.1: Waveform interpolation for two rate-changed events, $x_{i-1}^{\beta}(m)$ and $x_i^{\beta}(m)$. The waveform segments within the interpolation interval κ are severely out of phase.

The effective length of each rate-changed event $y_i^{\beta}(m)$ after interpolation is $\lfloor L_i/\beta \rfloor$ as desired. We recall that the length of rate-changed events is denoted as L_i^{β}. The rate-changed signal $x^{\beta}(n)$ is then constructed by concatenating the $y_i^{\beta}(m)$ as in (2.108).

The interpolation interval κ should be kept as short as possible to preserve the character of the individual rate-changed events. Ideally,

$$\kappa \ll L_i^{\beta}. \tag{3.40}$$

However, as a rule, the larger the phase mismatch between two rate-changed events, the longer κ must be to smooth the discontinuity. One way to alleviate the discontinuity is to *align* the rate-changed events prior to interpolation.

The starting index of the i-th rate-changed event is offset by ℓ_i samples such that the next κ samples provide the "best match" for the last κ samples of the $(i-1)$-th rate-changed event. This approach is shown in Figure 3.2. The modified

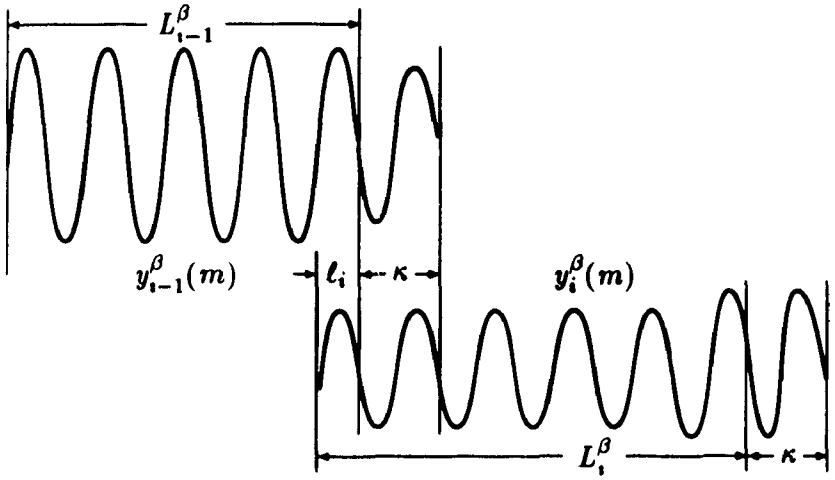


Figure 3.2: Waveform interpolation with time-alignment for two "extended" rate-changed events, $y^{\beta}_{i-1}(m)$ and $y^{\beta}_{i}(m)$. The waveform segments within the interpolation interval κ (in the middle of the diagram) are in phase as a result of discarding the first ℓ_i samples from $y^{\beta}_{i}(m)$.

version of $y^{\beta}_{i}(m)$ is defined as

$$z^{\beta}_{i}(m) = y^{\beta}_{i}(m + \ell_i) \qquad \text{for } 0 \le m < L^{\beta}_{i} + \kappa - \ell_i, \tag{3.41}$$

where ℓ_i maximizes the short-time correlation function

$$\max_{\ell} \left\{ \sum_{r=0}^{\kappa-1} y^{\beta}_{i-1}(r + L^{\beta}_{i-1})\, y^{\beta}_{i}(r + \ell_i) \right\}. \tag{3.42}$$

Though the first ℓ_i samples are discarded from $y^{\beta}_{i}(m)$, the interpolation interval κ can be kept shorter than it would be if the events were not time-aligned.

In order to preserve the character of the rate-changed events and to limit the amount of distortion at event boundaries, the segmentation rules should be designed to satisfy the condition

$$L_i \gg \beta(\kappa + \ell_i). \tag{3.43}$$

3.5.2 Phase Modulation

The phase modulation approach also assumes that $x^\beta(n)$ is constructed from individually rate-changed events. In essence, the method consists of adjusting the phase values of the STFT frequency components such as to prevent the phase error from accumulating beyond a single event. Contrary to the waveform interpolation approach, the phase modulation method eliminates waveform discontinuities *prior* to synthesis and does not discard any samples.

We recall that the initial phase offsets of the frequency components for each event are preserved under rate-change modifications, i.e.

$$\arg X_i(0,\omega) = \arg X_i^\beta(0,\omega). \tag{3.44}$$

Consequently, the quantities $X_i(0,\omega)$ provide an exact description of the phase relationship among the frequency components at specific points along $x(n)$.

We propose to restrict the accumulation of phase error by forcing the i-th rate-changed event to be phase-continuous with the $(i+1)$-th *original* event. The amount of phase correction required at the i-th event boundary is given by

$$\Theta_i(\omega) = \arg\Big[\lim_{f_s\to-\infty}\Big\{X_{i+1}(0,\omega) - X_i^\beta(L_i^\beta - 1,\omega)\Big\}\Big], \tag{3.45}$$

The amount of phase correction does not exceed π in either the positive or negative direction. Since $X_{i+1}(0,\omega)$ and $X_i^\beta(L_i^\beta,\omega)$ correspond roughly to the same sample instant along $x^\beta(n)$, the quantity $\Theta_i(\omega)$ can in practice be approximated by

$$\Theta_i(\omega) \approx d_\pi\Big[\arg X_{i+1}(0,\omega) - \arg X_i^\beta(L_i^\beta,\omega)\Big]. \tag{3.46}$$

The $d_\pi[\cdot]$ operator adds or removes multiples of 2π to its argument until the result lies in the $-\pi$ to π range. The STFT value $X_i^\beta(L_i^\beta,\omega)$ serves only in the computation of $\Theta_i(\omega)$, not in the actual synthesis of $x^\beta(n)$.

We define a phase-modulated version of $x_i^\beta(m)$ as

$$y_i^\beta(m) = \int_{-\pi}^{+\pi} X_i^\beta(m,\omega) \exp[j\mu_i(m,\omega)]\, e^{j\omega m}\, d\omega \qquad 0 \le m < L_i^\beta. \tag{3.47}$$

The proportionality constant $1/2\pi h(0)$ has been absorbed in $X_i^\beta(m,\omega)$ for brevity. The phase modulation term $\mu_i(m,\omega)$ distributes, according to some rule, the phase correction amount $\Theta_i(\omega)$ over L_i^β samples to ensure the $y_i^\beta(m)$ become phase-continuous. The rate-changed signal $x^\beta(n)$ is then constructed by concatenating the $y_i^\beta(m)$ as in (2.108).

In order to limit the distortion in each synthesized sample, it is suggested that $\Theta_i(\omega)$ be uniformly distributed over the i-th event, i.e.

$$\mu_i(m,\omega) = \frac{1}{L_i^\beta} d_\pi \left[\arg X_{i+1}(0,\omega) - \arg X_i^\beta(L_i^\beta,\omega)\right] m. \tag{3.48}$$

The average frequency of $X_i^\beta(m,\omega)$ is therefore shifted by a fixed quantity which does not exceed $\pm\pi/L_i^\beta$. The term $\mu_i(m,\omega)$ is a function of present and *future* phase data. Unlike the phase modulation term $\mu_i(\omega)$ given by (2.110), the effect of $\mu_i(m,\omega)$ does not extend beyond the i-th rate-changed event.

Since the amount of phase modulation is not constant across the frequency spectrum, we are compromising the spectral structure of the rate-changed event $x_i^\beta(m)$ to ensure the TSM system has finite memory while enforcing phase continuity at the event boundaries. The phase modulation amount could be made arbitrarily small by increasing L_i, but only at the cost of increased temporal structure distortion over the i-th rate-changed event.

Preliminary experiments using (3.47) and (3.48) indicate that there is no overlap between the perceptual tolerance regions of temporal and spectral distortion: one form of distortion is always apparent in the rate-changed output. Spectral distortion is perceived as occasional beating among the vocal tract harmonics in voiced speech and as occasional smearing in unvoiced speech. Temporal distortion

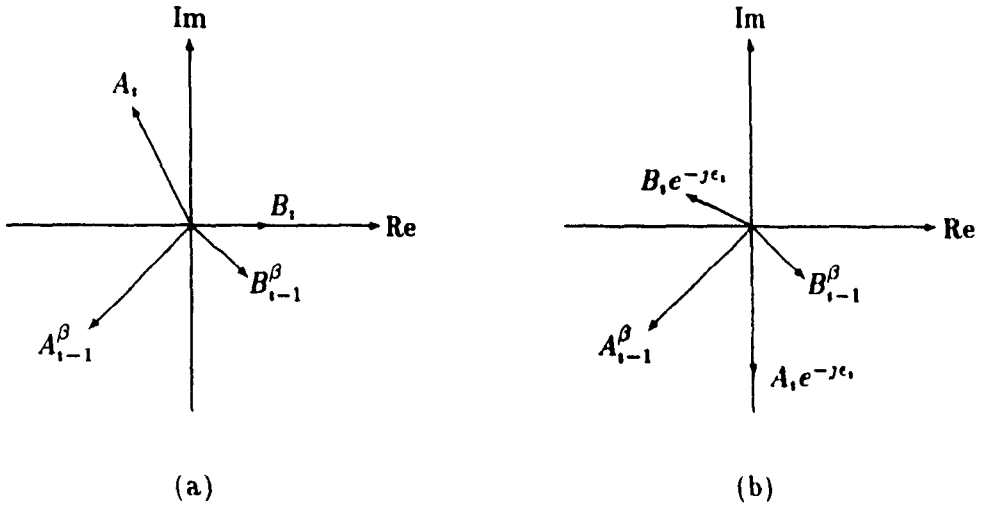


Figure 3.3: a) Phase and magnitude relationship between two frequency components, A and B, both on the original and new time-scales. The phase difference between A_i and A^{β}_{i-1} represents the amount of phase correction required for A in the $(i-1)$-th rate-changed event. Ditto for B_i and B^{β}_{i-1}. b) Application of weighting criteria to reduce the amount of phase correction for the perceptually most important frequency component, A. Vectors on the original time-scale are rotated by the constant $-\epsilon_i$.

modifies the perceptual character of some phoneme onsets and sustained sounds in more subtle ways. The method must be developed further to achieve a better structural compromise.

Perceptual Weighting

It is reasonable to suppose that the amount of perceived spectral distortion is proportional to the amount of phase modulation applied to the perceptually important frequency components of $x^{\beta}_i(m)$.

Hence, we propose to offset the phase of $X^{\beta}_{i+1}(m,\omega)$ by a constant ϵ_{i+1} to reduce the phase correction amount $\Theta_i(\omega)$ for the perceptually important frequency components. The constant ϵ_{i+1} can be interpreted as the complex plane rotation factor for the $(i+1)$-th original event. An example is shown in Figure 3.3. Ac-

cording to our earlier postulate in Section 2.7, the constant $\epsilon_{\iota+1}$ does not alter the waveform structure of $x^{\beta}_{\iota+1}(m)$.

The revised definition for $y^{\beta}_{\iota}(m)$ is

$$y^{\beta}_{\iota}(m) = \int_{-\pi}^{+\pi} X^{\beta}_{\iota}(m,\omega) \exp\left[j\left(\mu_{\iota}(m,\omega) + \epsilon_{\iota}\right)\right] e^{j\omega m}\, d\omega \qquad (3.49)$$
$$0 \le m < L^{\beta}_{\iota},$$

where

$$\epsilon_{\iota} = \int_{-\pi}^{+\pi} W_{\iota-1}(\omega) d_{2\pi}\left[\arg X_{\iota}(0,\omega) - \arg X^{\beta}_{\iota-1}(L^{\beta}_{\iota-1},\omega)\right] d\omega \qquad (3.50)$$

$$\mu_{\iota}(m,\omega) = \frac{1}{L^{\beta}_{\iota}} d_{\pi}\left[\arg X_{\iota+1}(0,\omega) - \epsilon_{\iota+1} - \arg X^{\beta}_{\iota}(L^{\beta}_{\iota},\omega)\right] m. \qquad (3.51)$$

The $d_{2\pi}[\cdot]$ operator adds or removes multiples of 2π to its argument until the result lies in the 0 to 2π range. The $W_{\iota}(\omega)$ quantity represents the perceptual weight of the frequency component ω. The constant ϵ_{ι} is the weighted average of the phase differences at the $(\iota-1)$-th event boundary. Contrary to the phase modulation component $\mu_{\iota}(m,\omega)$, the constant ϵ_{ι} depends on *past* STFT data.

Equation (3.49) may be intuitively interpreted as follows: $\mu_{\iota}(m,\omega)$ ensures that $y^{\beta}_{\iota}(m)$ is phase-continuous with $y^{\beta}_{\iota+1}(m)$ while ϵ_{ι} reduces the perceptual impact of phase modulation in $y^{\beta}_{\iota-1}(m)$. The system memory is still limited to a single event. We now define the perceptual weighting function.

Ignoring the masking and spectral energy spreading effects of the often-used critical band model of the human ear [12], a reasonable definition for the perceptual weighting function is

$$W_{\iota}(\omega) = \lambda_{\iota} |X_{\iota}(0,\omega)|^2, \qquad (3.52)$$

where the normalization constant λ_{ι} is given by

$$\lambda_{\iota} = \left[\int_{-\pi}^{+\pi} |X_{\iota}(0,\omega)|^2\, d\omega\right]^{-1}. \qquad (3.53)$$

However, (3.52) fails to take into account that the ear is most sensitive to spectral distortion in the low-frequency range of the auditory spectrum due to the narrow bandwidth of the critical bands in that region [31]. For example, an average frequency deviation of $\pm\pi/L_i^\beta$ applied to frequency components below 1kHz rather than above is much more objectionable to the ear. As it stands now, (3.52) tends to favor frequency components near the first formant region (1kHz), where most speech energy is concentrated. It is suggested that the weighting function be amended to reflect the position of a frequency component in the auditory spectrum.

The revised definition is

$$W_i(\omega) = \lambda_i |X_i(0,\omega)|^2 C(\omega), \tag{3.54}$$

where

$$\lambda_i = \left[\int_{-\pi}^{+\pi} |X_i(0,\omega)|^2 C(\omega)\, d\omega\right]^{-1}. \tag{3.55}$$

Given the logarithmic response of the ear, it is reasonable to presume an exponential form for $C(\omega)$ which favors lower frequencies,

$$C(\omega) = \begin{cases} \omega^{-p} & \text{for } \omega > 0 \\ 0 & \text{for } \omega = 0, \end{cases} \tag{3.56}$$

where $p \geq 1$. Increasing the weighting exponent p tends to concentrate the total weight near $\omega = 0$. However, p should not be so large as to cancel the effect of the spectral energy weights.

A heuristic upper bound for p can be defined as the value of p which attenuates the spectral energy weights by an amount equal to the average dynamic range D (in dB) of the signal over the frequency range having the highest spectral energy concentration, i.e.

$$10\log\Big(C(\omega_{min})/C(\omega_{max})\Big) < D, \tag{3.57}$$

or

$$p < \frac{D}{10\log(\omega_{min}/\omega_{max})}. \tag{3.58}$$

Letting $\omega = k\Omega_N$, we obtain the corresponding expression for the sampled transform implementation

$$p < \frac{D}{10\log(k_{min}/k_{max})}. \tag{3.59}$$

For voiced speech, the frequency range of interest is 0 to 1kHz. Assuming the average dynamic range of the harmonics in the first formant region is about 50dB, we obtain (with $k_{min} = 1$) an upper bound for p expressed as a function of the number of frequency samples N and the sampling frequency f_s,

$$p < 5/\log(N \times 1000\text{Hz}/f_s). \tag{3.60}$$

3.6 Overall Design

We now indicate the preferred design options for implementing a complete TSM system. We begin by summarizing the basic steps involved in computing the rate-changed STFT $X^{\beta}(n,\omega)$ for the proposed TSM synthesis methods. In order to simplify the notation, it is assumed that no spectral sampling takes place.

Portnoff's method requires, in the case of time-scale expansion,

- Bandlimited $D : I$ interpolation of the real and imaginary parts of $X(sR,\omega)$ to obtain $X(\beta sR,\omega)$.
- Computation and unwrapping of the phase sequence corresponding to the linearly time-scaled DSTFT. The resulting sequence is either $\theta(\beta sR,\omega)$ or $\theta_\pi(\beta sR,\omega)$, depending on the desired FM component accuracy.

- Estimation of the FM component from the unwrapped phase sequence obtained in the preceding step.
- Multiplication of the FM component estimate $\hat{\nu}(\beta sR,\omega)$ by $1/\beta$.
- A polar-to-rectangular conversion to obtain the real and imaginary parts of the rate-changed DSTFT estimate $\hat{X}^{\beta}(sR,\omega)$.
- Bandlimited $1:R$ interpolation of the real and imaginary parts of $\hat{X}^{\beta}(sR,\omega)$ to obtain $X^{\beta}(n,\omega)$.

The computational steps are the same for time-scale compression, where only the implementation order varies.

The polar synthesis method, in combination with the incremental parameter modification scheme, requires

- Computation of the magnitude and unwrapped phase sequences corresponding to $X(sR,\omega)$.
- Linear $1:R$ interpolation of the magnitude sequence $M(sR,\omega)$ to obtain an estimate for $M(n,\omega)$.
- Linear or cubic $1:R$ interpolation of the phase sequence $\theta(sR,\omega)$ to obtain an estimate for $\theta(n,\omega)$.
- Estimation of the rate-changed magnitude sequence $M^{\beta}(n,\omega)$.
- Estimation of the rate-changed unwrapped phase sequence $\theta^{\beta}(n,\omega)$.

We recall that the rate-changed quantities of the last two steps are obtained through a "sample interval" insertion and deletion process without the use of multiplications and explicit interpolation procedures.

The polar incremental method is crude compared to Portnoff's original approach. However, the primary source of distortion in rate-changed speech is the structural deterioration caused by the non-linear STFT phase modification. There is therefore no point in striving to obtain an exact representation of a rate-changed STFT which is non-ideal in the first place. Furthermore, preliminary experiments using polar synthesis with $\beta = 1$ suggest that exact signal representations may be perceptually redundant.

For these reasons, we elect to base our TSM system design on the polar incremental approach. More rapid structural deterioration of the rate-changed speech is, however, anticipated. Waveform structure compensation is therefore an important issue.

The waveform interpolation method, which consists of overlapping and interpolating consecutive rate-changed waveform events, is straightforward. In practice, however, it is often difficult to hide the disturbances caused by interpolated waveform events which are severely out of phase. While the time-alignment procedure discussed earlier alleviates the problem, the distortion becomes less obvious if the waveform event boundaries correspond to phoneme boundaries or are located in low energy regions. Voicing detection was found to be adequate for segmenting noise-free speech. Unfortunately, the resulting TSM system was not robust. More sophisticated segmentation algorithms would be required for processing noisy speech.

The phase modulation method, on the other hand, compromises the spectral structure of the rate-changed signal to eliminate temporal structure deterioration at event boundaries. This is achieved by uniformly distributing the phase difference between consecutive rate-changed events over time. Consequently, the phase disturbances are never concentrated in one particular region of the signal. The weighting scheme suggested earlier can significantly reduce the amount of phase

correction required for perceptually important frequency components. In general, the perceptual impact of phase modulation can be further reduced by increasing the length of the original waveform events. However, the length of the original event also determines the storage requirements of the TSM system. The phase data for an entire rate-changed waveform event must be buffered before the phase modulation computations can take place. Assuming the number of distinct DSTFT harmonics is $N/2$, the total amount of phase data storage required for processing the i-th event is

$$S_\theta = \frac{N}{2} L_i^\beta b_\theta, \tag{3.61}$$

where L_i^β denotes the length of the i-th rate-changed waveform event. The constant b_θ represents the amount of storage per phase coefficient. The buffering requirement poses a serious practical limitation: *variable* length rate-changed waveform events cannot be synthesized at a *constant* rate, assuming no side information is available. We are therefore obliged to choose a constant waveform event length L and to disregard the issue of segmentation altogether.

Despite the latter compromise, phase modulation is the preferred method, as it is believed to be more robust and economical.

The block diagram for the final TSM system is shown in Figure 3.4. The mathematical representations for each component will be reviewed in sequence.

Analysis

Since the input sample index range of most practical N-point FFT algorithms is limited to $[0, N-1]$, we must modify the original STFT definition, which assumes that the analysis filter $h(n)$ slides past the input signal $x(n)$ and is centered about index n. In the following definition, the position of $h(n)$ instead appears fixed in

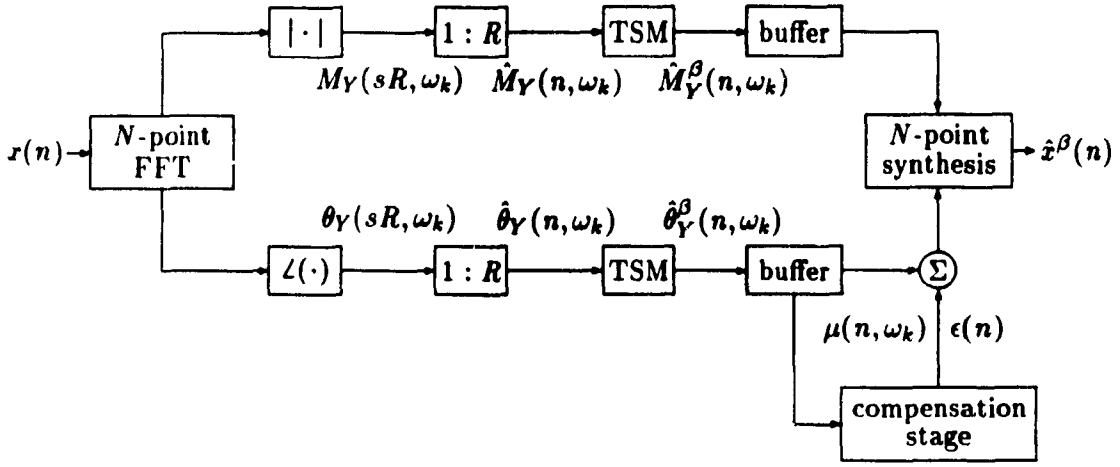


Figure 3.4: Block diagram for the proposed polar incremental TSM system. The subscript Y reflects the change in the original STFT definition. Note that $\omega_k = \Omega_N k$.

time and centered within the summation interval, while $x(n)$ slides past $h(n)$,

$$\begin{aligned} Y(sR, k\Omega_N) &= \sum_{m=0}^{N-1} h(m - N/2)x(sR + m - N/2)e^{-jk\Omega_N m} \\ &= \exp\left[jk\Omega_N(sR - N/2)\right] X(sR, k\Omega_N). \end{aligned} \tag{3.62}$$

Expressed in terms of the polar quantities previously defined, the magnitude and unwrapped phase sequences corresponding to $Y(sR, k\Omega_N)$ are, respectively,

$$\begin{aligned} M_Y(sR, k\Omega_N) &= |Y(sR, k\Omega_N)| \\ &= M(sR, k\Omega_N) \end{aligned} \tag{3.63}$$

and

$$\begin{aligned} \theta_Y(sR, k\Omega_N) &= \arg Y(sR, k\Omega_N) + 2\pi I_Y(sR, k\Omega_N) \\ &= k(sR - N/2)\Omega_N + \theta(sR, k\Omega_N). \end{aligned} \tag{3.64}$$

For a given k, the phase quantities $\theta_Y(sR, k\Omega_N)$ and $\theta(sR, k\Omega_N)$ differ only by a constant and a linear term. Substituting (3.64) in (3.16) yields the new phase unwrapping criterion

$$\left|\nabla_s^b \theta_Y(sR, k\Omega_N) - k\Omega_N R\right| < \pi. \tag{3.65}$$

The Hamming window defined by

$$h(n) = 0.54 - 0.46\cos\left[\frac{2\pi(n + M/2)}{M}\right] \qquad -M/2 \le n < M/2 \tag{3.66}$$

is chosen as the analysis filter. The length M is fixed at $\lfloor 20 \times 10^{-3} f_s \rfloor$ to afford adequate frequency resolution while the speech parameters appear reasonably constant over the analysis interval. The temporal sampling interval R is bounded by $M/4$ to prevent phase unwrapping errors and frequency-aliasing of $Y(sR, k\Omega_N)$. The number of frequency samples N is chosen such that $N > M$ to prevent time-aliasing of the windowed speech segments. A power of 2 is selected for N so that the efficiency of radix-2 FFT algorithms may be exploited to compute $Y(sR, k\Omega_N)$.

Parameter Interpolation

The estimate for the magnitude sequence $M_Y(n, k\Omega_N)$ is obtained by linearly interpolating its samples as in [10],

$$\hat{M}_Y(n, k\Omega_N) = M_Y(sR, k\Omega_N) + \xi(n, s)\nabla_s^f M_Y(sR, k\Omega_N), \tag{3.67}$$

where

$$s = \lfloor n/R \rfloor \tag{3.68}$$

$$\xi(n, s) = \frac{n - sR}{R}. \tag{3.69}$$

While $\hat{\theta}_Y(n, k\Omega_N)$ could be computed in the same manner, cubic interpolation would provide a smoother estimate. The idea was first proposed by Almeida and Silva [32] who used cubic interpolation in their harmonic wave synthesizer, and was later adopted by McAulay and Quatieri [10].

The approach assumes that $\hat{\theta}_Y(n, k\Omega_N)$ is obtained over the range $sR \leq n \leq (s+1)R$ by sampling the continuous cubic polynomial

$$\theta_k(t) = a + bt + ct^2 + dt^3 \qquad 0 \leq t \leq R \tag{3.70}$$

at unity time intervals, with the boundary conditions

$$\theta_k(0) = \theta_Y(sR, k\Omega_N) \tag{3.71}$$

$$\theta_k(R) = \theta_Y\big((s+1)R, k\Omega_N\big). \tag{3.72}$$

The average frequency of the k-th harmonic is approximately $k\Omega_N$ because the linear phase component $e^{jk\Omega_N n}$ tends to dominate $\theta_Y(n, k\Omega_N)$. Using the fact that the instantaneous frequency is the derivative of the phase, we may postulate further constraints on $\theta_k(t)$, i.e.

$$\dot{\theta}_k(0) = \dot{\theta}_k(R) \approx k\Omega_N. \tag{3.73}$$

Under the constraints defined by equations (3.71), (3.72) and (3.73), the coefficient solution set corresponding to (3.70) is

$$a = \theta_Y(sR, k\Omega_N) \tag{3.74}$$

$$b = k\Omega_N \tag{3.75}$$

$$c = \frac{3}{R^2}\left[\nabla_s^f \theta_Y(sR, k\Omega_N) - k\Omega_N\right] \tag{3.76}$$

$$d = -\frac{2}{R^3}\left[\nabla_s^f \theta_Y(sR, k\Omega_N) - k\Omega_N\right]. \tag{3.77}$$

We note that

$$\lim_{R\to\infty} c = \lim_{R\to\infty} d = 0, \tag{3.78}$$

indicating that cubic interpolation should be no more effective than linear interpolation for large temporal sampling intervals. Cubic interpolation is more useful in a SSM system, where in general $\dot{\theta}_k(0) \neq \dot{\theta}_k(R)$ because the base frequencies of the sinusoidal components vary over time. Both the linear and cubic methods will be implemented for comparison.

Parameter Modification

Rate-change modifications for the estimated phase and magnitude sequences are implemented using the incremental parameter modification equations given by (3.37) and (3.38). The corresponding estimators for the final implementation are

$$\hat{\theta}_Y^{\beta}(n, k\Omega_N) = \begin{cases} \theta_Y(0, k\Omega_N) + \sum_{r=1}^{n} \nabla_{\lfloor \beta r \rfloor}^{b} \hat{\theta}_Y(\lfloor \beta r \rfloor, k\Omega_N) & \text{for } n > 0 \\ \theta_Y(0, k\Omega_N) & \text{for } n = 0. \end{cases} \tag{3.79}$$

$$\hat{M}_Y^{\beta}(n, k\Omega_N) = \hat{M}_Y(\lfloor \beta n \rfloor, k\Omega_N). \tag{3.80}$$

Phase Modulation

The phase modulation equations for the final implementation follow from (3.50), (3.51), (3.54), (3.55), (3.56) and (3.60). For convenience, the constant L^{β} will be used to denote the rate-changed event length $\lfloor L/\beta \rfloor$. It will also be assumed that L is a multiple of the temporal sampling interval R. The equations are

$$\epsilon(n) = \sum_{k=0}^{N/2-1} W_{i-1}(k) d_{2\pi}\left[\arg Y(iL, k\Omega_N) - \hat{\theta}_Y^{\beta}(iL^{\beta}, k\Omega_N)\right] \tag{3.81}$$

$$\mu(n, k\Omega_N) = \frac{1}{L^{\beta}} d_{\pi}\left[\arg Y\big((i+1)L, k\Omega_N\big) - \epsilon(n + L^{\beta}) - \hat{\theta}_Y^{\beta}\big((i+1)L^{\beta}, k\Omega_N\big)\right](n - iL^{\beta}), \tag{3.82}$$

where

$$i = \lfloor n/L^{\beta} \rfloor \tag{3.83}$$

$$W_i(k) = \lambda_i |Y(iL, k\Omega_N)|^2 C(k) \tag{3.84}$$

$$\lambda_i = \left[\sum_{k=0}^{N/2-1} |Y(iL, k\Omega_N)|^2 C(k)\right]^{-1} \tag{3.85}$$

$$C(k) = k^{-p} \tag{3.86}$$

$$p < 5/\log(N \times 1000\text{Hz}/f_s). \tag{3.87}$$

The values of L and p will be determined experimentally.

Synthesis

The estimate of the rate-changed version of $\theta_Y(n, k\Omega_N)$ can be related to that of $\theta(n, k\Omega_N)$ using (3.33) and (3.64),

$$\hat{\theta}_Y^{\beta}(n, k\Omega_N) = \hat{\theta}^{\beta}(n, k\Omega_N) + k(n - N/2)\Omega_N. \tag{3.88}$$

Substituting this result in the sampled transform implementation of (3.14) gives, with $\Omega_N = 2\pi/N$, the final synthesis equation

$$\hat{x}^{\beta}(n) = \frac{2}{Nh(0)} \sum_{k=0}^{N/2-1} \hat{M}_Y^{\beta}(n, k\Omega_N) \cos[\hat{\theta}_Y^{\beta}(n, k\Omega_N) + \pi k]. \tag{3.89}$$

Chapter 4

Simulation of a Time-Scale Modification System

A software simulation of the polar incremental TSM system proposed in Section 3.6 was conducted using the *C* programming language on a general purpose workstation. Three audio sources form the basis of the experimental data presented in this chapter: a male speaker, a female speaker and orchestral music.

4.1 Experimental Procedure

The experimental procedure was as follows: single channel 16-bit digital audio files were selected from a database, rate-changed by the simulation software, stored, then played back on a digital audio reproduction device. Informal listening tests were conducted using either headphones or a single loudspeaker.

4.2 Results

The experimental results are presented as waveform plots so that the structural impact of TSM may be appreciated. Subjective terms are used to assess the quality

of rate-changed speech.

The experiments were initially performed using audio files having a source sampling rate f_s of 8kHz. Some sort of time-varying phase distortion could be heard in the rate-changed speech, along with occasional quavering in the voiced portions and some smearing in the unvoiced portions. The problem largely disappeared when the sampling rate was increased to 16kHz. We recall that any increase in f_s does *not* improve the frequency resolution (in Hz) of the short-time Fourier analyzer if the duration (in ms) of the analysis filter is fixed. However, the *granularity* of the incremental parameter modification section is decreased. By this we mean that the structural impact of repeating and deleting "sample intervals" diminishes as the sampling rate increases Similarly, the effect of the waveform compensation section (i.e. phase modulation) becomes less noticeable if the duration (in ms) of the waveform events remains unchanged as f_s increases. The reason is that the amount of phase correction for the i-th rate-changed event is distributed over larger values of L^{β}, the length (in samples) of the rate-changed waveform events. Finally, less frequency aliasing and phase unwrapping errors occur for perceptually important frequency components as f_s increases because they are shifted away from the Nyquist frequency. Thus, for the sake of interest, we have chosen to report the experimental data relating to audio files having a source sampling rate of 16kHz. Consequently, the analysis filter length M was set to 320 (20ms) and the number of frequency samples N, to 512 (first power of 2 greater than M). The upper bound for the perceptual weighting factor p of the phase modulation algorithm, for the given f_s, is approximately 5.

Female Speaker

Figure 4.1 depicts an original utterance by a female speaker. The average pitch of the voiced portions is about 190Hz.

In order to verify the integrity of the polar synthesizer, we processed the utterance in the absence of any parameter modification ($\beta = 1$). Figure 4.2 shows the result of synthesizing the utterance with linear and cubic DSTFT phase interpolation. The original and synthesized signals were virtually indistinguishable in both cases. It was therefore decided to conduct the remaining experiments using linear phase interpolation.

Figure 4.3 illustrates a compressed version ($\beta = 2.0$) of the same utterance and Figure 4.4, an expanded version ($\beta = 0.5$). The perceptual quality of the rate-changed signals was excellent across the tested time-scaling factor range, $0.5 \leq \beta \leq 2.0$. Occasional smearing could be noticed at certain unvoiced phoneme boundaries for time-scale factors near 0.5, but otherwise the quality of the signals appeared to match that of the original utterance. If some form of distortion occurs in the case of time-compression, it is presumably masked by the accelerated rate of articulation.

One striking observation was made: better quality was obtained using *higher* values of the temporal sampling interval R rather than lower ones. For $R = 1$, the rate-changed signal had a slightly reverberant quality. It would appear that preserving the characteristics of the original DSTFT phase *exactly* is not desirable under a rate-change modification.

Figure 4.5 plots the first backward difference of the unwrapped phase sequence of a DSTFT harmonic over the 400-800ms portion of the utterance shown in Figure 4.1. The diagram reflects the variations in the instantaneous frequency of that harmonic over time. We chose a harmonic corresponding to a frequency where many phase irregularities are likely to occur. Despite the local disturbances, the average first backward difference holds steady at about $2\pi k/N$, the base frequency of the k-th DSTFT harmonic. The abrupt excursions (which never exceed π) are caused by voicing transitions and, presumably, by spectral energy shifts

among adjacent harmonics. The unwrapped phase sequences for harmonics having the highest spectral energy, usually those in the 0-1kHz range, tended to be the smoothest. Thus the use of linear phase interpolation appears justified for these harmonics.

The parameter values

- $p = 2$ or $p = 3$,
- $L = \lfloor 100 \times 10^{-3} f_s \rfloor$ for $\beta > 1$, and $L = \lfloor 100\beta \times 10^{-3} f_s \rfloor$ for $\beta < 1$,

were found to give good results. We also noted that the TSM system tends to attenuate rate-changed signals. In order to determine the source of the attenuation, we set the complex plane rotation factor $\epsilon(n)$ to zero and compared the outputs of the original and modified TSM systems. Thus it was determined that both $\epsilon(n)$ and the TSM algorithm attenuate rate-changed signals in distinct ways. For a single rate-changed waveform event, it appears that $\epsilon(n)$ contributes a fixed amount of attenuation, whereas the TSM algorithm contributes a time-varying component which probably reflects the degree of structural deterioration.

Figure 4.6 illustrates the result of concatenating individually rate-changed events when the waveform structure compensation stage is bypassed. The processing parameters are the same as those used in the time-scale expansion example of Figure 4.4. The location of the discontinuities (every 100ms) correspond to the waveform event boundaries. As we expected, structural deterioration is most severe in the neighborhood of sharp transients. For example, the onset characteristics of the voiced phoneme following the consonant burst labeled as *'D'* are very different from those of the original utterance shown in Figure 4.1.

Male Speaker

Figure 4.7 depicts an original utterance by a male speaker. The average pitch of the voiced portions is about 125Hz.

Figure 4.8 illustrates a compressed version ($\beta = 1.25$) of the same utterance. Mild quavering was occasionally noticed in the voiced portions of the rate-changed signals for $1 < \beta \leq 2.0$, especially near unvoiced-to-voiced speech boundaries, along with some degree of smearing in the unvoiced portions. The perceptual quality of the rate-changed signals was high otherwise, comparable to that of the original.

Figure 4.9 shows an expanded version ($\beta = 0.67$) of the original utterance. In general, the TSM algorithm did not perform as well for $\beta < 1$. Both the quavering and the smearing became more obvious as the time-scaling factor was decreased. For $\beta = 0.5$, the rate-changed signal was somewhat reverberant. The quality of expanded signals (in the male speaker category) may be described as "good" since the rate-change modification does not significantly impair the intelligibility of the speaker.

A spectrogram analysis of the expanded signals revealed that the harmonic structure of the original utterance had not be faithfully reproduced. Only the general formant structure had remained intact under the rate-change modification. The wideband spectrograms of Figure 4.10 show a voiced portion of the original utterance and the corresponding voiced portion of an expanded version. The vertical striations in the spectrogram of the expanded signal are less defined and more irregular.

We recall that the minimum pitch bound for obtaining a "clean" STFT band-pass representation of voiced speech is inversely proportional to the duration of the analysis window. The minimum pitch bound for a short-time Fourier analyzer

which uses a 20ms Hamming window was set at 200Hz. Since the original utterance violates this bound by about 80Hz, the distortion in the expanded signal may be attributed to insufficient frequency resolution in the Fourier analyzer. We tried increasing the analysis filter duration but this merely aggravated the reverberation in the rate-changed signals. In contrast, the female speaker utterance shown in Figure 4.1 does not violate the minimum pitch bound by more than 10Hz; consequently, the rate-changed versions were of better quality.

Music

We attempted to compress a particularly busy and rich passage of orchestral music. Some pitches were as low as 65Hz. The results were disappointing, even for mild compression factors. Severe beating marred the rate-changed output. Increasing the duration of the analysis filter to 32ms resulted in some improvement in the lower-pitched portions, yet to the detriment of higher-pitched ones. We did not find any particular values of L and p which could eliminate the distortion. It would appear, therefore, that the waveform structure of music is much more fragile than that of speech. Significant improvements would likely result if the sampling rate were increased, as we did initially for speech.

Background Noise

We simulated a multiple speaker environment by mixing the same sample files in varying proportions. The quality of each voice in the rate-changed mix was very similar to that obtained by rate-changing the voices individually. Lastly, injecting the same sample music in the speech files to simulate background noise resulted in no significant loss of quality in the rate-changed output.

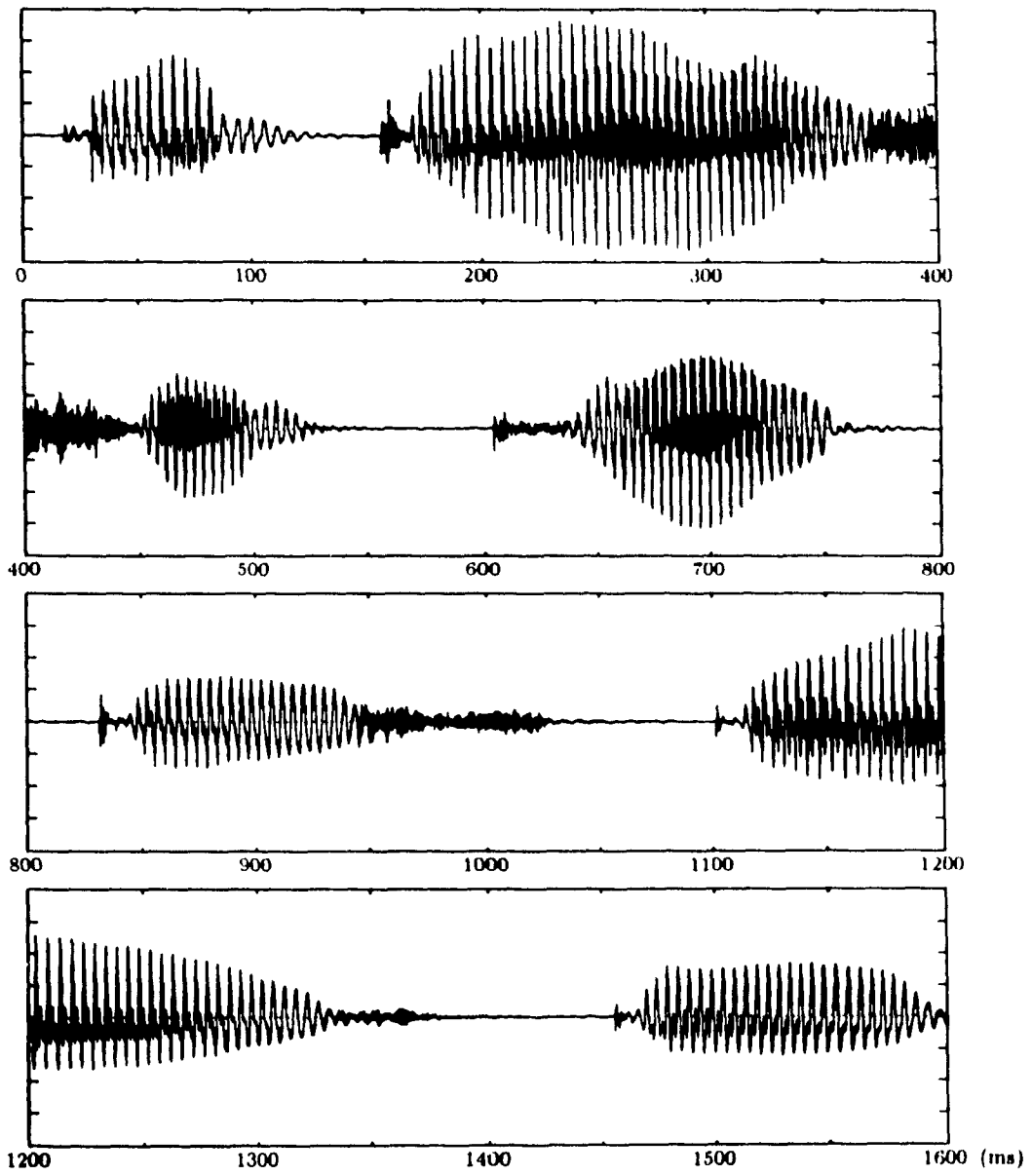


Figure 4.1: Original speech signal: "*A dash of pepper spoils bee...*[f stew]". Female speaker, f_s = 16kHz, 16-bit samples.

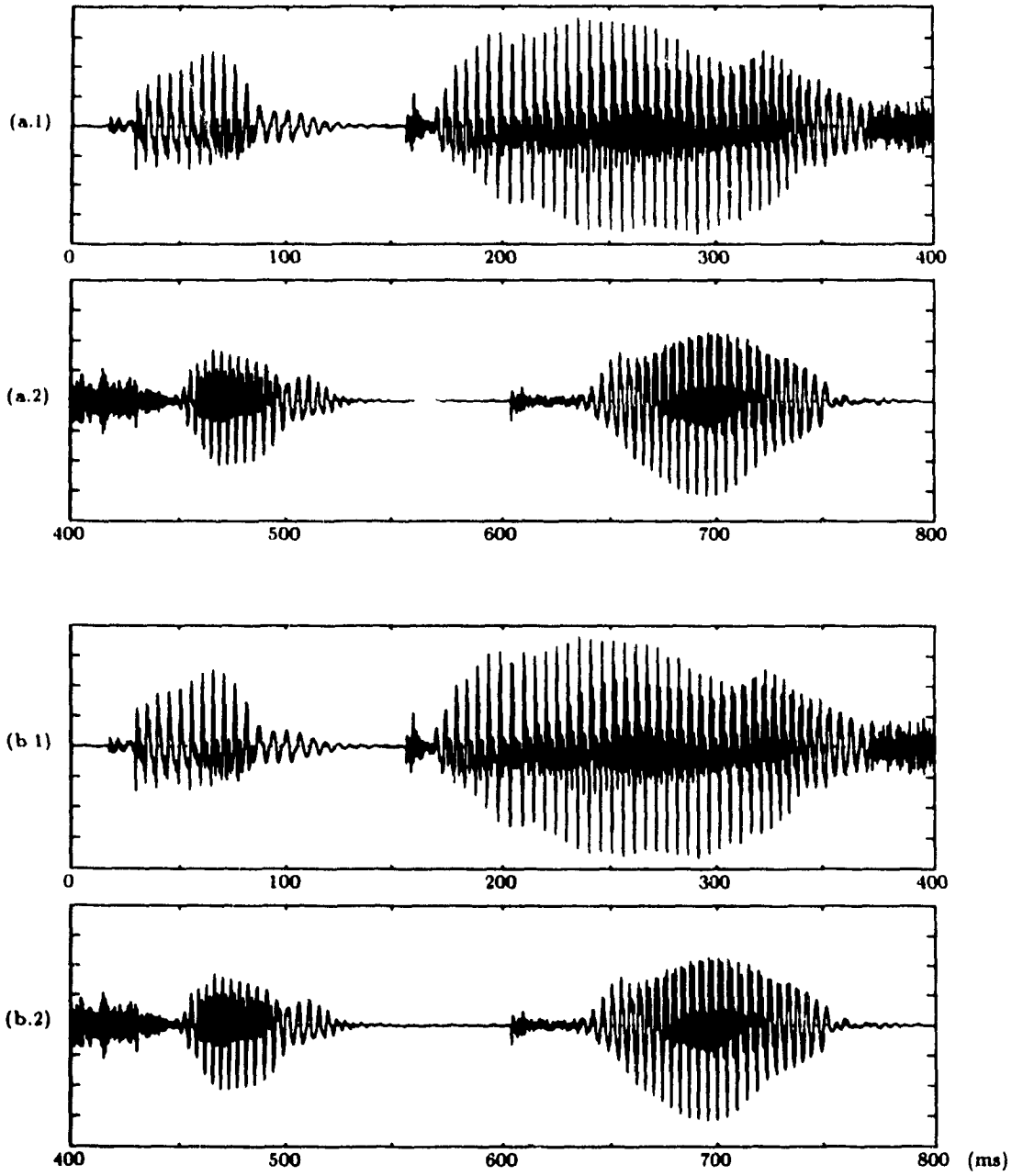


Figure 4.2: Comparison of polar synthesis methods (without any parameter modification). Speech signal: "*A dash of pe...*[pper spoils beef stew]". Female speaker, f_s = 16kHz, 16-bit samples. Processing parameters: $\beta = 1$, $M = 320$, $N = 512$, $R = 80$. a) Linear phase interpolation. b) Cubic phase interpolation.

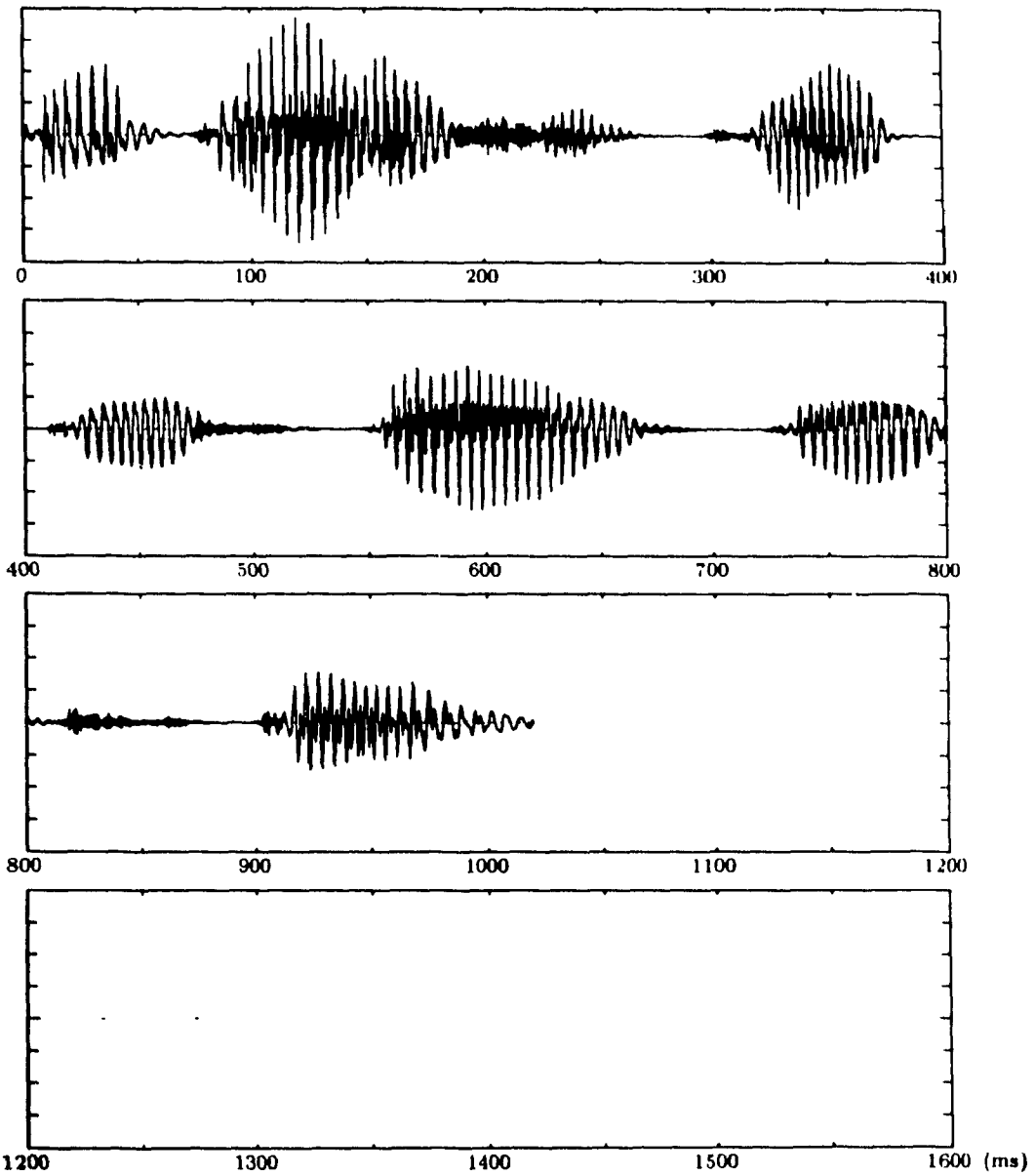


Figure 4.3: Rate-changed speech signal ($\beta = 2.0$) with phase modulation: "*A dash of pepper spoils beef stew*". Processing parameters: $M = 320$, $N = 512$, $R = 80$, $L = 1600$, $p = 2$, linear phase interpolation. Female speaker, $f_s = 16$kHz, 16-bit samples.

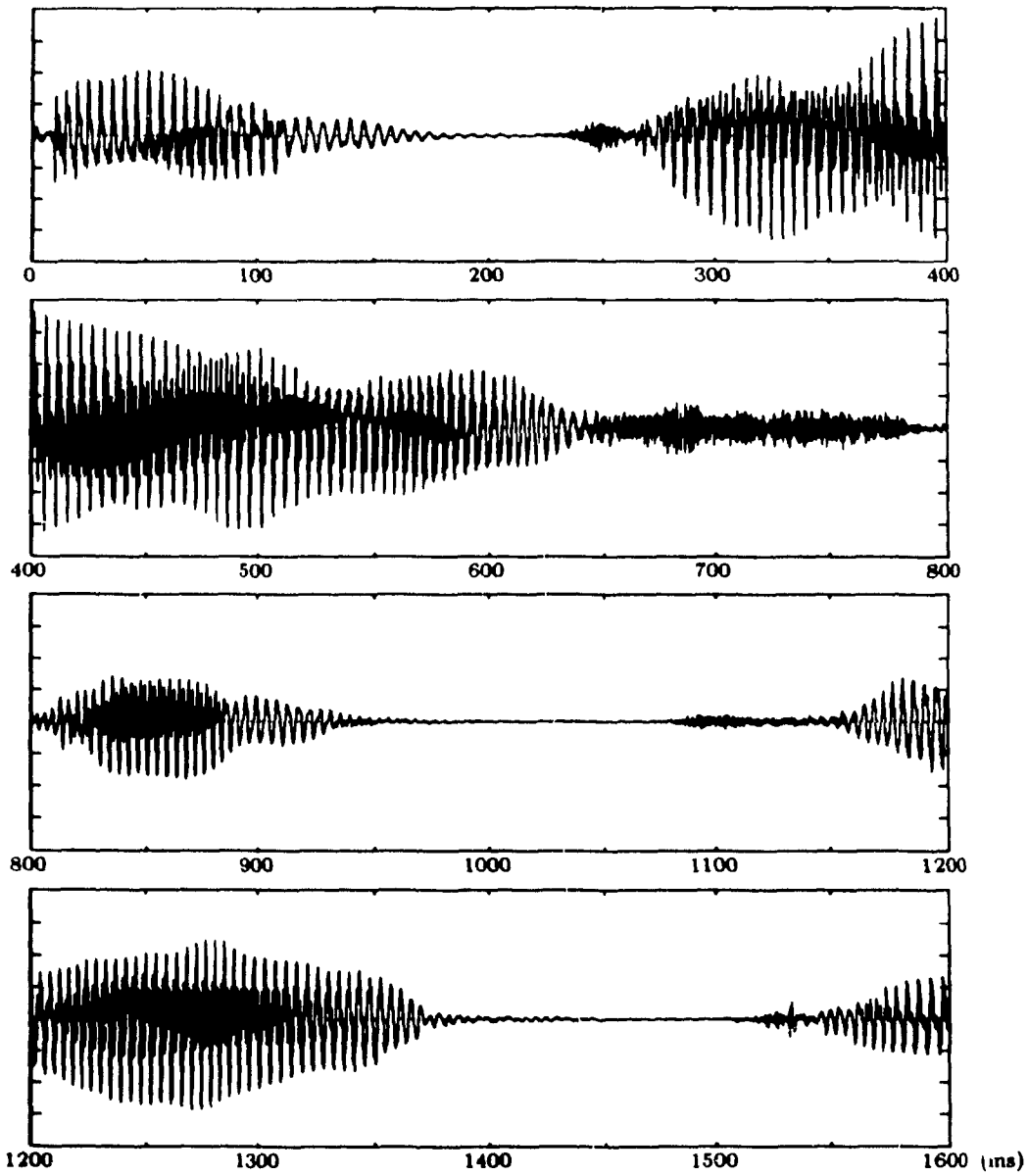


Figure 4.4: Rate-changed speech signal ($\beta = 0.5$) with phase modulation: "*A dash of peppe...*[r spoils beef stew]". Processing parameters: $M = 320$, $N = 512$, $R = 80$, $L = 800$, $p = 3$, linear phase interpolation. Female speaker, $f_s = 16$kHz, 16-bit samples.

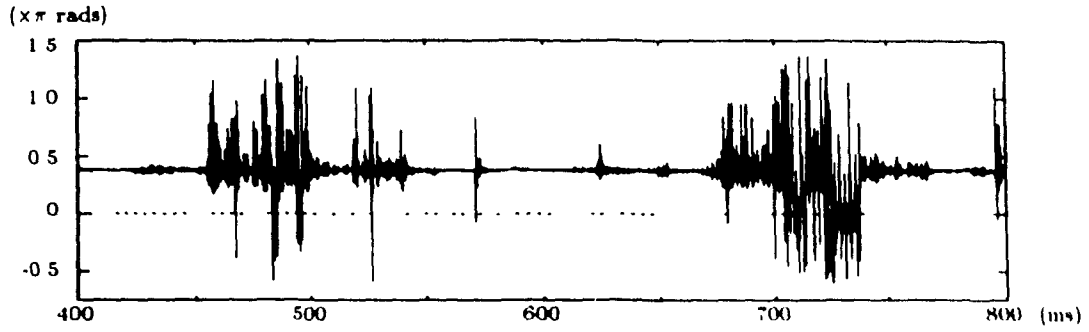


Figure 4.5: First backward difference of unwrapped phase of a DSTFT harmonic, $k = 96$ (3kHz), over 400–800ms portion of speech signal shown in Figure 4.1. The average value of the first backward difference is approximately $2\pi k/N$, the base frequency of the k-th DSTFT harmonic. Processing parameters: $M = 320$, $N = 512$, $R = 1$. Female speaker, $f_s = 16$kHz, 16-bit samples.

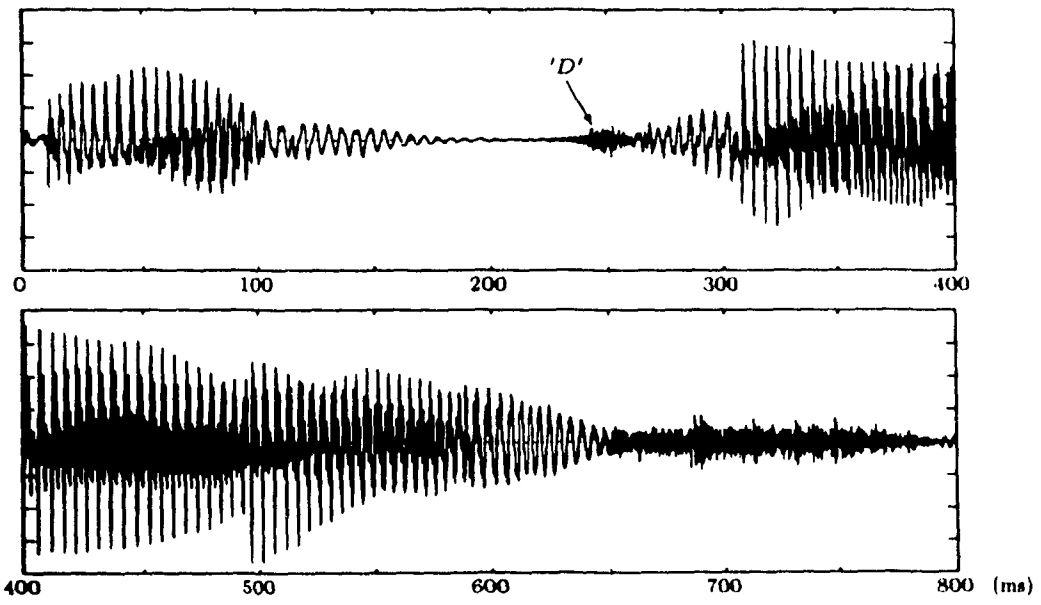


Figure 4.6: Concatenation of individually rate-changed waveform events ($\beta = 0.5$), without waveform structure compensation. Speech signal: "*A dash...*[of pepper spoils beef stew]". The discontinuities (every 100ms) correspond to waveform event boundaries. Processing parameters: $M = 320$, $N = 512$, $R = 80$, linear phase interpolation. Female speaker, $f_s = 16$kHz, 16-bit samples.

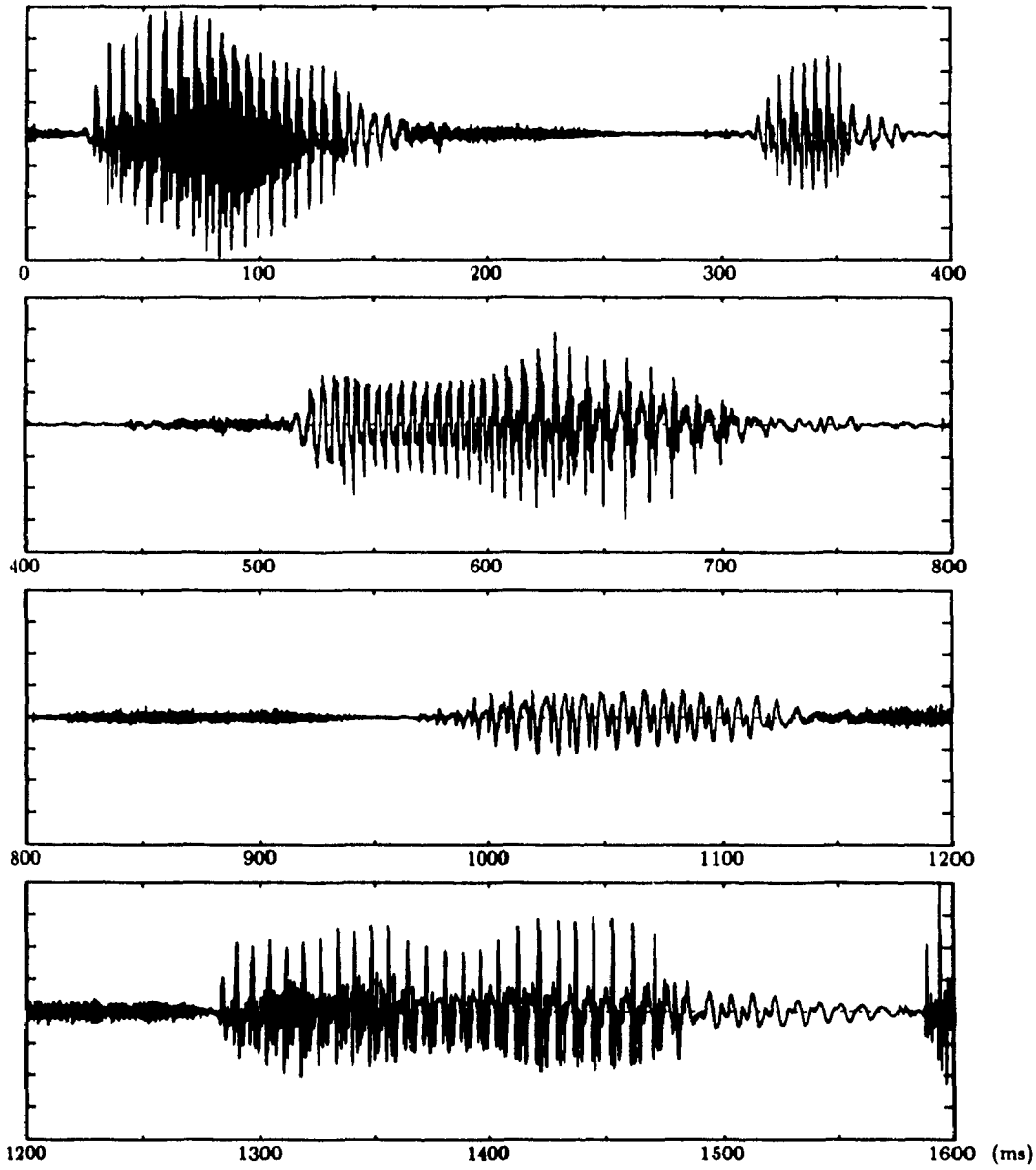


Figure 4.7: Original speech signal: "*Press the pants and sew a b...*[utton on the vest]". Male speaker, f_s = 16kHz, 16-bit samples.

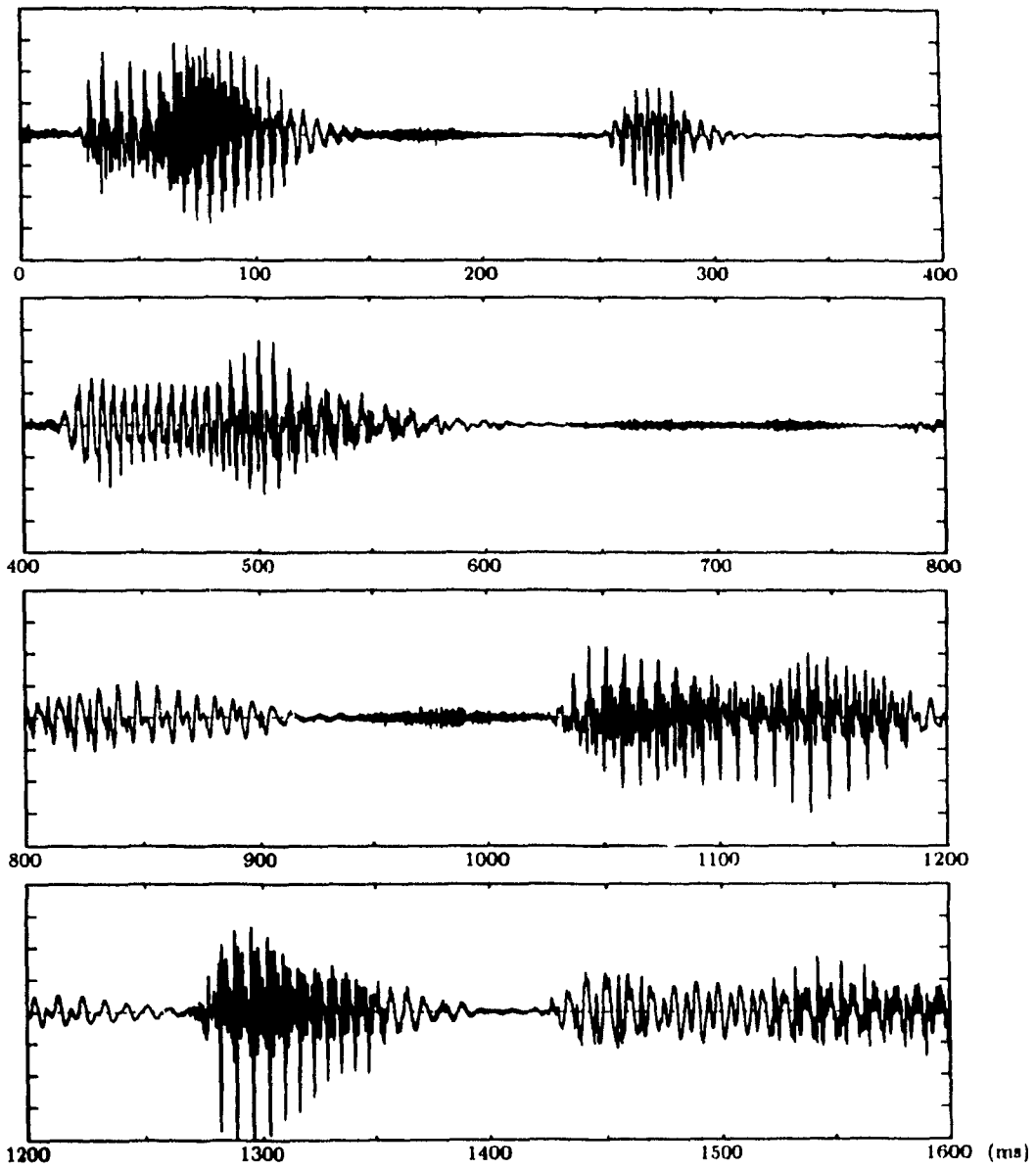


Figure 4.8: Rate-changed speech signal ($\beta = 1.25$) with phase modulation: "*Press the pants and sew a button on the...*[vest]". Processing parameters: $M = 320$, $N = 512$, $R = 80$, $L = 1600$, $p = 3$, linear phase interpolation. Male speaker, $f_s = 16$kHz, 16-bit samples.

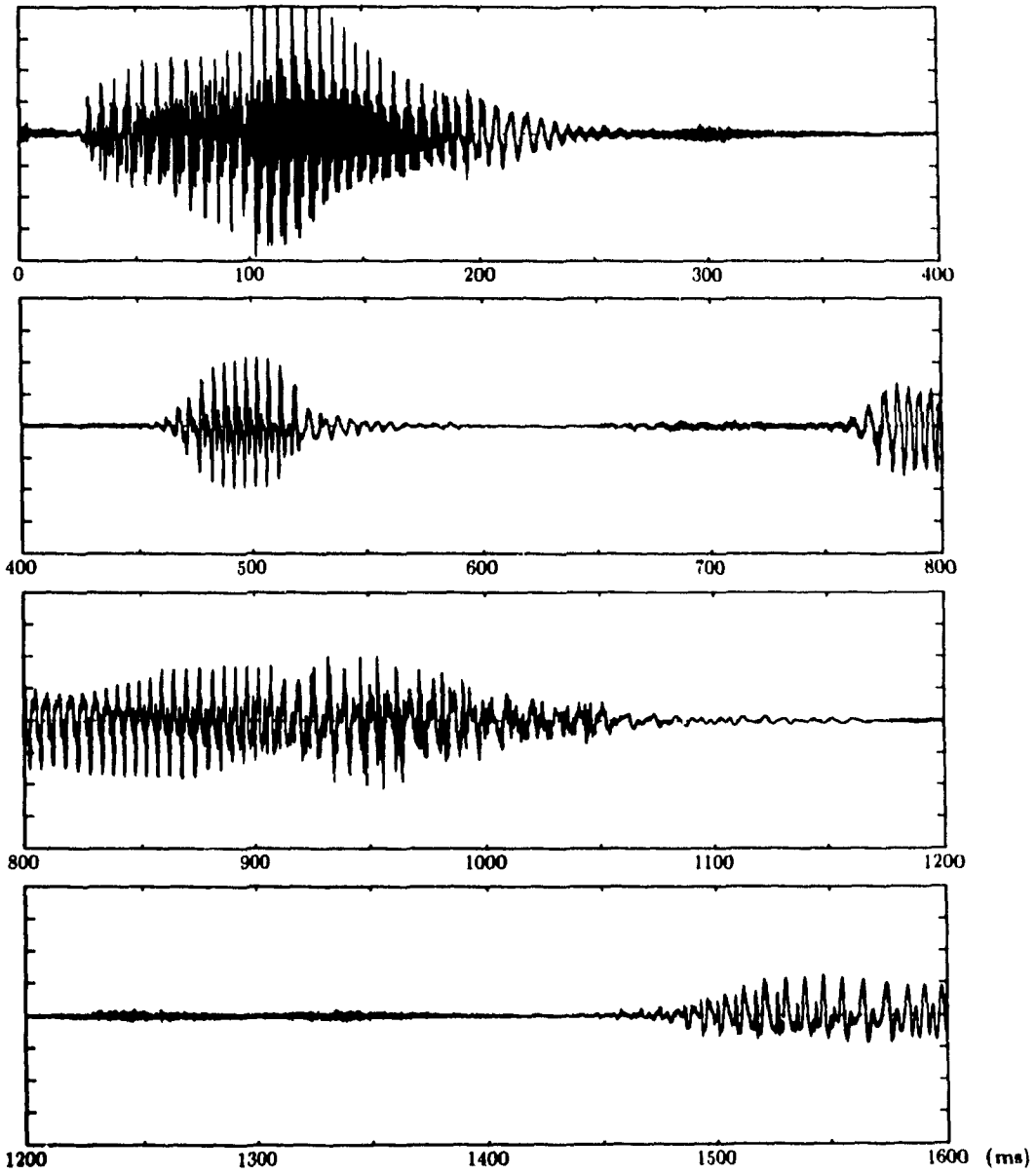


Figure 4.9: Rate-changed speech signal ($\beta = 0.67$) with phase modulation: "*Press the pants an...*[d sew a button on the vest]". Processing parameters: $M = 320$, $N = 512$, $R = 50$, $L = 1000$, $p = 2$, linear phase interpolation. Male speaker, $f_s = 16$kHz, 16-bit samples.

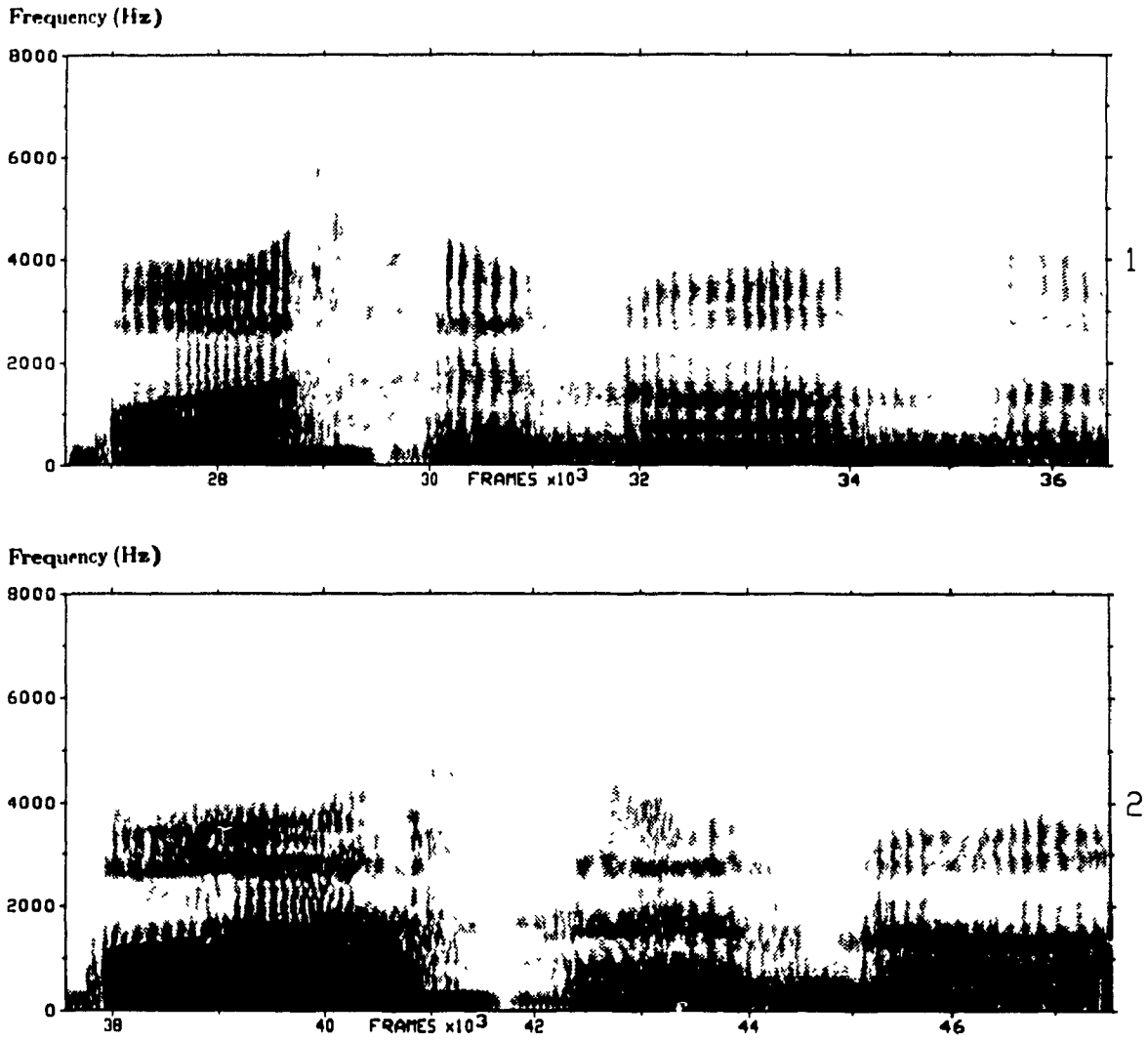


Figure 4.10: Spectrogram analysis of signal: "[Press the pants and sew a] ... ***button on*** ...[the vest]". Male speaker, f_s = 16kHz, 16-bit samples. 1) Original. 2) Expanded version ($\beta = 0.67$) with phase modulation, same processing parameters as in Figure 4.9. Note that the harmonic structure of the rate-changed signal is a distorted version of the original, i.e. the vertical striations in the second spectrogram are not as well defined as those in the first one.

Chapter 5

Conclusion

We have developed a simple time-frequency representation in which the temporal features of a signal $x(n)$ can be, in essence, manipulated independently from its spectral features to effect high-quality rate-change modifications at a low computational cost.

We began our analysis by postulating the same speech model which forms the basis of Portnoff's *time-scale modification* (TSM) approach [2]. It was assumed that a speech signal is generated by applying an excitation source to a linear time-varying filter which represents the spectral features of the vocal tract. For voiced speech, the excitation source is a unit-sample train where the unit-sample spacing corresponds to the local pitch period. In the case of unvoiced speech, the excitation source is a white noise sequence. Voiced speech was modeled as a sum of harmonically related complex exponentials, whereas unvoiced speech was characterized by its second order statistics, namely its auto-correlation and its power spectrum. Both the impulse response of the vocal tract and the pitch of the voiced speech excitation source were assumed to be nearly fixed for a duration which does not exceed 20ms.

The speech model was modified to represent rate-changed speech. It was found that rate-change modifications can be achieved by linearly time-scaling the speech

parameters. For voiced speech, however, one important exception was noted: the values of the instantaneous phase of the excitation source must be scaled by $1/\beta$ to preserve the original pitch of the speech. We recall that β denotes the time-scaling factor.

It was shown how the *short-time Fourier transform* (STFT) can be used to estimate the necessary speech parameters for implementing the rate-change modification. The key design requirements for the analysis filter $h(n)$ were identified. The length of $h(n)$ should be sufficiently short so that the speech parameters appear nearly fixed for the duration of the analysis interval. Proper resolution of voiced speech spectra requires that the bandwidth of the analysis filter be less than half the source pitch. It was argued that the same analysis filter bandwidth is adequate for estimating the time-varying power spectrum of unvoiced speech.

The STFT rate-change modification for voiced speech which Portnoff proposed consists of a linear time-scaling operation and a non-linear phase modification aimed at preserving the instantaneous frequency of the excitation source, as in the rate-changed speech model. Portnoff argued that the same STFT rate-change modification can be used for unvoiced speech.

It was shown that, in general, the non-linear STFT modification deteriorates the structure of rate-changed signals over time. We argued that the problem is further compounded by phase estimation errors, by sharp transients and, more importantly, by the *infinite memory* property of Portnoff's TSM method, which causes phase errors to accumulate indefinitely. Thus the need for waveform structure compensation in rate-changed signals was identified.

Several design options for the components of a complete TSM system were examined on the basis of complexity, audio quality and robustness. The temporal sampling requirement of the *discrete* STFT (DSTFT) raised a number of interpolation issues, both in the parameter modification and DSTFT synthesis stages.

Portnoff suggested the use of bandlimited interpolation on the real and imaginary parts of the original and rate-changed DSTFT. We argued that it may not be perceptually appropriate to treat the real and imaginary parts of a DSTFT as independent bandlimited sequences. We also pointed out that Portnoff's synthesis equation is inefficient in that it requires a polar to rectangular coordinate conversion for every DSTFT harmonic.

It was proposed instead to process the polar parameters (i.e. magnitude and unwrapped phase) of the DSTFT, both in the parameter modification and DSTFT synthesis stages. Since polar parameter sequences are not necessarily bandlimited, exact signal representations are generally impossible. However, the computational requirements for interpolating polar parameters are believed to be modest for achieving high-quality synthesis. For example, in the absence of any parameter modification, it was found that the synthesized and original signals are virtually indistinguishable when the polar parameters of the DSTFT are linearly interpolated. The proposed polar synthesis equation resembles the one used by McAulay and Quatieri for their sinusoidal speech model (SSM) [10]. The main difference is that the sinusoidal components of our implementation employ fixed rather than time-varying base frequencies.

A novel incremental parameter modification scheme which exploits the smoothness of the polar parameters was suggested, eliminating the need for *explicit* linear time-scaling and multiplications by $1/\beta$. Rate-change modifications are effectively achieved by periodically deleting or repeating "sample intervals" in the original signal. The scheme allows *variable* TSM to be implemented easily. However, the incremental approach shuns the original speech model in that it preserves the instantaneous frequency of the DSTFT harmonics themselves instead of that of the excitation source, leading, in principle, to more rapid structural deterioration of the rate-changed signal.

The primary source of distortion in rate-changed speech is the structural deterioration caused by the non-linear STFT modification. Portnoff's method does not alleviate the problem. We therefore chose the polar incremental synthesis approach for its efficiency and decided to implement a waveform structure compensation section. Two options were considered.

The waveform interpolation method consists of overlapping and interpolating consecutive rate-changed speech segments. A speech segment was referred to as a "waveform event".

The preferred method, phase modulation, is more robust and consists of modulating the rate-changed DSTFT harmonics of a waveform event such that their phase *relationship* on the new time-scale at intervals of $\lfloor L/\beta \rfloor$ samples matches that of the original DSTFT harmonics (on the unity time-scale) at intervals of L samples. We recall that the constant L denotes the length of the waveform events. A perceptual weighting scheme was proposed to reduce the impact of the resulting spectral distortion. The phase data for at least one waveform event must be buffered to implement this waveform structure compensation method.

The simulation results indicated that the proposed TSM system (polar incremental synthesis + phase modulation) is capable of generating high-quality rate-changed versions of speech signals recorded under a variety of conditions. The tested time-scale factor range was $0.5 \leq \beta \leq 2.0$, which should be sufficient for most applications. The best quality should be achieved for speech signals where the voiced portions satisfy the minimum pitch bound (in Hz) which ensures proper resolution of voiced speech spectra. This bound is inversely proportional to the duration (in ms) of the analysis filter of the short-time Fourier analyzer. The bound for our particular implementation (20ms Hamming window) is 200Hz. The duration of the analysis window cannot be increased because the variations in the speech parameters over analysis intervals greater than 20ms are no longer neg-

ligible. Consequently, some distortion due to insufficient frequency resolution is expected for lower pitched speech signals, such as those produced by male speakers. Time-scale compression tends to mask this distortion because the rate of articulation is accelerated. For time-scale expansion, however, quavering in the voiced portions and smearing in the unvoiced portions become more apparent as β is decreased. The perceptual quality of expanded speech (in the male speaker category) can be rated as "good" because the speaker remains quite intelligible.

We noticed a dramatic improvement in the perceptual quality of rate-changed speech by increasing the source sampling rate f_s from 8kHz to 16kHz. Since the duration (in ms) of the analysis window is fixed, any increase in f_s does *not* improve the frequency resolution (in Hz) of the short-time Fourier analyzer. There are, however, at least three ways in which f_s affects the quality of rate-changed speech.

- An increase in f_s reduces the granularity of the incremental parameter modification algorithm. Granularity refers to the structural impact of deleting and repeating "sample intervals".

- The effect of the waveform structure compensation section (i.e. phase modulation) becomes less noticeable if the duration (in ms) of the waveform events remains unchanged as f_s increases. The reason is that the amount of phase correction is distributed over a larger number of samples.

- As f_s increases, less frequency aliasing and phase unwrapping errors occur for perceptually important frequency components because they are shifted away from the Nyquist frequency.

The experiments suggested that, under a rate-change modification, it is not desirable to preserve the characteristics of the original DSTFT phase *exactly*. It

was also found that the current implementation is unsuitable for arbitrary music sources.

The initial success of the proposed TSM algorithm for speech signals suggests that time-frequency models should not strive in obtaining exact signal representations, as these may be superfluous. We have found that the quality of rate-changed signals synthesized from the STFT-based approach is compromised mainly by the time-frequency resolution of the analysis filter. Unfortunately, this compromise is irremediable.

Future research efforts should perhaps focus on eliminating the need for waveform structure compensation, as this constitutes a convenient fix, not a real solution. The optimal estimate for the rate-changed unwrapped STFT phase $\theta^{\beta}(n,\omega)$ should ideally be a non-linear function of past and future rate-changed samples in the vicinity of n on the *new* time-scale. It is not likely, therefore, that $\theta^{\beta}(n,\omega)$ can be estimated from phase data calculated on the original time-scale alone, without some form of recursion or analysis-by-synthesis.

Since the proposed TSM approach produces a high-quality *local* estimate of the rate-changed signal, perhaps a recursive method incorporating a polar incremental TSM algorithm would require fewer iterations than the *least-squares error estimation* (LSEE) class of TSM algorithms [3, 4, 5, 6] to determine the optimal phase and magnitude fit of the original DSTFT harmonics on a new time-scale. However, the prospect of recursion compromises our initial objective of low computational cost, leaving us to wonder whether the STFT framework is too rigid for TSM applications. Other signal representations [33] aimed at modeling the human auditory process rather than the sound source might well be worth exploring.

Bibliography

[1] G. Fairbanks, W. L. Everitt and R. P. Jaeger, "Method for time or frequency compression-expansion of speech," *IRE Trans. Professional Group on Audio*, vol. AU-2, pp. 7–12, Jan.–Feb. 1954.

[2] M. R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Transactions on Acoust., Speech, Signal Processing*, vol. 29, pp. 374–390, June 1981.

[3] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoust., Speech, Signal Processing*, vol. 32, pp. 236–242, Feb. 1984.

[4] S. Roucos and A. M. Wilgus, "High-quality time-scale modification for speech," *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Tampa, Florida, pp. 493–496, 1985.

[5] M. Abe, S. Tamura and H. Kuwabara, "A new speech modification method by signal reconstruction," *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Glasgow, pp. 592–595, 1989.

[6] E. Hardam, "High-quality time-scale modification of speech signals using fast synchronized overlap-add algorithms," *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Albuquerque, NM, pp. 409–412, 1990.

[7] M. K. Asi and B. E. A. Saleh, "A linear filter for time scaling of speech," *Proc. Int. Conf. Acoust., Speech, Signal Processing*, New York, NY, pp. 79–82, 1988.

[8] M. K. Asi and B. E. A. Saleh, "A linear periodically time-varying filter for time-frequency scaling of speech," *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Albuquerque, NM, pp. 405–408, 1990.

[9] T. F. Quatieri and R. J. McAulay, "Speech transformations based on a sinusoidal representation," *IEEE Transactions on Acoust., Speech, Signal Processing*, vol. 34, pp. 1449–1464, Dec. 1986.

[10] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoust., Speech, Signal Processing*, vol. 34, pp. 744–754, Aug. 1986.

[11] N. S. Jayant, "High-quality coding of telephone speech and wideband audio," *IEEE Commun. Mag.*, pp. 10–20, Jan. 1990.

[12] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Sel. Areas in Commun.*, vol. 6, pp. 314–323, Feb. 1988.

[13] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, p. 45, 1989.

[14] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, pp. 101–112, 1989.

[15] T. A. Ramstad, "Digital methods for conversion between arbitrary sampling frequencies" *IEEE Transactions on Acoust., Speech, Signal Processing*, vol. 32, pp. 577–591, June 1984.

[16] D. O'Shaughnessy, *Speech Communication*, Addison-Wesley, Reading, MA, p. 39, 1987.

[17] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, p. 98, 1978.

[18] R. W. Schafer and L. R. Rabiner, "Digital representations of speech signals," *Proc. IEEE*, vol 63, pp. 662–677, Apr. 1975.

[19] M. R. Portnoff, "Short-time Fourier analysis of sampled speech," *IEEE Transactions on Acoust., Speech, Signal Processing*, vol. 29, pp. 364–373, June 1981.

[20] A. V. Oppenheim, A. S. Willsky with I. T. Young, *Signals and Systems*, Prentice-Hall, Englewood Cliffs, NJ, p. 316, 1983.

[21] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, NY, p. 225, 1984.

[22] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, NY, p. 272, 1984.

[23] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, p. 251, 1978.

[24] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, p. 118, 1978.

[25] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, p. 266, 1978.

[26] J. M. Heinz and K. N. Stevens, "On the properties of voiceless fricative consonants," *J. Acoust. Soc. Amer.*, vol. 33, pp. 589–596, May 1961.

[27] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, p. 450, 1989.

[28] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, pp. 527–530, 1989.

[29] D. O'Shaughnessy, *Speech Communication*, Addison-Wesley, Reading, MA, pp. 206–207, 1987.

[30] G. Oetken, T. W. Parks and H. W. Schüssler, "New results in the design of digital interpolators," *IEEE Transactions on Acoust., Speech, Signal Processing*, vol. 23, pp. 301–309, June 1975.

[31] D. O'Shaughnessy, *Speech Communication*, Addison-Wesley, Reading, MA, pp. 148–150, 1987.

[32] L. B. Almeida and F. M. Silva, "Variable-frequency synthesis: an improved harmonic coding scheme," *Proc. Int. Conf. Acoust., Speech, Signal Processing*, San Diego, CA, pp. 27.5.1–27.5.4, 1984.

[33] W. Heinbach, "Aurally adequate signal representation: the part-tone-time-pattern," *Acustica*, vol. 67, pp. 113–121, Dec. 1988.