

AD/

ESD-TR-85-313

Technical Report
717

Speech Transformations Based on a Sinusoidal Representation

T.F. Quatieri
R.J. McAulay

16 May 1986

Lincoln Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

*L*EXINGTON, *M*ASSACHUSETTS



Prepared for the Department of the Air Force
under Electronic Systems Division Contract F19628-85-C-0002.

Approved for public release; distribution unlimited.

ADA 169740

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, with the support of the Department of the Air Force under Contract F19628-85-C-0002.

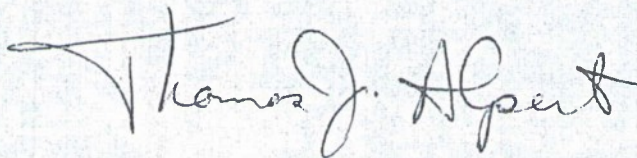
This report may be reproduced to satisfy needs of U.S. Government agencies.

The views and conclusions contained in this document are those of the contractor and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the United States Government.

The ESD Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

A handwritten signature in black ink that reads "Thomas J. Alpert". The signature is written in a cursive style with a large, stylized initial 'T'.

Thomas J. Alpert, Major, USAF
Chief, ESD Lincoln Laboratory Project Office

Non-Lincoln Recipients

PLEASE DO NOT RETURN

Permission is given to destroy this document
when it is no longer needed.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

**SPEECH TRANSFORMATIONS BASED
ON A SINUSOIDAL REPRESENTATION**

*T.F. QUATIERI
R.J. McAULAY*

Group 24

TECHNICAL REPORT 717

16 MAY 1986

Approved for public release; distribution unlimited.

LEXINGTON

MASSACHUSETTS

ABSTRACT

In this report, a new speech analysis/synthesis technique is presented which provides the basis for a general class of speech transformations including time-scale modification, frequency scaling, and pitch modification. These modifications can be performed with a time-varying change, permitting continuous adjustment of a speaker's fundamental frequency and rate of articulation. The method is based on a sinusoidal representation of the speech production mechanism that has been shown to produce synthetic speech that preserves the waveform shape and is essentially perceptually indistinguishable from the original. Although the analysis/synthesis system originally was designed for single-speaker signals, it is equally capable of recovering and modifying nonspeech signals such as music, multiple speakers, marine biologic sounds, and speakers in the presence of interferences such as noise and musical backgrounds.

TABLE OF CONTENTS

| | |
|--|-----|
| Abstract | iii |
| List of Illustrations | vii |
| 1. INTRODUCTION | 1 |
| 2. ANALYSIS/SYNTHESIS BASED ON A SINE-WAVE SPEECH MODEL | 5 |
| 3. TIME-SCALE MODIFICATION | 19 |
| 4. FREQUENCY TRANSFORMATIONS | 39 |
| 5. JOINT TIME-FREQUENCY MODIFICATIONS | 49 |
| 6. DISCUSSION | 53 |
| References | 55 |

LIST OF ILLUSTRATIONS

| Figure No. | | Page |
|---------------|---|------|
| 2.1 | Sinusoidal Representation of Speech Production | 6 |
| 2.2 | Block Diagram of Sinusoidal Analysis | 9 |
| 2.3 | Requirement of Unwrapped System Phase | |
| | (a) Interpolation of Wrapped System Phase | |
| | (b) Interpolation of Unwrapped System Phase | 11 |
| 2.4 | Block Diagram of Sinusoidal Synthesis | 15 |
| 2.5 | Reconstruction of Signal from Single Speaker | |
| | (a) Original | |
| | (b) Reconstruction | 16 |
| 2.6 | Reconstruction of Signal from Two Speakers | |
| | (a) Original | |
| | (b) Reconstruction | 16 |
| 2.7 | Reconstruction of Speech in a Music Background | |
| | (a) Original | |
| | (b) Reconstruction | 17 |
| 3.1 | Time Warping with Fixed Rate Change $\rho > 1$ | 20 |
| 3.2 | Model of Time-Scale Modification | 20 |
| 3.3 | Time-Scale Expansion of a Single Sine Wave | 21 |
| 3.4 | Functional Mappings for Time-Scale Expansion | 23 |
| 3.5 | Block Diagram of Uniform Rate-Change System | 24 |
| 3.6 | Flow Diagram of Computer Implementation of Uniform Rate-Change System | 26 |
| 3.7 | Time-Scale Modification of Synthetic Waveform | |
| | (a) Original | |
| | (b) Reconstruction ($\rho = 1.0$) | |
| | (c) Expansion ($\rho = 1.5$) | |
| | (d) Compression ($\rho = 0.5$) | 27 |
| 3.8 | Time-Scale Expansion of Speech | |
| | (a) Original | |
| | (b) Expansion ($\rho = 2$) | 27 |

| Figure No. | | Page |
|-----------------------|--|-------------|
| 3.9 | Time-Scale Compression of Speech (a) Original (b) Compression ($\rho = 0.5$) | 28 |
| 3.10 | Time-Scale Expansion of Speech in Music (a) Original (b) Expansion ($\rho = 1.5$) | 28 |
| 3.11 | Time-Scale Expansion of Combined Male and Female Speech (a) Original (b) Expansion ($\rho = 2$) | 29 |
| 3.12 | Piecewise Constant Rate Change (a) Rate Change Function (b) Time-Warp | 30 |
| 3.13 | System Phase Mapping for the Piecewise Constant Rate Change of Figure 3.12 | 31 |
| 3.14 | Flow Diagram of Computer Implementation of Nonuniform Rate-Change System | 35 |
| 3.15 | Speech Segments from the Passage, "She Fell From the Car," with Superimposed Spectral Derivative. The Spectral Derivative Has Been Normalized to Lie Between Zero and Unity and Was Assumed Constant over an Analysis Frame | 37 |
| 4.1 | Time-Frequency Illustration of Frequency Compression (a) Original (b) Frequency Compressed | 40 |
| 4.2 | Frequency Compression of the Spectral Magnitude (a) Original (b) Compression | 41 |
| 4.3 | Sinusoidal Model for Pitch Modification | 43 |
| 4.4 | Time-Frequency Illustration of Pitch Modification (a) Original (b) Pitch-Scaled (Lowered) | 43 |

| Figure No. | | Page |
|------------|---|------|
| 4.5 | Pitch Modification of Synthetic Waveform | |
| | (a) Original | |
| | (b) Increase in Pitch ($\beta = 1.5$) | |
| | (c) Decrease in Pitch ($\beta = 0.5$) | 44 |
| 4.6 | Pitch Modification of Speech in the Frequency Domain | |
| | (a) Original | |
| | (b) Pitch-Scaled Spectral Magnitude ($\beta = 1.5$) | 45 |
| 4.7 | Pitch Modification of Speech in the Time Domain | |
| | (a) Original | |
| | (b) Pitch-Scaled ($\beta = 0.8$) | 46 |
| 5.1 | Joint Frequency Scaling and Time-Scale Modification | |
| | (a) Original | |
| | (b) Frequency Compression and Time-Scale Expansion | |
| | (c) Inversion of Figure 5.1(b) | 50 |

SPEECH TRANSFORMATIONS BASED ON A SINUSOIDAL REPRESENTATION

1. INTRODUCTION

In a number of important applications, it is desirable to transform a speech waveform to a signal which is more useful than the original¹. In time-scale modification², for example, the rate of articulation is slowed down to make degraded speech more comprehensible. Alternatively, speech is speeded up in order to quickly scan a passage or compress an utterance into an allocated time interval. In other applications, the speech is compressed or expanded in frequency. In particular, frequency compression is useful in bandwidth reduction³ or in placing the speech into a desired frequency range as an aid to the hearing impaired⁴. Another application requires that the fundamental frequency of the speaker be modified while preserving the shape of the envelope of the short-time speech spectrum. This operation is useful in psychoacoustic research⁵ or in correcting pitch disjunctions in concatenated speech segments⁶. In a number of these applications, it is sometimes desired to perform speech modifications which vary in time or to perform modifications simultaneously. For example, in time-scale modification it is of interest to have the means to continuously adjust a speaker's rate of articulation, while in concatenating speech segments both the time scale and pitch may require modification.

In this report, a speech analysis/synthesis system is presented which forms the basis of a general class of such transformations. The system is based on a sinusoidal representation of speech which incorporates a model of speech production, but which is independent of the speech state and of the pitch. The reconstruction requires an estimate of the excitation and vocal tract contributions of the amplitude and phase of each component of the underlying sine-wave model. Functional representations of these parameters are derived from short-time Fourier transform samples which correspond to the sine-wave components. The resulting analysis/synthesis system thus represents a refinement of a purely sine-wave-based analysis/synthesis procedure described in a previous report⁷. The new system has been applied to obtain high-quality time-scale modification, frequency scaling, and scaling of fundamental frequency. These operations can be performed independently of one another or simultaneously and can also be applied with time-varying changes. For example, a speaker's pitch can be continuously changed while continuously changing the rate of articulation. Furthermore, the system does not break down either for a large class of nonspeech signals or for speech corrupted by interferences such as a second speaker or acoustic background noise in the sense that the background is not perceived as different from the original.

Numerous other methods have been proposed for modification of the speech waveform. One of the earlier approaches, based on classical vocoders^{1,8}, utilizes pitch and voiced/unvoiced decisions in the excitation and an estimate of the vocal tract system function. Although this procedure is suitable for a wide range of speech transformations, errors in pitch and voiced/unvoiced state decisions typically introduce artifacts into the synthetic modified speech. A more recent approach that does not require pitch extraction and voiced/unvoiced decisions manipulates

an excitation obtained by deconvolving the original speech with a vocal tract spectral envelope estimate⁹. This procedure thus relies on the speech production model, as do classical vocoders, but avoids many of the problems inherent in the vocoder approach. Another class of methods widely used in the application of time-scale modification is based on the Fairbank's method¹⁰. This technique periodically repeats or discards segments of the speech waveform, a method prone to boundary discontinuities. Refinements of this technique involve "pitch synchronous" splicing of the waveform¹¹. A further improvement was introduced by Neuberg¹² who smoothly merged adjacent speech segments to reduce discontinuities. Another approach to guaranteeing a smooth synthetic waveform from speech segments uses an iterative method for reconstructing the modified waveform from the short-time Fourier transform magnitude^{13, 14}. Although computationally burdensome for practical applications, this method yields very high-quality rate-altered speech.

A number of approaches to analysis/synthesis based on sine-wave models have been discussed in the literature. Hedelin¹⁵ proposed a pitch-independent sine-wave model for use in coding the baseband signal for speech data-rate compression. The amplitudes and frequencies of the underlying sine waves are estimated using Kalman filtering techniques and the sine-wave phases are obtained by integrating the instantaneous frequencies. The use of this system for speech transformations was not explored. Other high-quality systems based on a sine-wave representation have been applied to time-scale modification of speech^{16, 17} and frequency scaling¹⁷. The system by Portnoff¹⁶, a refinement of the phase vocoder¹, in addition, represents each sine-wave component by vocal cord excitation and vocal tract system contributions. In contrast to Hedelin's approach and the approach taken in this report, these analysis/synthesis systems are based on an underlying representation which constrains the sine-wave components to be harmonically related. Furthermore, the analysis in these systems does not explicitly estimate the sine-wave components, but rather views them as outputs of a bank of uniformly-spaced bandpass filters. The synthesis can be viewed as summing the modified output of this filter bank. A system which may be applicable to time-scale modification is also based on a harmonic sine-wave model and uses sine-wave generators explicitly in the synthesizer¹⁸. To compensate for the inadequacies in the harmonic model, a residual waveform is computed which in turn must be time-scale modified, a problem which has not yet been investigated.

As indicated above, in contrast to earlier sine-wave based systems, the analysis/synthesis system of this report explicitly estimates the amplitude and phase of the vocal cord excitation and vocal tract system function contributions to each sine wave. These estimates are obtained from the short-time Fourier transform evaluated at frequencies corresponding to the location of spectral peaks. Since the frequencies of the sine waves are not constrained to be harmonic, pitch is not required in the analysis. The synthesis uses the amplitude and phase estimates to obtain a *functional* representation of the time evolution of each parameter. This particular functional representation is the key to all speech transformations and allows for a flexibility (i.e., joint time-varying transformations) not present in other high-quality systems.

This report is organized as follows. The sinusoidal speech representation upon which the analysis/synthesis system is based is described in Section 2. Then in Section 3 the problem of

time-scale modification is addressed in two parts. First the time scale is allowed to change uniformly and then the generalization to a time-varying adjustment of the time scale is developed. The generalized system is used in attempting to more closely simulate actual rate-changed speech by adapting the time-scale to features of the speech waveform. In Section 4 the problems of frequency scaling and pitch modification are addressed. As in Section 3 both uniform and time-varying modifications are considered. Finally, in Section 5 the time and frequency modifications are combined into a single system capable of performing the operations jointly.

2. ANALYSIS/SYNTHESIS BASED ON A SINE-WAVE SPEECH MODEL

In this section, the sinusoidal representation of speech presented in Reference 2 is first reviewed. A new analysis/synthesis system is then developed which refines the analysis/synthesis procedure in Reference 2 by separating the vocal cord excitation and vocal tract system contributions underlying each component of the sine-wave model.

2.1 The Sinusoidal Representation

In the speech production model, the speech waveform $s(t)$ is assumed to be the output of passing a glottal excitation waveform $e(t)$ through a linear time-varying system with impulse response $h(t, \tau)$, representing the characteristics of the vocal tract. Mathematically, this can be written as

$$s(t) = \int_0^t h(t, t - \tau)e(\tau)d\tau \quad (2.1)$$

The excitation will be represented by a sum of sine waves of arbitrary amplitudes, frequencies, and phases:

$$e(t) = \sum_{\ell=1}^{L(t)} a_{\ell}(t)\cos[\Omega_{\ell}(t)] \quad (2.2a)$$

where

$$\Omega_{\ell}(t) = V_{\ell}(t) + \phi_{\ell} \quad (2.2b)$$

with

$$V_{\ell}(t) = \int_{t_{\ell}}^t \omega_{\ell}(\sigma)d\sigma \quad (2.2c)$$

where t_{ℓ} is the onset time of the ℓ th sine wave and where $L(t)$ is the number of sine-wave components at time t . For the ℓ th component, $a_{\ell}(t)$ and $\omega_{\ell}(t)$ are the slowly time-varying amplitude and frequency (i.e., the parameters are essentially constant over a 20-30 ms analysis window), $V_{\ell}(t)$ is the changing contribution to the excitation phase [e.g., for a steady-state tone, $V_{\ell}(t)$ is a linear ramp,] and ϕ_{ℓ} is the fixed phase offset which accounts for the fact that the sine waves will generally not be in phase. The vocal tract transfer function is given by the Fourier transform of the system response $h(t, \tau)$ for each t and will be denoted by $H(\omega, t)$:

$$H(\omega, t) = M(\omega, t)\exp[j\Phi(\omega, t)] \quad (2.3)$$

where $M(\omega, t)$ and $\Phi(\omega, t)$ are the amplitude and phase of $H(\omega, t)$, respectively. With these definitions, the speech production mechanism can be represented in the frequency domain, as depicted in Figure 2-1. Since the vocal tract system function is linear, each sine-wave component of the excitation is independently affected by the system function.

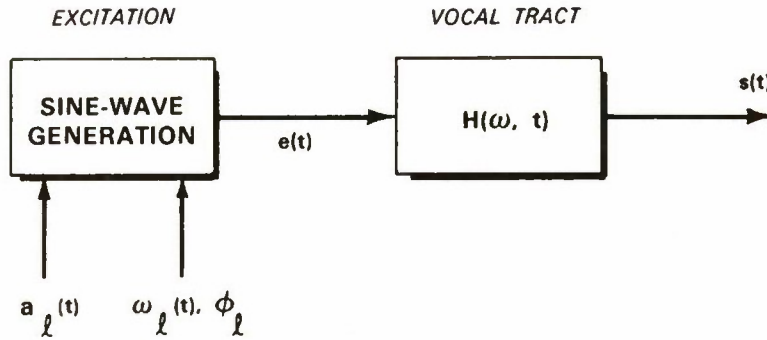


Figure 2-1. Sinusoidal representation of speech production.

This approach to modeling the excitation and vocal tract leads to a particularly simple representation of the speech waveform. Let the amplitude and phase of the system function along each frequency track $\omega_l(t)$ of the excitation be denoted by

$$M_l(t) = M[\omega_l(t), t] \quad (2.4a)$$

and

$$\Phi_l(t) = \Phi[\omega_l(t), t] \quad (2.4b)$$

Then using (2.1) through (2.4) results in the sinusoidal representation⁷

$$s(t) = \sum_{l=1}^{L(t)} A_l(t) \cos[\theta_l(t)] \quad (2.5a)$$

where

$$A_l(t) = a_l(t) M_l(t) \quad (2.5b)$$

and

$$\begin{aligned} \theta_l(t) &= \Omega_l(t) + \Phi_l(t) \\ &= V_l(t) + \phi_l + \Phi_l(t) \end{aligned} \quad (2.5c)$$

represent the amplitude and phase of each sine-wave component along the frequency track $\omega_l(t)$. The accuracy of this representation is subject to the caveat that the parameters are slowly varying relative to the duration of the vocal tract impulse response.

2.2 Analysis

The objective of the analysis is to estimate the model parameters (2.5) at an analysis frame rate sufficient to track articulatory changes, typically 5-20 ms. The procedure begins with estimating from a high-resolution spectral analysis the frequencies $\omega_l(t)$ and the composite

amplitudes $A_{\ell}(t)$ and phases $\theta_{\ell}(t)$ at the analysis frame rate. The second step in analysis separates the system and excitation components of (2.5b) and (2.5c).

Let $x(n)$ represent samples of the speech waveform, $w(n)$ the analysis window and R the analysis frame rate in samples, where the time sampling interval is assumed normalized to unity. Since the measured speech waveform is to be processed digitally, the sampled data notation is used primarily throughout this section. The frequencies of the glottal excitation $e(n)$ in (2.2) at time kR , associated with the k th analysis frame, are chosen to correspond to the $L(kR)$ largest peaks in the magnitude of the short-time Fourier transform, $|X(\omega, kR)|$, where

$$X(\omega, kR) = \sum_m w(kR - m)x(m)\exp(-jm\omega) \quad (2.6)$$

is the Fourier transform of the windowed speech segment $w(kR - n)x(n)$. The window sequence $w(n)$ is Hamming in shape and is nonzero over a range $0 \leq n < N$ corresponding to typically between 20 to 30 ms. In practice, the window duration is adaptive, being set to 2.5 times the speaker's measured average pitch. A minimum window width of 20 ms is used to guarantee adequate representation of unvoiced speech. Although ideally provision might be made for making the window duration a function of the instantaneous pitch, this refinement was found unnecessary for high-quality synthetic speech. The number of peaks $L(kR)$ is typically about 40 to 60 over a 4 kHz range. The maximum number of peaks that can be specified is limited by a threshold that is also a function of the measured average pitch. In particular, the maximum number of peaks was set at 60 peaks for a low-pitch speaker (60-100 Hz), 50 peaks for a medium-pitch speaker (100-200 Hz), and 40 peaks for a high-pitch speaker (200-300 Hz). In general, the performance was affected by the choice of this threshold only when too few peaks were allowed. The locations of the largest peaks were estimated by simply searching for a change of slope from positive to negative in the uniformly spaced samples of the short-time Fourier transform magnitude computed using the Discrete Fourier Transform (DFT). In practice, the DFT was evaluated using a 512-point Fast Fourier Transform (FFT) which gave adequate frequency resolution.

The amplitudes and phases (modulo 2π) of the component sine waves are given by the appropriate samples of the high-resolution DFT corresponding to $X(\omega, kR)$ at the chosen frequencies. Specifically, if ω_{ℓ}^k is the ℓ th frequency estimate on the k th analysis frame, i.e.,

$$\hat{\omega}_{\ell}^k = \omega_{\ell}(kR) \quad (2.7)$$

where “ $\hat{}$ ” denotes estimate, then the corresponding estimated amplitudes and phases, denoted by \hat{A}_{ℓ}^k and $\hat{\theta}_{\ell}^k$, respectively, are given as

$$\hat{A}_{\ell}^k = |X(\hat{\omega}_{\ell}^k, kR)| \quad (2.8a)$$

and

$$\hat{\theta}_{\ell}^k = \arg[X(\hat{\omega}_{\ell}^k, kR)] \quad (2.8b)$$

where “ \arg ” denotes principal phase value. In Reference 7 this estimator was shown to have certain robustness properties.

The next step in the analysis is to decompose the measured amplitudes \hat{A}_ℓ^k and phases $\hat{\theta}_\ell^k$ into vocal tract and excitation components whose amplitudes and phases are combined as in (2.5b) and (2.5c), respectively. The approach is to first obtain at each analysis frame, estimates of the vocal tract amplitude and phase as functions of frequency, i.e., $\hat{M}(\omega, kR)$ and $\hat{\Phi}(\omega, kR)$ (In practice, of course, only uniform samples from the DFT are available.). It will be shown below that the method of homomorphic deconvolution is particularly appropriate for obtaining these estimates in the context of the sine-wave representation. Assuming that the functions of frequency $\hat{M}(\omega, kR)$ and $\hat{\Phi}(\omega, kR)$ have been determined, then the system amplitude and phase estimates at the selected frequencies $\hat{\omega}_\ell^k$ are given by:

$$\hat{M}_\ell^k = \hat{M}(\hat{\omega}_\ell^k, kR) \quad (2.9a)$$

and

$$\hat{\Phi}_\ell^k = \hat{\Phi}(\hat{\omega}_\ell^k, kR) \quad (2.9b)$$

Finally, the excitation parameter estimates at each analysis frame boundary are obtained as

$$\hat{a}_\ell^k = \hat{A}_\ell^k / \hat{M}_\ell^k \quad (2.10a)$$

and

$$\hat{\Omega}_\ell^k = \hat{\theta}_\ell^k - \hat{\Phi}_\ell^k \quad (2.10b)$$

A block diagram of the primary steps in the analysis is depicted in Figure 2-2. The dotted lines stemming from the frequency estimator indicate that the frequency selection, amplitude division and phase subtraction in (2.9) and (2.10) are performed at only the estimated frequencies $\hat{\omega}_\ell^k$.

The remaining problem is to estimate $M(\omega, kR)$ and $\Phi(\omega, kR)$ as functions of frequency from the high resolution short-time Fourier transform $X(\omega, kR)$. There exist a number of established ways for separating out the system magnitude from the high-resolution spectrum, such as all-pole modeling¹⁹ and homomorphic deconvolution²⁰. The phase separation problem, on the other hand, is less well-understood and thus more difficult to solve²¹. However, if the vocal tract transfer function is assumed to be minimum phase then the logarithm of the system magnitude and the system phase form a Hilbert transform pair. With this constraint, a phase estimate $\hat{\Phi}(\omega, kR)$ can thus be derived from the logarithm of a magnitude estimate $\hat{M}(\omega, kR)$ through the Hilbert transform²⁰. Furthermore, the resulting phase estimate will be smooth and unwrapped as a function of frequency, a property that will be useful in performing speech synthesis. Although this minimum phase condition considerably simplifies the problem, the condition holds only approximately since the vocal tract transfer function may contain zeros outside the unit circle in the z-plane.

It follows that homomorphic deconvolution is particularly well-suited to the above estimation problem since an estimate of the system amplitude from the high-resolution spectrum and the computation of the Hilbert transform from this amplitude estimate can be performed simultaneously in this technique²⁰. The Fourier transform of the logarithm of the high-resolution magnitude is first computed to obtain the "cepstrum". In order to remove the effects due to

pitch, a right-sided window with duration proportional to the average pitch period, is then applied. The imaginary component of the inverse Fourier transform of the resulting sequence is the desired phase and the real part is the smooth log-magnitude. In practice, uniformly spaced samples of the Fourier transform are computed with the FFT whose length was chosen to be 512 points which was sufficiently large to avoid aliasing in the cepstrum. Thus, as illustrated in Figure 2-2, the high-resolution spectrum used to estimate the sine-wave frequencies is also used to estimate the vocal-tract system function.

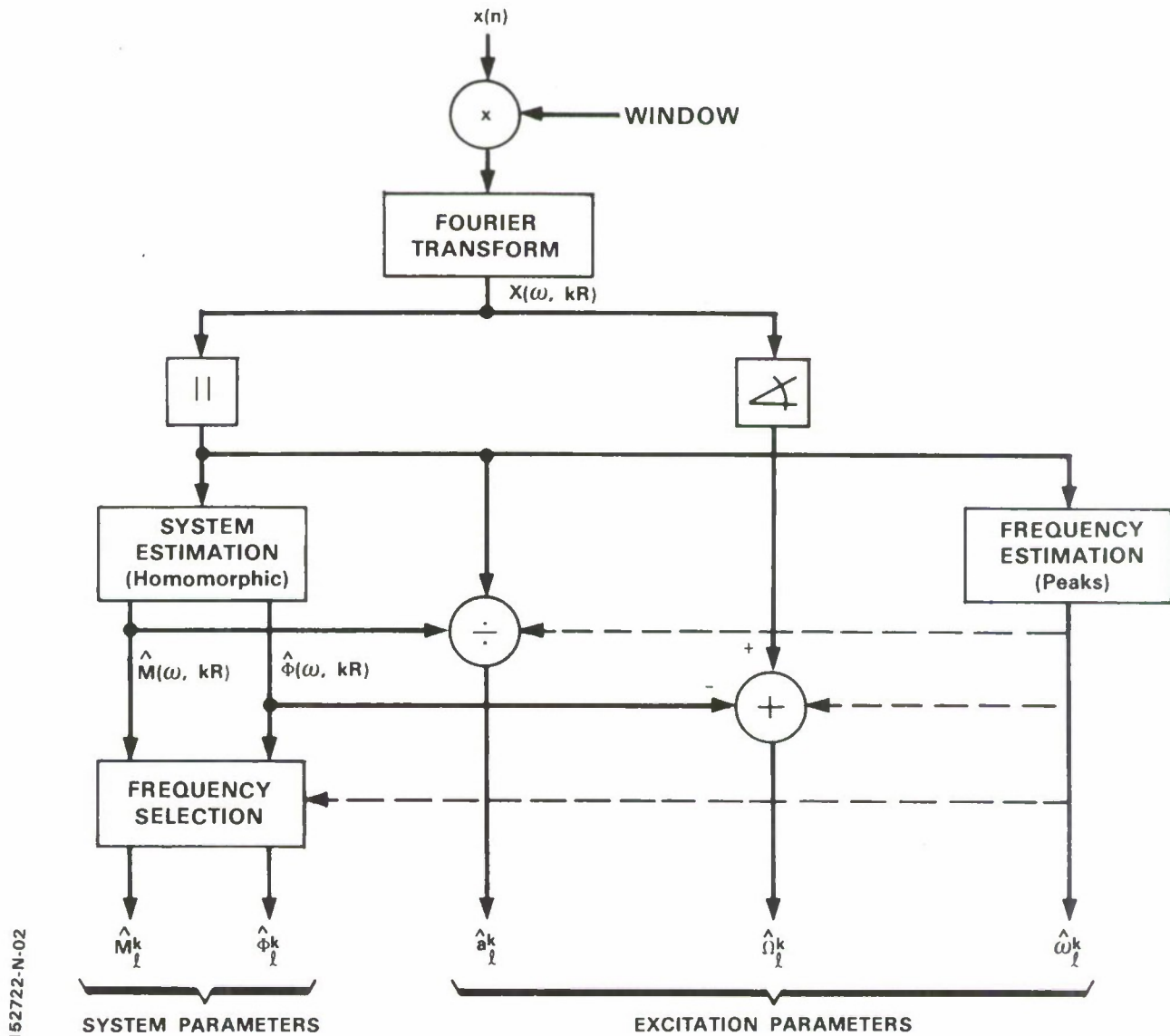


Figure 2-2. Block diagram of sinusoidal analysis.

2.3 Synthesis

In speech synthesis, the goal is to reconstruct an approximation that is as “close as possible” to the original speech. In the context of the sine-wave representation of the speech production mechanism, the synthesis first requires the matching of samples of the excitation and vocal tract contributions of each sine wave computed at consecutive frame boundaries. The matching procedure is followed by interpolation of the resulting pairs of amplitude and phase samples of the excitation and vocal tract functions and, lastly, the generation of sine waves based on the interpolated components.

The first step can be accomplished by associating the excitation frequencies measured on one frame with those obtained on a successive frame. Since the excitation and system amplitudes and phases are specified at the excitation frequencies, the matching of these parameters over consecutive frames follows directly. An algorithm for matching the location of spectral peaks was proposed for use in the synthesis procedure in Reference 7 which uses a purely sine-wave-based model (i.e., the excitation and system contributions of each sine-wave component are not explicitly represented). However, since the matching requirements are similar, the matching algorithm in Reference 7 can also be used here. The essence of the procedure is a nearest-neighbor association of frequencies. However, in practice, the location of spectral peaks, and thus of the frequencies, will change as the pitch changes and there will be rapid changes in both the location and number of peaks corresponding to rapidly-varying regions of speech, due to voiced/unvoiced transitions and to voiced fricatives for example. Since the nearest-neighbor association of frequencies is not sufficient to account for such rapid spectral movements, the frequency matching also incorporates a birth-death process of the component sine waves⁷. As a result of the matching algorithm, all of the amplitudes and phases of the excitation and system components measured for an arbitrary frame k at frequencies ω_ℓ^k are associated with a corresponding set of parameters for frame $k + 1$.

The next step in the synthesis is to interpolate the matched excitation and system parameters over a frame duration. The interpolation procedures are based on the assumption that the excitation and system functions are “slowly-varying” across each frame along frequency tracks $\omega_\ell(t)$. It follows that this slowly-varying constraint implies a slowly-varying excitation and system amplitude, and hence it suffices to interpolate samples of these functions linearly over a frame duration. Letting \hat{M}_ℓ^k and \hat{M}_ℓ^{k+1} denote a successive pair of system amplitude estimates for the ℓ th frequency track, then the system amplitude estimate across the k th frame is given by

$$\hat{M}_\ell(t) = \hat{M}_\ell^k + (\hat{M}_\ell^{k+1} - \hat{M}_\ell^k)t/T \quad (2.11a)$$

where T is the frame duration and $0 \leq t \leq T$ is the time into the k th frame. (Note that for simplicity the k dependence of $\hat{M}_\ell(t)$ has not been made explicit. Throughout the remainder of the report, the dependence of functional estimates on the frame number k is generally assumed.) Likewise, the excitation amplitude estimate $\hat{a}_\ell(t)$ over the k th frame is given by

$$\hat{a}_\ell(t) = \hat{a}_\ell^k + (\hat{a}_\ell^{k+1} - \hat{a}_\ell^k)t/T \quad (2.11b)$$

where \hat{a}_ℓ^k and \hat{a}_ℓ^{k+1} denote a pair of excitation amplitude estimates. Since the analysis and synthesis is performed digitally, in practice only samples of the continuous-time functions $\hat{M}_\ell(t)$ and $\hat{a}_\ell(t)$ are computed in obtaining a discrete-time realization of the synthetic waveform. Nevertheless, the continuous-time functional representation (2.11) is given here since it provides the key to performing time-scale modification in Section 3.

Since the vocal tract system is assumed slowly-varying over consecutive frames, it is reasonable to assume that its phase is slowly-varying as well and thus linear interpolation of the system phase samples will also suffice. However, the characteristic of "slowly-varying" is more difficult to achieve for the system phase than for the system magnitude. This is because an additional constraint must be imposed on the measured phase; namely that the phase be smooth and unwrapped as a function of frequency at each frame boundary. This requirement is illustrated in Figure 2-3. There it is shown that if the system phase is obtained modulo 2π , then linear interpolation can result in a falsely rapidly-varying system phase between frame boundaries.

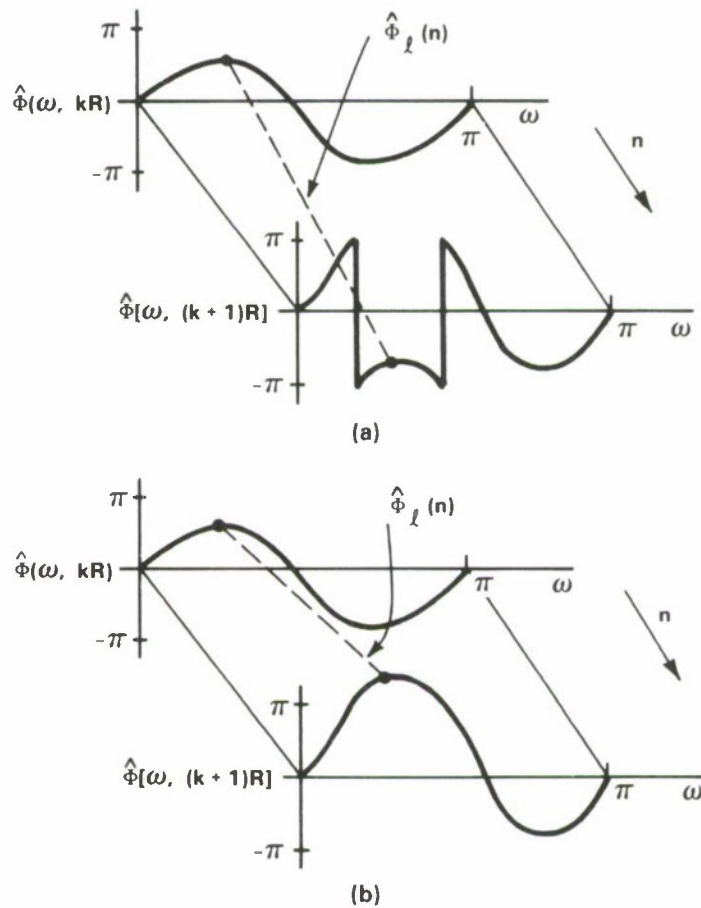


Figure 2-3. Requirement of unwrapped system phase. (a) Interpolation of wrapped system phase. (b) Interpolation of unwrapped system phase.

The importance of the use of a homomorphic analyzer in the previous section is now evident. The system phase estimate $\hat{\Phi}_\ell(\omega, kR)$ derived from the homomorphic analysis is unwrapped in frequency and thus is slowly-varying when the system amplitude (from which it was derived) is slowly-varying. Linear interpolation of samples of this function then results in a phase trajectory which reflects the underlying vocal tract movement. The interpolation scheme is similar to that in (2.11), and is given by

$$\hat{\Phi}_\ell(t) = \hat{\Phi}_\ell^k + (\hat{\Phi}_\ell^{k+1} - \hat{\Phi}_\ell^k)t/T \quad (2.12)$$

where as before $0 \leq t \leq T$ is the time into the k th frame and where in a discrete-time realization only samples of (2.12) are required.

Unfortunately, such a simple approach cannot be used to interpolate the phase and frequency of the excitation. Since the phase of $X(\omega, kR)$ in (2.8b) is measured modulo 2π , then the excitation phase $\hat{\Omega}_\ell^k$ in (2.10b) may contain 2π discontinuities. Thus in interpolating the excitation phase, phase unwrapping must be performed. In addition, since the excitation phase is the integral of the instantaneous frequency as seen in (2.2b), the interpolation must yield a phase which is consistent with the frequencies measured at each frame boundary. This problem, which was originally addressed in Reference 7, was solved by using a cubic polynomial for the interpolation function, namely

$$\hat{\Omega}_\ell(t) = a + bt + ct^2 + dt^3 \quad (2.13)$$

with $t = 0$ corresponding to frame k and $t = T$ corresponding to frame $k + 1$. Thus as before t represents the time into the k th frame. Since the instantaneous frequency is the derivative of the phase, then

$$\hat{\omega}_\ell(t) = \dot{\hat{\Omega}}_\ell(t) = b + 2ct + 3dt^2 \quad (2.14)$$

The solution requires constraining the cubic function (2.13) and its derivative (2.14) to equal the excitation phase and frequencies measured at the frame boundaries. The notion of applying a cubic polynomial to interpolate the excitation phase between frame boundaries was independently proposed by Almeida and Silva¹⁸ for use in their harmonic sine-wave synthesizer. However, since only the principal value of the phase can be measured, provision must be made for unwrapping the phase subject to the above constraints on the cubic phase interpolation function. This leads to invoking an additional constraint which requires that the unwrapped phase be maximally smooth. The criterion of "smoothness" is defined as the minimization of the second derivative of $\hat{\Omega}_\ell(t)$ over the analysis frame duration. This approach is similar to the method for interpolating samples of the estimated composite phase $\hat{\theta}_\ell(t)$ that was developed in Reference 7. The resulting phase function not only satisfies all the endpoint constraints, but also resolves any 2π phase ambiguities, thus unwrapping the excitation phase in time along each frequency track. Note that since the excitation frequency equals the phase derivative, a quadratic frequency trajectory can be computed directly from this procedure.

Since the above analysis began with the assumption of initial estimates of the excitation and the system amplitudes and phases corresponding to the start of frame k , it is necessary to specify the initialization of the frame interpolation procedure. At the birth of a track, since a matched

peak was not found at frame k , the frequency between frame k and frame $k + 1$ is assumed fixed at the measured frequency $\hat{\omega}_\ell^{k+1}$ [7]. The system amplitude and phase are simply set to the values measured on frame k at the frequencies $\hat{\omega}_\ell^{k+1}$, i.e., $\hat{M}(\hat{\omega}_\ell^{k+1}, kR)$ and $\hat{\Phi}(\hat{\omega}_\ell^{k+1}, kR)$. On the other hand, the excitation amplitude is set to zero to avoid unnatural discontinuities, since upon the birth of a track the measured excitation function at $\hat{\omega}_\ell^{k+1}$ on frame k may be arbitrary. To initiate the interpolation of the excitation phase at the birth of a track, the phase $\hat{\Omega}_\ell^{k+1}$ is defined to be the measured phase and the startup phase on frame k is defined to be

$$\hat{\Omega}_\ell^k = \hat{\Omega}_\ell^{k+1} - \hat{\omega}_\ell^{k+1}R \quad (2.15)$$

where R is the number of samples traversed in going from frame $k + 1$ back to frame k (a discrete-time realization is assumed here) and where the frequency over frame $k + 1$ is assumed fixed at $\hat{\omega}_\ell^{k+1}$. This procedure insures that the phase interpolation constraints are satisfied initially. Note also that $\hat{\Omega}_\ell^k$ provides an estimate of the phase offset ϕ_ℓ of (2.2b). Likewise the startup time kR gives an estimate of the sine-wave onset time t_ℓ of the ℓ th sine wave. This estimate of the onset time always falls on a frame boundary.

It was noted earlier in presenting the sine-wave model that the excitation phase consists of two components, a constant term and a time-varying term as given in (2.2) where the time t is continuously running. Similarly, the estimate of the excitation phase over the k th frame can be written in terms of a constant and a time-varying term. Although this decomposition of the excitation phase is not necessary for signal synthesis, it is developed here since it will be used in later sections on speech modifications. Specifically, the excitation phase over the k th frame is written as

$$\begin{aligned} \hat{\Omega}_\ell(t) &= \int_{t_\ell}^t \hat{\omega}_\ell(\sigma) d\sigma + \hat{\phi}_\ell \\ &= \int_0^t \hat{\omega}_\ell(\sigma) d\sigma + \int_{t_\ell}^0 \hat{\omega}_\ell(\sigma) d\sigma + \hat{\phi}_\ell \end{aligned} \quad (2.16)$$

where t falls in the range $[0, T]$ which defines the k th frame. Letting Σ_ℓ^k denote the phase due to the time-varying frequency accumulated up to frame k , i.e.,

$$\begin{aligned} \Sigma_\ell^k &= \int_{t_\ell}^0 \omega_\ell(\sigma) d\sigma \\ &= \Sigma_\ell^{k-1} + \int_{-T}^0 \omega_\ell(\sigma) d\sigma \end{aligned} \quad (2.17a)$$

and if $\hat{V}_\ell(t)$ denotes the phase due to the time-varying frequency accumulated over frame k , i.e.,

$$\hat{V}_\ell(t) = \int_0^t \hat{\omega}_\ell(\sigma) d\sigma \quad (2.17b)$$

then the excitation phase in (2.16) can be written as

$$\hat{\Omega}_\ell(t) = \hat{V}_\ell(t) + \sum_\ell^k + \hat{\phi}_\ell \quad (2.18)$$

The time-varying component $\hat{V}_\ell(t)$ in (2.18) is obtained by integrating the instantaneous frequency, the result of which is given by the time-varying component of the cubic (2.13). The constant component consists of the phase offset estimate $\hat{\phi}_\ell$ and the accumulated phase component \sum_ℓ^k which from (2.17a) and (2.17b) can be obtained recursively as

$$\sum_\ell^{k+1} = \sum_\ell^k + \hat{V}_\ell(T) \quad (2.19)$$

The constant component is also the phase at the right-hand boundary of the previous frame and is given by the value of the parameter a in (2.13). In the case where the track is just initiated, the accumulated phase $\sum_\ell^k = 0$ and the constant $a = \hat{\phi}_\ell = \hat{\Omega}_\ell^k$ of (2.15).

The final synthetic waveform (in discrete time) is given by

$$\hat{s}(n) = \sum_{\ell=1}^{L(n)} \hat{A}_\ell(n) \cos[\hat{\theta}_\ell(n)] \quad (2.20a)$$

where

$$\hat{A}_\ell(n) = \hat{a}_\ell(n) \hat{M}_\ell(n) \quad (2.20b)$$

and

$$\hat{\theta}_\ell(n) = \hat{\Omega}_\ell(n) + \hat{\Phi}_\ell(n) \quad (2.20c)$$

where $L(n)$ is the number of sine waves estimated at time n and where, since the functional estimates in (2.20) were derived above on a frame-by-frame basis, the index $n = 0, 1, 2, \dots, R - 1$ is interpreted as the discrete time into the k th frame. The functions $\hat{a}_\ell(n)$, $\hat{M}_\ell(n)$, $\hat{\Omega}_\ell(n)$, and $\hat{\Phi}_\ell(n)$ come from samples of the continuous-time functions in (2.11) through (2.13). A block diagram of the overall synthesis structure is illustrated in Figure 2-4. The dotted lines stemming from the frequency matcher indicate that the matched frequencies are required by the linear and cubic interpolation procedures. Note that the functional contributions in (2.20) are estimated using only two consecutive frames. Thus in a computer implementation of (2.20) the speech waveform can be processed in block fashion, requiring the storage of only one high-resolution short-time Fourier transform and two sets (corresponding to two consecutive frames) of system and excitation parameters in (2.9) and (2.10).

In order to evaluate the performance of this new approach to analysis and synthesis of speech, a non-real-time floating point computer simulation of this system was developed. The speech processed in the simulation was low-passed filtered at 5 kHz, digitated at 10 kHz, analyzed at a 5 ms frame rate and synthesized over a 4 kHz range. A 10 ms analysis frame, however, was also found adequate for reconstruction. Informal listening demonstrated that for both male and female speakers, the synthetic speech was nearly perceptually indistinguishable

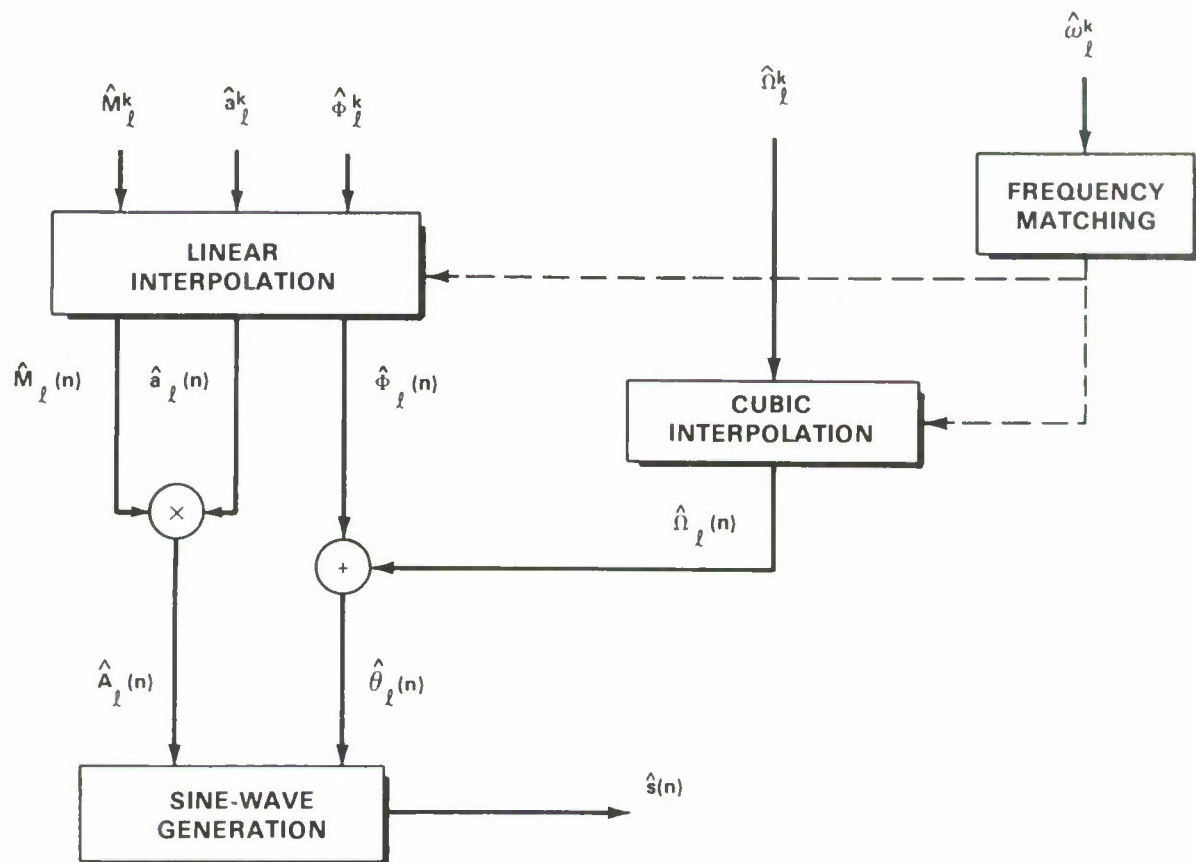
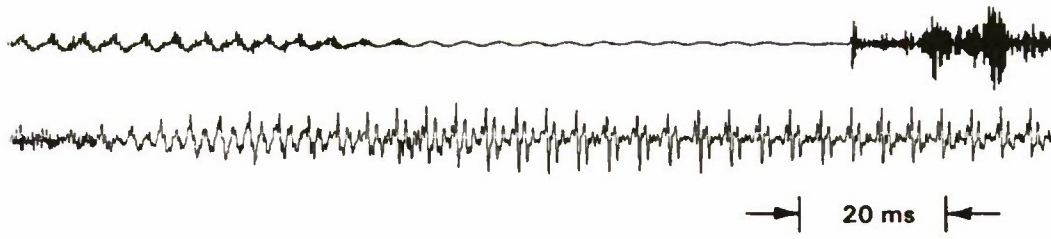


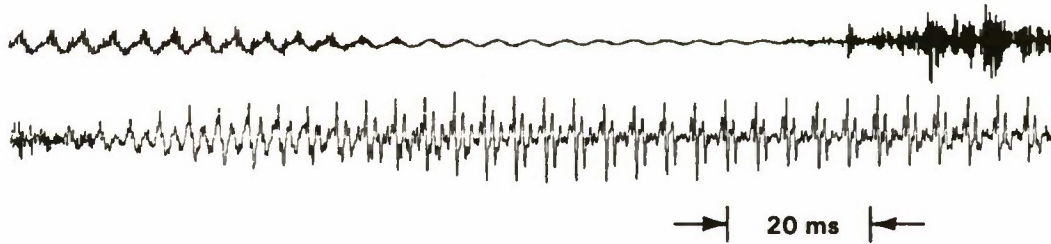
Figure 2-4. Block diagram of sinusoidal synthesis.

from the original. As illustrated in Figure 2-5, the original waveform structure is also essentially preserved in the reconstruction. In this example, representing a segment from a male speaker, the duration of the analysis window was 25 ms and a threshold of 50 peaks was specified.

Although the system was originally designed for single-speaker signals, the reconstruction does not break down for multiple speakers nor for nonspeech sounds such as music and marine biologic sounds. Figure 2-6 for example depicts the reconstruction of a waveform consisting of the sum of two speech signals, one from a male and one from a female speaker. Furthermore, in the presence of acoustic background noise (down to 0 dB S/N) and other interferences such as music, the speech and interference are virtually perceptually identical to the originals. A segment of the reconstruction for the case of a female speaker in a music background is illustrated in Figure 2-7. In these different cases, where appropriate, the number of peaks is assumed equal to the sum of the required number of peaks for each signal type. For example, for the female speaker in music, the total number of peaks was set at about 80 over a 4 kHz range, 40 peaks for each signal type.

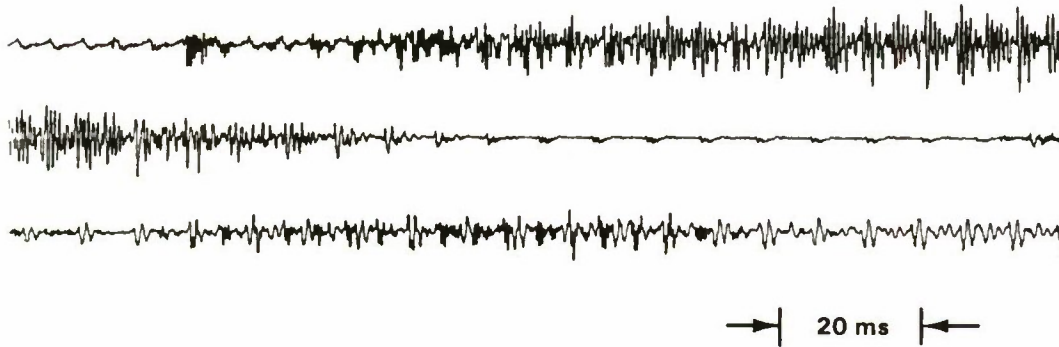


(a)

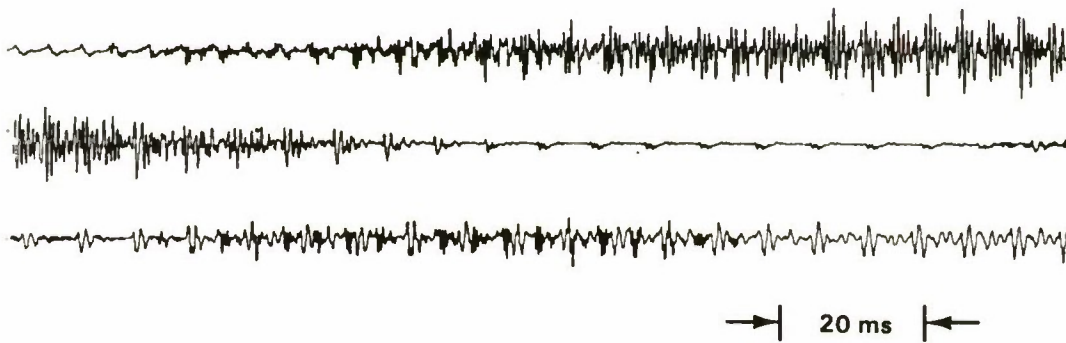


(b)

Figure 2-5. Reconstruction of signal from single speaker. (a) Original. (b) Reconstruction.



(a)



(b)

Figure 2-6. Reconstruction of signal from two speakers. (a) Original. (b) Reconstruction.

147343-R-01

149385-R-01

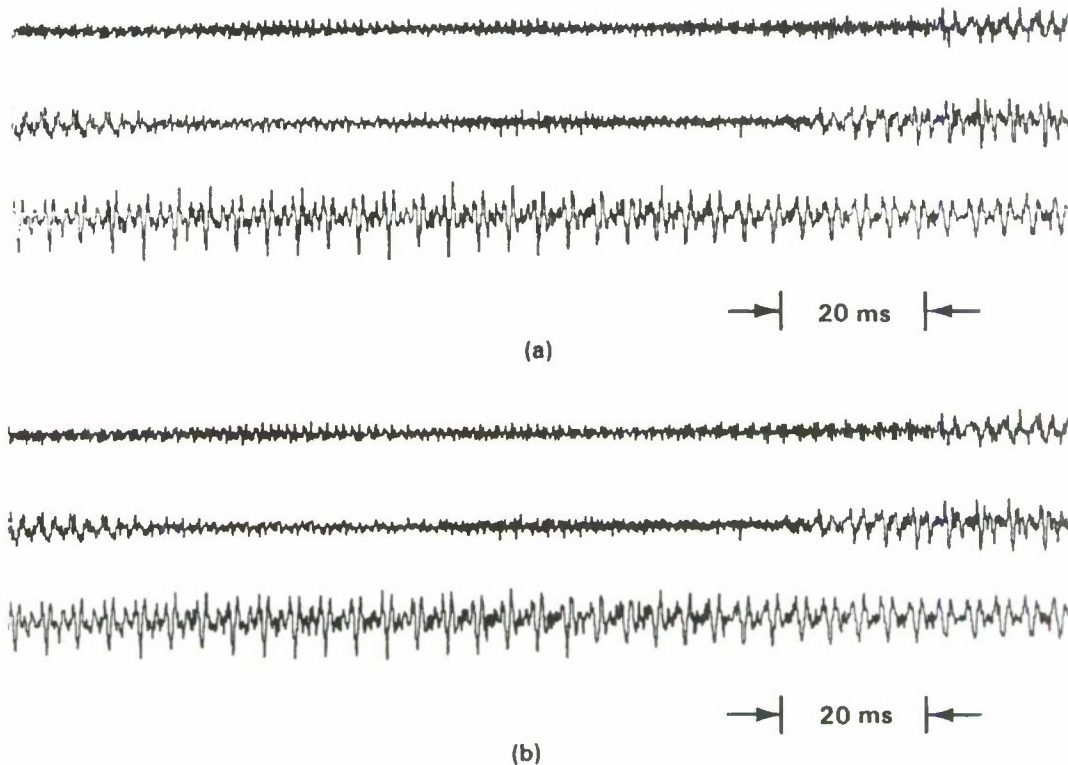


Figure 2-7. Reconstruction of speech in a music background. (a) Original. (b) Reconstruction.

2.4 Application to Speech Modification

Since the analysis/synthesis procedure has been expressed in terms of a functional model describing the behavior of each sine wave component, it is now possible to explore speech modifications simply by transforming each of the functional descriptors. In performing these modifications, the excitation and vocal tract amplitude and phase of each of the sine-wave components will be manipulated in different ways. For example, in time-scale modification the frequency trajectories of the excitation sine waves will be stretched or compressed in time, while the vocal tract components will be made to move faster or slower. In pitch modification, the spacing between the excitation frequency trajectories (which defines pitch) is made smaller or larger, while preserving the vocal tract spectral characteristics. The first of these transformations to be developed will be for time-scale modification.

3. TIME-SCALE MODIFICATION

The goal of time-scale modification is to maintain the perceptual quality of the original speech while changing the apparent rate of articulation. This requires that the pitch contour, and thus the frequency trajectories of the excitation, be stretched or compressed in time, and that the vocal tract, and thus the amplitude and phase of the vocal tract transfer function, be changed at a slower or faster rate than the rate of normally spoken speech. Thus both the pitch trajectory and the spectral characteristics of the speaker are essentially preserved. The synthesis method of the previous section is ideally suited for this transformation since it involves summing sine waves composed of the excitation and vocal tract system contributions for which explicit functional expressions have been derived.

In this section, a method is first presented for performing a fixed rate change based on the analysis and synthesis system of Section 2. The method is motivated by a sine-wave model for time-scale modification of speech. With the fixed rate-change case as a stepping stone, a similar development follows for time-varying rate change where the time scale is continuously adjusted. As well as providing additional flexibility, an adjustable time scale can lead to a more natural change in the rate of articulation (than achieved with a fixed time-scale modification) by allowing the time scale to adapt to various features of the speech waveform.

3.1 Fixed Rate Change

For an arbitrary time-scale transformation, the time t_0 corresponding to the original articulation rate is mapped to the transformed time t'_0 through the mapping

$$t'_0 = W(t_0) \tag{3.1}$$

For a fixed rate change ρ , the mapping (3.1) is reduced to the linear relation $W(t_0) = \rho t_0$. The case $\rho > 1$ corresponds to slowing down the rate of articulation by means of a time-scale expansion, while the case $\rho < 1$ corresponds to speeding up the rate of articulation by means of a time-scale compression. Speech "events" which take place at a time t'_0 according to the new time scale will have occurred at $\rho^{-1}t'_0$ in the original time scale. The time-scale transformation $W(\)$ is illustrated in Figure 3-1 for the case $\rho > 1$ where the time scale is expanded. The time scales of Figure 3-1 can be thought of as representing two simultaneous time counters, one running with respect to the original articulation rate and the other with respect to the transformed rate.

In the sine-wave model for time-scale modification, the "events" which are time-scaled are the system amplitudes and phases, $M(\omega, t)$ and $\Phi(\omega, t)$, and the excitation amplitudes and frequencies, $a_q(t)$ and $\omega_q(t)$, of each underlying sine wave. The system parameters are manipulated such that the vocal tract articulators move faster or slower in time. The excitation parameters are modified so that frequency trajectories are stretched or compressed while maintaining pitch. The model for time-scale modified speech is illustrated schematically in Figure 3-2 and represents a simple modification of the input/output model for unmodified speech that was depicted in

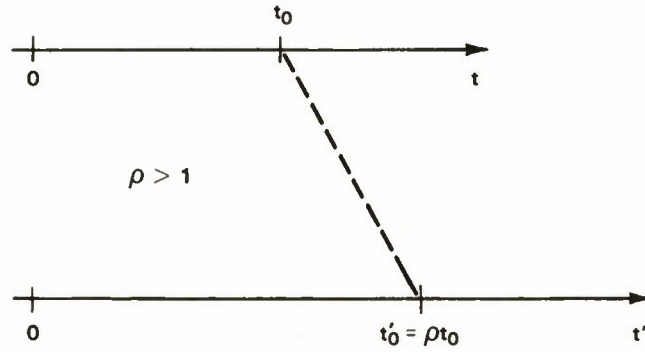


Figure 3-1. Time warping with fixed rate change $\rho > 1$.

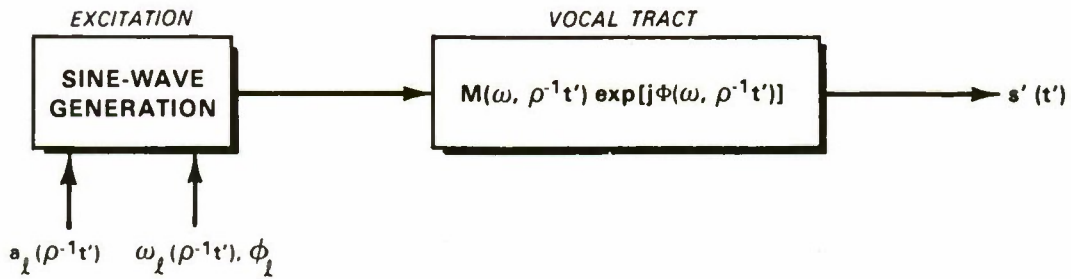


Figure 3-2. Model of time-scale modification.

Figure 2-1. Based on the mathematical sine-wave model for speech production in (2.5), Figure 3-2 is easily transformed to a mathematical model for time-scale modified speech, denoted by $s'(t)$, and is given by

$$s'(t) = \sum_{\ell=1}^{L(t)} A'_{\ell}(t) \cos [\theta'_{\ell}(t)] \quad (3.2a)$$

where

$$A'_{\ell}(t) = A_{\ell}(\rho^{-1}t) = a_{\ell}(\rho^{-1}t) M_{\ell}(\rho^{-1}t) \quad (3.2b)$$

and

$$\theta'_{\ell}(t) = \Omega'_{\ell}(t) + \Phi_{\ell}(\rho^{-1}t) \quad (3.2c)$$

with

$$\Omega'_{\ell}(t) = \int_{t_{\ell}}^{t'} \omega_{\ell}(\rho^{-1}\tau) d\tau + \phi_{\ell} \quad (3.2d)$$

where the system functions $M_\ell(t)$ and $\Phi_\ell(t)$ were defined in Section 2.4 along the frequency trajectories $\omega_\ell(t)$, and where the time scale in (3.2) corresponds to the transformed time scale of Figure 3-1. The onset time of each excitation sine wave, t_ℓ , in the original time scale is mapped to the new onset time, $t'_\ell = \rho t_\ell$ where the excitation phase takes on the initial value ϕ_ℓ . Note that with a change of variables $\sigma = \rho^{-1}\tau$, (3.2d) can be written as

$$\begin{aligned} \Omega'_\ell(t') &= \int_{t'_\ell}^{\rho^{-1}t'} \omega_\ell(\sigma) d\sigma / \rho^{-1} + \phi_\ell \\ &= V_\ell(\rho^{-1}t') / \rho^{-1} + \phi_\ell \end{aligned} \quad (3.2e)$$

where $V_\ell(t)$ is the time-varying contribution to the excitation phase given in (2.2b) and (2.2c). The initial phase offset in (3.2d) is consistent with preserving the pulse-like nature of the excitation function during voicing. To see how the excitation function is preserved, consider an excitation function given by a periodic pulse train where the first pulse begins at time = t_0 . This excitation can be represented by a sum of sinusoidal components $\cos[\ell\omega_0(t - t_0)]u(t - t_0)$ where $u(t)$ is the unit step function. Clearly, in this case, $t_\ell = t_0$ and $\phi_\ell = 0$ for all ℓ and thus the choice of the phase offset ϕ_ℓ results in the first pulse occurring at the time $t'_\ell = \rho t_\ell$ in the transformed time scale. Figure 3-3 illustrates how a single sine-wave burst with initial phase offset $\phi_\ell = 0$ changes as it is time-scale modified. When all sine waves begin at the same time t_ℓ but before the

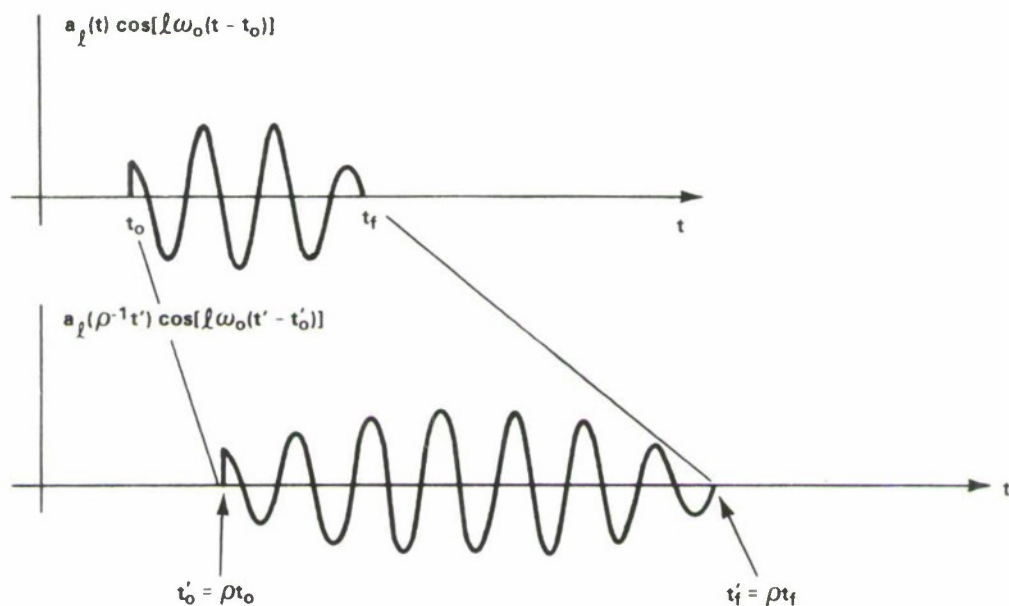


Figure 3-3. Time-scale expansion of a single sine wave.

first pulse which occurs at time t_0 , then the first pulse in the transformed time scale will occur at a distance $t'_0 - t'_l$ from t'_l when the phase offset is set as in (3.2d). In practice, however, a periodic excitation pulse train is not impulse-like and the sine-wave onset times may be different. When sine waves begin at different times t_l , then initializing the excitation phase with ϕ_l in the transformed time scale may introduce phase dispersion in the periodic pulse train. This dispersion, without knowledge of the first pulse position, may be difficult to avoid. (Note that in the experimental system other phase offset models, e.g., ϕ_l/ρ^{-1} , were found to result in a pulse-like excitation during voicing).

The modifications of (3.2) correspond to stretching (or compressing) the frequency tracks of the excitation (and thus the pitch contour), and the slower (or faster) movement of the vocal tract articulators. As in the model for speech production without modification, the system magnitude and phase need be specified only along the frequency trajectories of the excitation function. The functional mappings along these trajectories are illustrated in Figure 3-4 for time-scale expansion over a time duration $0 \leq t \leq t_f$. Both the excitation frequency and phase trajectories are depicted to illustrate the preservation of pitch and the special modification of the excitation phase function.

With this time-scale model as a basis, it is straightforward to construct a time-scale modification system for fixed rate change using the analysis/synthesis structure of Section 2. The estimates (2.11) through (2.18), obtained in the synthesis stage, provide functional forms for the parameters in (3.2). Since these functional estimates were derived on a frame-by-frame basis, it is natural to view the inverted time $\rho^{-1}t'$ as the *time into* the k th frame within the original time scale. The estimates of the time-scaled parameters (3.2) can then be obtained by evaluating the functional estimates at the time ($\rho^{-1}t'$ modulo T) where T is the original frame duration. In a discrete-time implementation, the inverted time is given by ($\rho^{-1}n'$ modulo R) where R is the number of samples in the original frame duration, where n' is the discrete- (transformed) time index and where the sampling interval is assumed unity. It follows that the time-scaled synthetic waveform (in discrete time) can be obtained over the k th frame by replacing the model parameters of (3.2) by their estimates:

$$\hat{s}'(n') = \sum_{\ell=1}^{L(n')} \hat{A}'_{\ell}(n') \cos[\hat{\Omega}'_{\ell}(n') + \hat{\Phi}'_{\ell}(n')] \quad (3.3a)$$

where

$$\hat{A}'_{\ell}(n') = \hat{A}_{\ell}[(\rho^{-1}n')_R] \quad (3.3b)$$

and

$$\hat{\Phi}'_{\ell}(n') = \hat{\Phi}_{\ell}[(\rho^{-1}n')_R] \quad (3.3c)$$

and

$$\hat{\Omega}'_{\ell}(n') = \hat{V}_{\ell}[(\rho^{-1}n')_R]/\rho^{-1} + (\Sigma_{\ell}^k)' + \hat{\phi}_{\ell} \quad (3.3d)$$

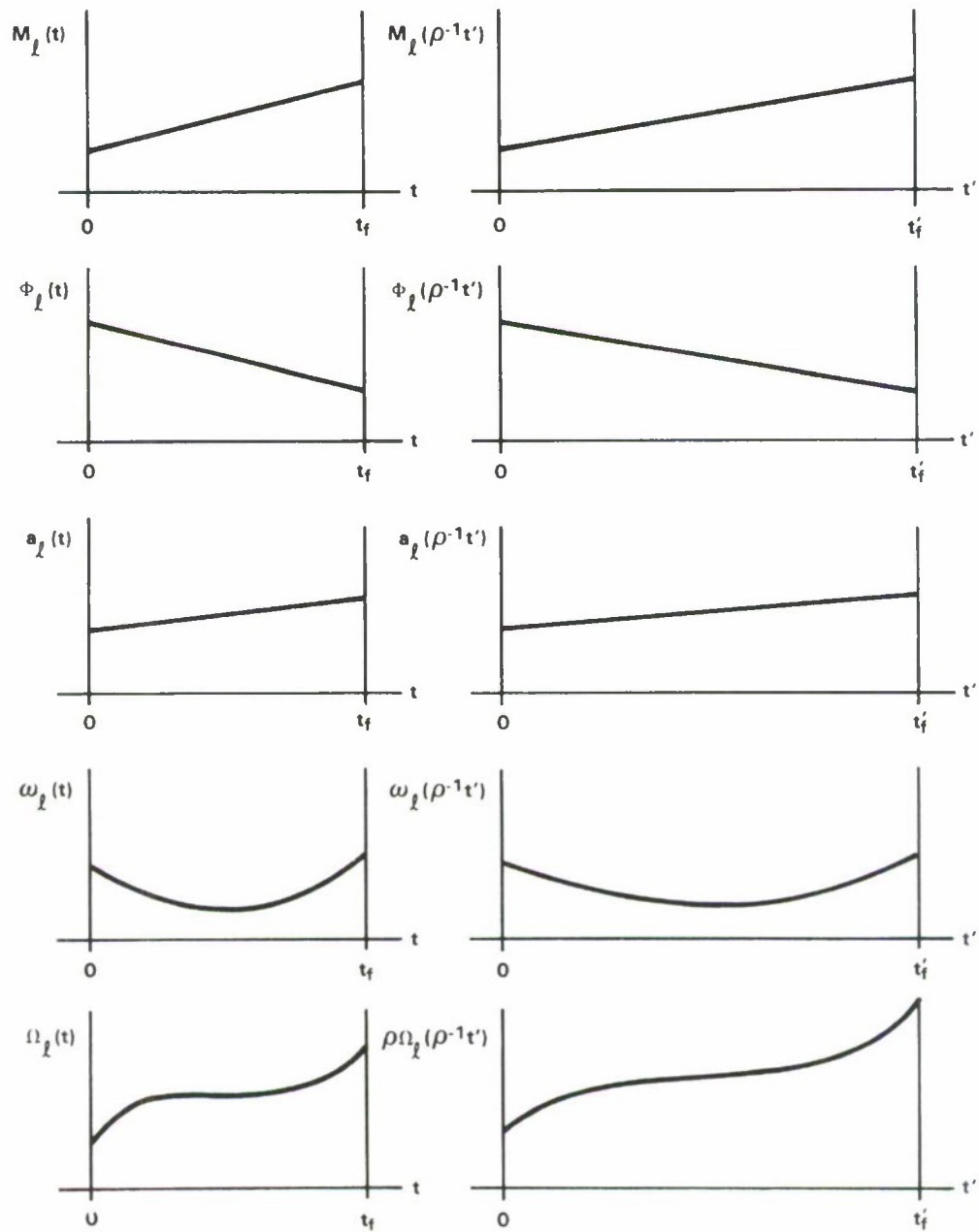


Figure 3-4. Functional mappings for time-scale expansion.

with $(\Sigma_{\ell}^k)'$ computed recursively as

$$(\Sigma_{\ell}^{k+1})' = (\Sigma_{\ell}^k)' + \hat{V}_{\ell}(R)/\rho^{-1} \quad (3.3e)$$

where $(*)_R$ denotes *modulo* R. Recall from (2.17) that $\hat{V}_{\ell}(t)$ represents the time-varying phase component which is added to the excitation phase obtained up to frame k. The accumulated phase due to the time-varying phase is denoted by Σ_{ℓ}^k and was obtained by adding the values of the time-varying phase at frame boundaries, i.e., $\hat{V}_{\ell}(R)$. In (3.3d) and (3.3e), $(\Sigma_{\ell}^k)'$ denotes this same accumulated phase function but now scaled by ρ^{-1} . The recursive computation of $(\Sigma_{\ell}^k)'$ in (3.3e) is initialized at zero. The other parameter values in (3.3) are obtained by sampling the estimates $\hat{M}_{\ell}(t)$, $\hat{a}_{\ell}(t)$, $\hat{V}_{\ell}(t)$, and $\hat{\Phi}_{\ell}(t)$ of (2.11) through (2.15) at time $t_{n'} = (\rho^{-1}n)_R$ over each frame. Since the parameter estimates of the unmodified synthesis are available as continuous functions of time, in theory, any rate change is possible. However, arbitrarily large or small rate changes may not be meaningful from a speech production or perceptual viewpoint. A block diagram of the synthesis component of the time-scale modification procedure represented by (3.3) is shown in Figure 3-5 where the *modulo* R notation has been eliminated. The analysis component is identical to that depicted in Figure 2.2.

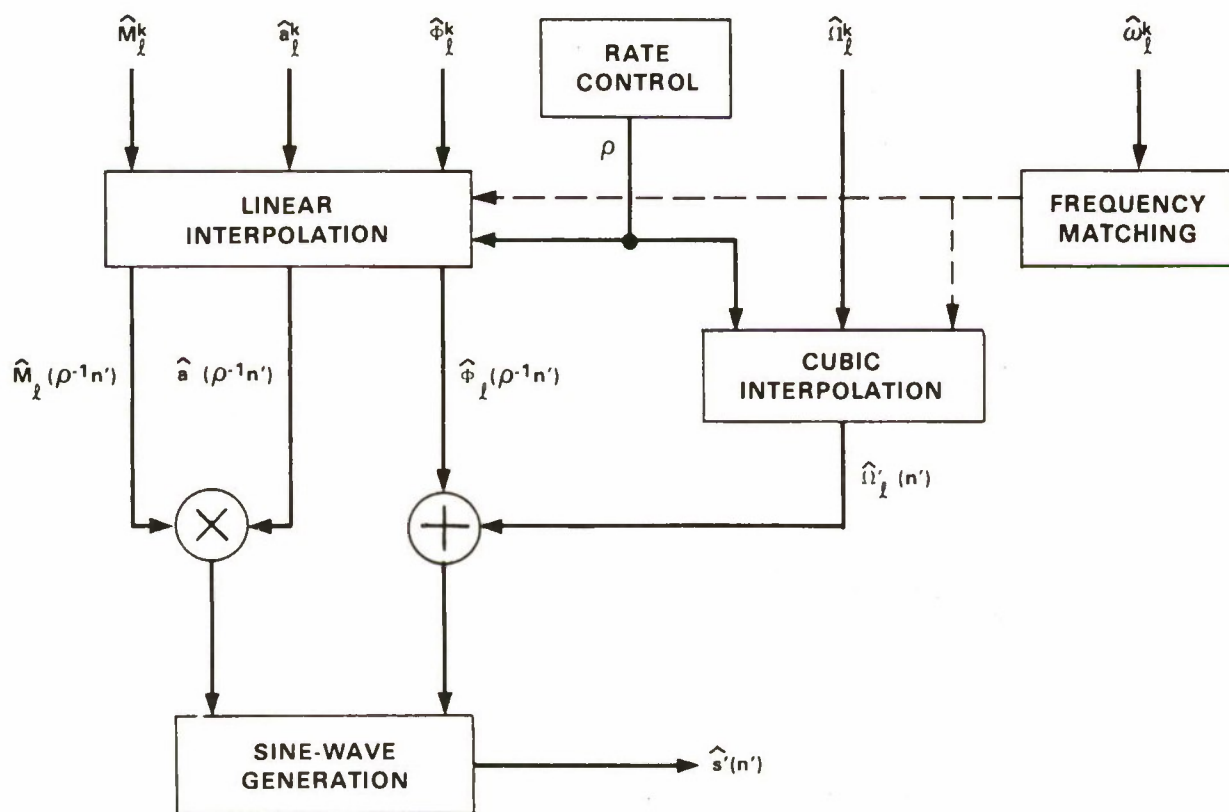


Figure 3-5. Block diagram of uniform rate-change system.

In a computer implementation of (3.3), as with the analysis/synthesis system of Section 2, the processing can be performed in block fashion. This is a natural computational strategy since each of the underlying functional estimates of (3.3) were derived using only two consecutive frames. The samples of the synthetic modified waveform are generated in segments by mapping back the time index n' through one analysis frame to the time $t_{n'} = (\rho^{-1}n')_R$. In particular, suppose that the k th frame has been entered. Then the amplitude and phase functional estimates are first derived using (2.11) through (2.18). The discrete time index n' (initiated at zero) is updated until the boundary of an analysis frame is crossed, i.e., $t_{n'} = (\rho^{-1}n')_R$ "wraps" back on itself. Until this time, the set of functional estimates is sampled at the inverse time $t_{n'}$ and the modified waveform is synthesized as in (3.3). When a new frame is entered, a new set of functional estimates (2.11) through (2.16) are sampled according to the inverse time mapping $t_{n'} = (\rho^{-1}n')_R$. When a new frame is entered, the accumulated time-varying excitation phase $(\Sigma_0^k)'$ is updated according to (3.3e) where at the onset of a new sine wave $(\Sigma_0^k)' = 0$, Figure 3-6 gives a flow diagram of the entire process. This block-oriented implementation has the advantage that storage of only two consecutive sets of amplitude and phase parameters is required.

To test the reconstruction procedure of (3.3), time-scale modifications ranging from a compression of two to an expansion of three were implemented with the frame-based method described above and outlined in Figure 3-6. The example in Figure 3-7 illustrates the response of the system to a synthetic waveform formed by convolving a periodic pulse train (with a 100 Hz fundamental) with an exponentially decaying response. Here a 5 ms frame interval, a 25 ms analysis window, and 40 peaks over a 4 kHz range were used. The rate-change factors are 1.0, 1.5, and 0.5. In processing speech, the parameters such as the number of peaks, sampling rates, window length, etc., were set equal to those used in the system for unmodified speech. Generally, the rate-changed synthetic speech was of high quality and free of artifacts such as glitches and reverberation. Furthermore, the natural quality and smoothness of the original speech were preserved through transitions such as voiced/unvoiced boundaries. Examples of the synthesis are illustrated in Figures 3-8 and 3-9 for the case of a single speaker. Figure 3-8 depicts an example of time-scale expansion by a factor of two during an unvoiced/voiced transition; while Figure 3-9 illustrates time-scale compression by a factor of two.

Although the original system was designed for a single speaker, as with the baseline system, the time-scale modification system was also found to perform successfully for nonspeech sounds and speech with various types of interference. This includes music, sounds emitted by whales, multiple speakers, speech in acoustic background noise (down to about 0 dB S/N) and speech with a musical background. Examples of time-scale modification of speech in music and of a waveform consisting of speech from overlapping male and female speakers are depicted in Figures 3-10 and 3-11, respectively. The modified waveforms were natural sounding and without artifacts. The characteristics of F15 cockpit noise, for example, are hardly altered in expansion and compression by a factor of two. When applying the system to music, the time-scaled music seems to emanate from instruments played at a faster or slower rate than normal. When reconstructing two simultaneous speakers, the modified speech seems to have been generated by "linear analysis/synthesis" (i.e., the sum of the modified waveforms equals the modification of the sum).

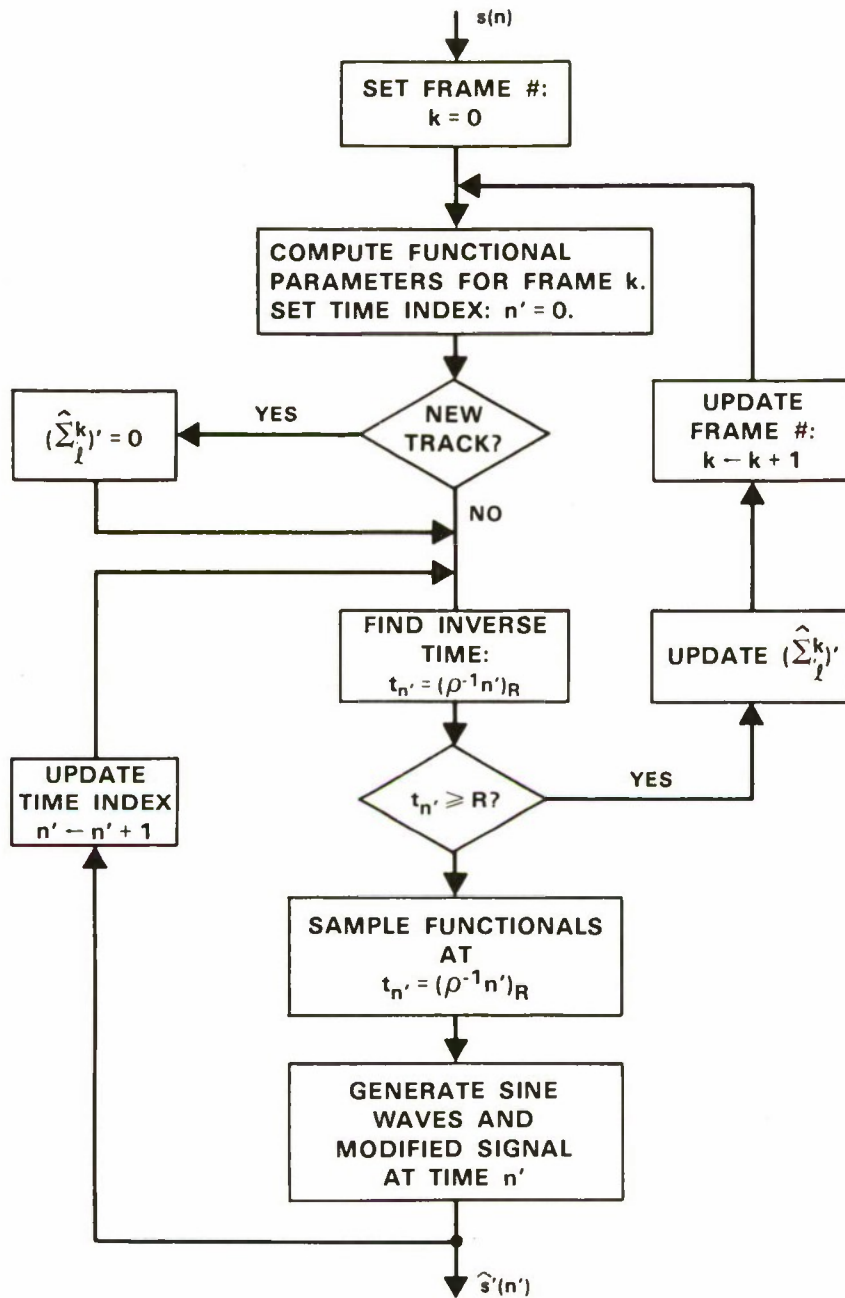


Figure 3-6. Flow diagram of computer implementation of uniform rate-change system.

157912-N-01

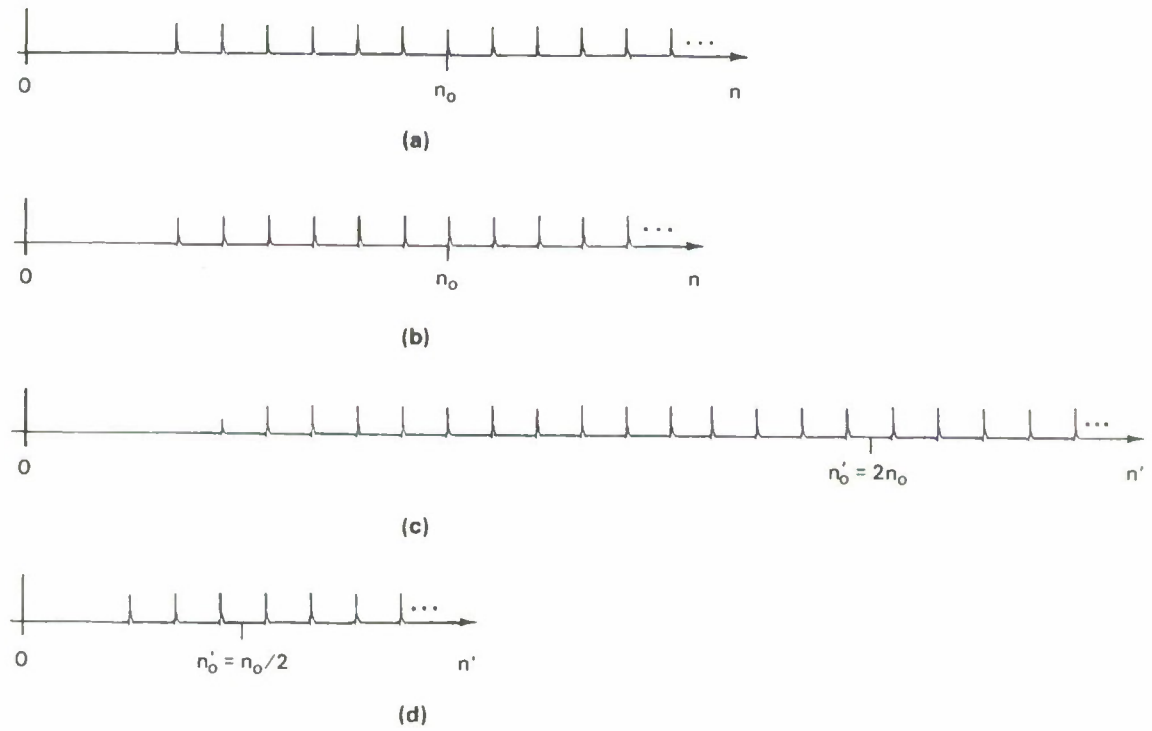


Figure 3-7. Time-scale modification of synthetic waveform. (a) Original, (b) Reconstruction ($\rho = 1.0$), (c) Expansion ($\rho = 1.5$), and (d) Compression ($\rho = 0.5$).

147345-R-01

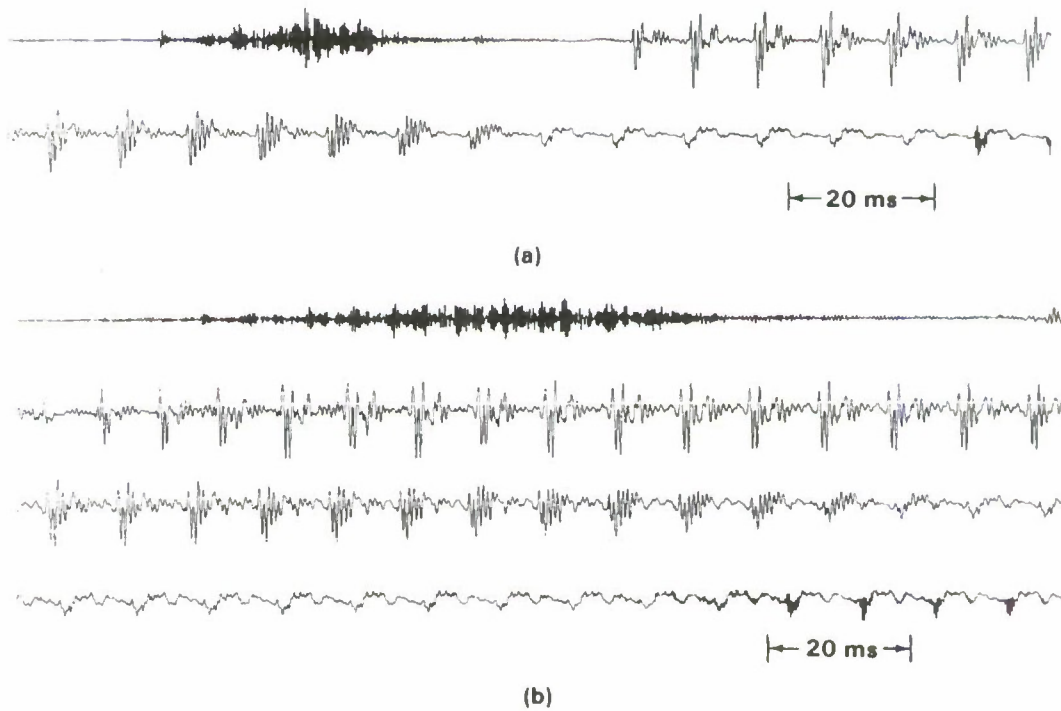


Figure 3-8. Time-scale expansion of speech. (a) Original, and (b) Expansion ($\rho = 2$).

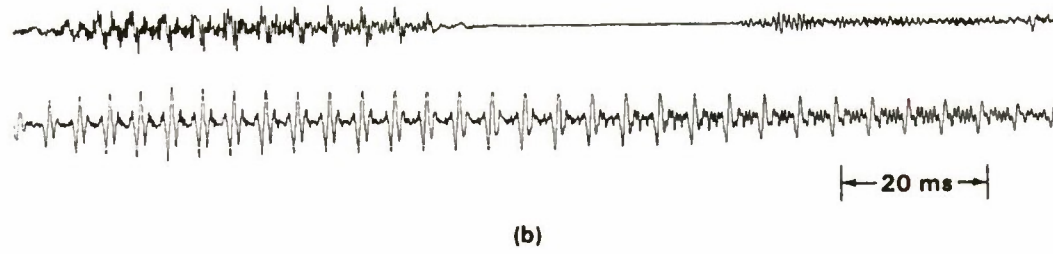
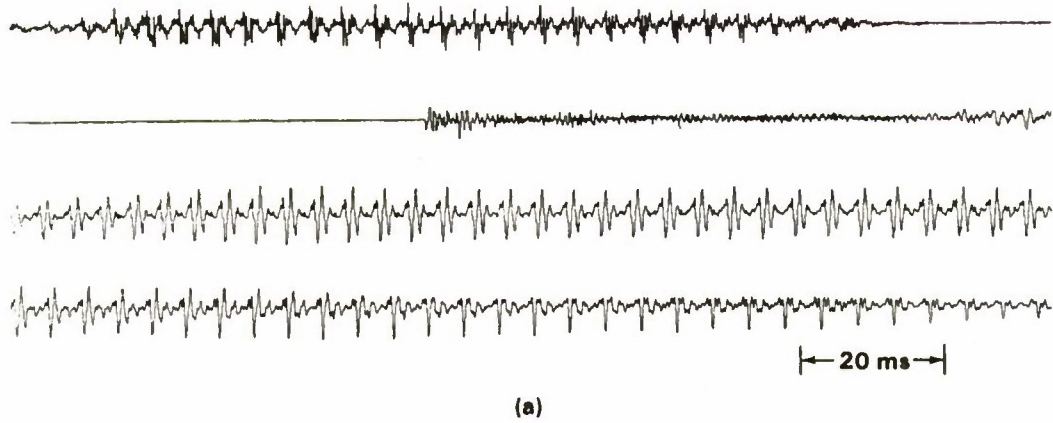


Figure 3-9. Time-scale compression of speech. (a) Original, and (b) Compression ($\rho = 0.5$).

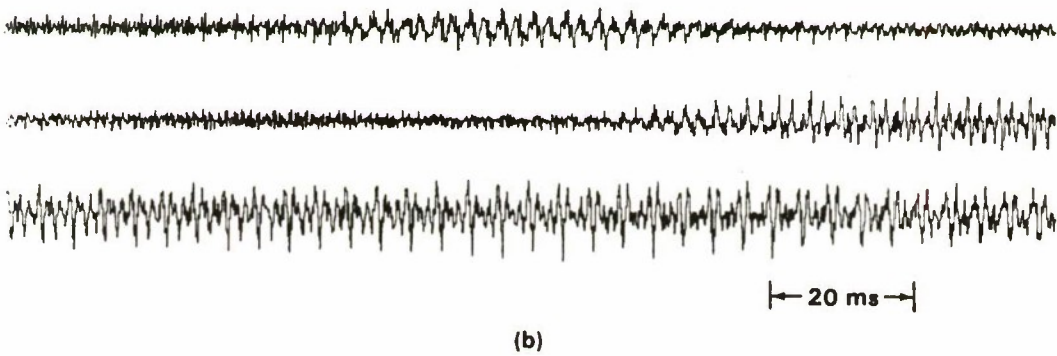
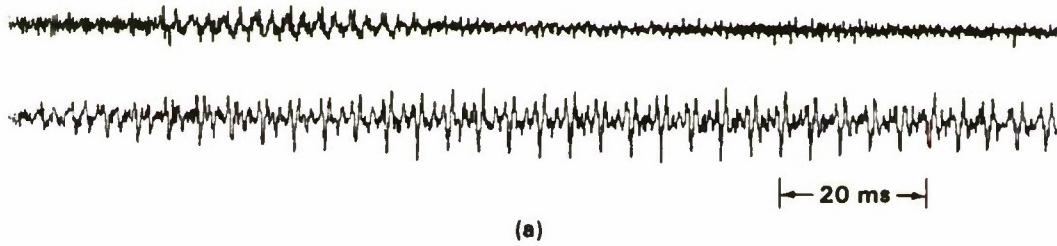
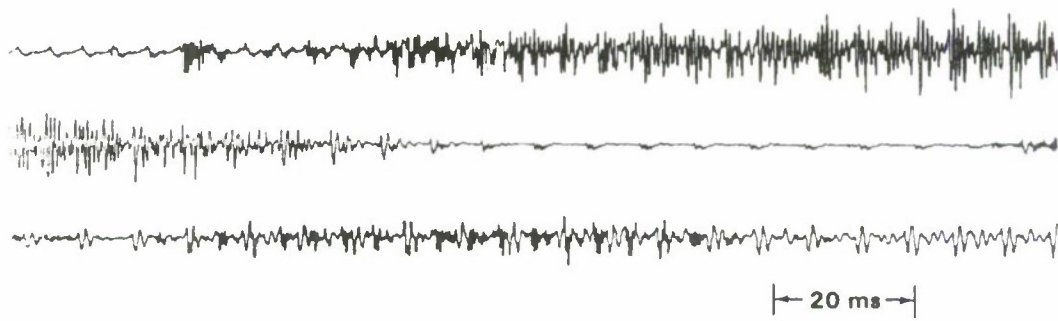


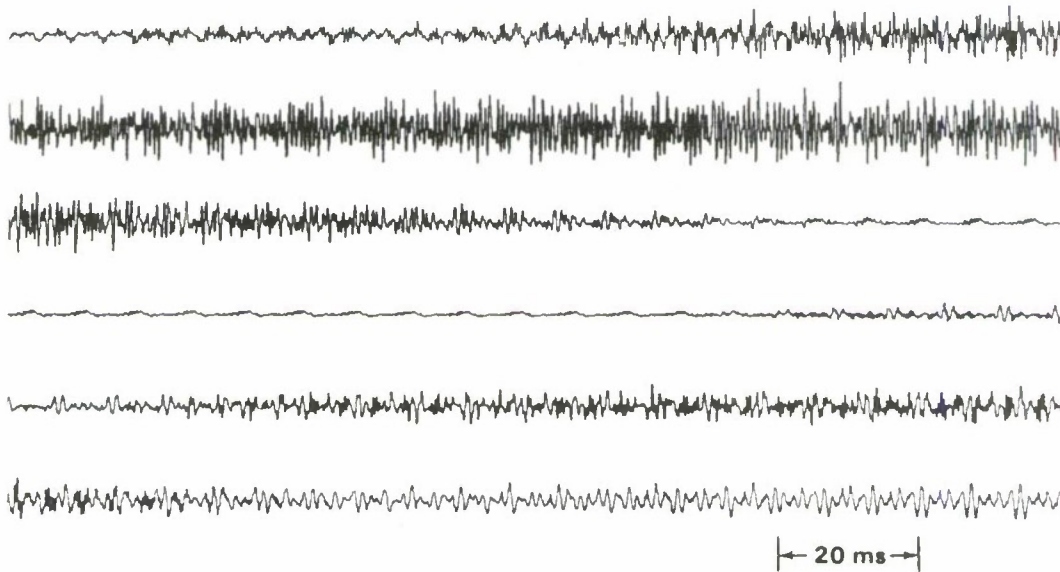
Figure 3-10. Time-scale expansion of speech in music. (a) Original, and (b) Expansion ($\rho = 1.5$).

147347-R-01

149389-R-01



(a)



(b)

Figure 3-11. Time-scale expansion of combined male and female speech. (a) Original, and (b) Expansion ($\rho = 2$).

3.2 Time-Varying Rate Change

With the proposed sinusoidal representation, it is also straightforward to model a time-varying rate change $\rho(t)$. Here the time-warping transformation is nonlinear and is given by

$$t' = W(t) = \int_0^t \rho(\tau) d\tau \quad (3.4)$$

where $\rho(\tau)$ is the desired time-varying rate change. Note that for a constant ρ , (3.4) reduces to the fixed rate-change case (3.1). In this generalization, each time-differential $d\tau$ is scaled by a different factor $\rho(\tau)$. An example of a nonuniform time-warp is illustrated in Figure 3-12 for a

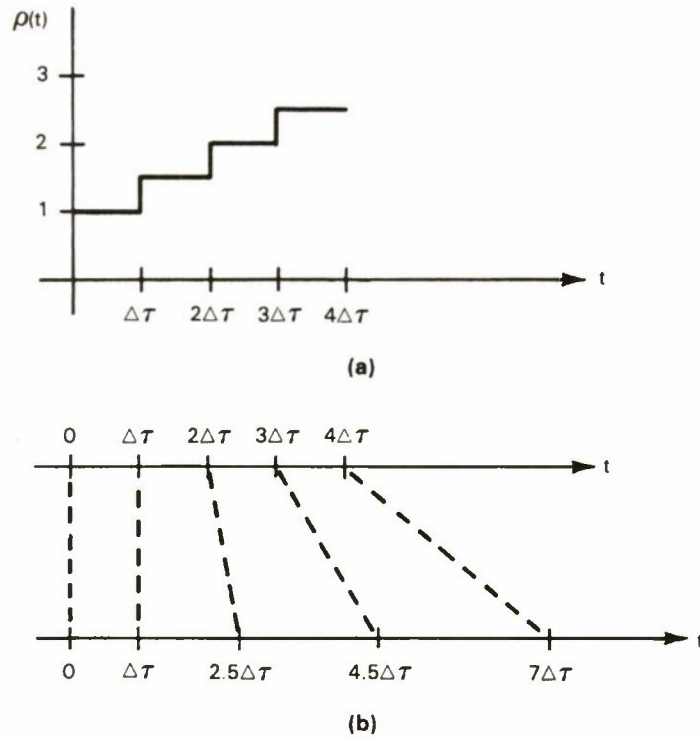


Figure 3-12. Piecewise constant rate change. (a) Rate Change Function, and (b) Time-warp.

piecewise-constant rate change. Speech events which take place at a time t' in the new time scale will have occurred at a time $t = W^{-1}(t')$ in the original time scale where $W^{-1}(\cdot)$ denotes the inverse mapping from the new time scale back to the original time scale.

The speech model for time-varying rate change is given by (3.2) where the inverse time $\rho^{-1}t'$ is now replaced by $W^{-1}(t')$:

$$s'(t') = \sum_{\ell=1}^{L(t')} A'_{\ell}(t') \cos[\phi'_{\ell}(t')] \quad (3.5a)$$

where

$$A'_{\ell}(t') = A_{\ell}[W^{-1}(t')] \quad (3.5b)$$

and

$$\theta'_{\ell}(t') = \Omega'_{\ell}(t') + \Phi_{\ell}[W^{-1}(t')] \quad (3.5c)$$

with

$$\Omega'_{\ell}(t') = \int_{t'_2}^{t'} \omega_{\ell}[W^{-1}(\tau)] d\tau + \phi_{\ell} \quad (3.5d)$$

where $A_{\varrho}(t)$ is the composite amplitude $a_{\varrho}(t)M_{\varrho}(t)$. As with the uniform rate-change case, the phase offset in (3.5d) at time t'_{ϱ} is set so the transformation is consistent with maintaining the shape of a perfectly periodic excitation pulse train in the vicinity of the transformed onset time t'_{ϱ} . Figure 3-13 illustrates nonuniform time-scaling applied to the system phase function with the piecewise constant $\rho(t)$ of Figure 3-12. Similar modifications are made to the system amplitudes and the excitation amplitudes and frequencies. Note that when the rate-change function $\rho(t)$ is set to a constant, all of the expressions in (3.5) reduce to those of the fixed rate-change case (3.2). An important difference, however, in the time-varying rate-change model from the fixed rate-change case is that, for an arbitrary $W^{-1}(t)$, the modified excitation phase is not simply related to the original excitation phase, as in (3.2e). Furthermore, the inverse time mapping $W^{-1}(t)$ generally is difficult to evaluate exactly. Thus in developing an implementation for time-varying time-scale modification based on the mathematical model (3.5), these two issues must first be addressed.

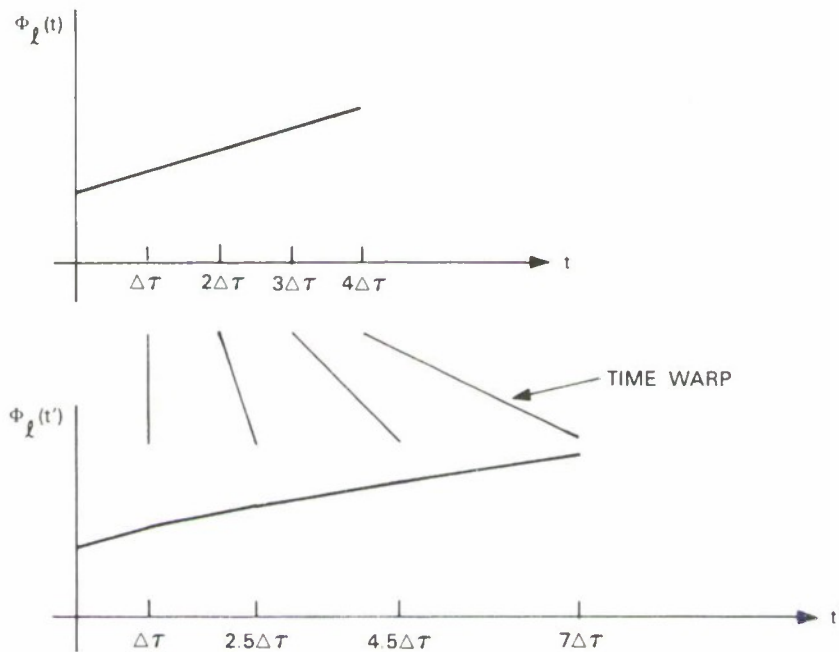


Figure 3-13. System phase mapping for the piecewise constant rate change of Figure 3.12.

One class of time-scale transformations that will be of particular interest invokes a piecewise-constant $\rho(t)$. This condition on $\rho(t)$ allows both exact time inversion and exact evaluation of the integral equation, (3.5d). In addition, it leads to a digital implementation which is a straightforward extension of the constant $\rho(t)$ case of the previous section. An example of a piecewise-constant $\rho(t)$ and a resulting time-scale transformation was illustrated in Figures 3-12 and 3-13. A more general piecewise-functional representation for $\rho(t)$, involving higher-order polynomials (e.g., linear and quadratic), will be discussed later in this section.

In showing that the piecewise constant condition leads to exact inversion of (3.4), for convenience in implementation, the time increment over which the rate change is held constant is assumed equal to the analysis frame duration T , i.e.,

$$\rho(t) = \text{constant} \quad \text{for } kT \leq t \leq (k+1)T \quad (3.6)$$

Nevertheless, there is no restriction on how small this increment can be made (regardless of the frame duration) since *continuous* functional representations of all model parameters are available. (Consequently, an arbitrarily smooth $\rho(t)$ can be approximated as closely as desired.) Given the condition in (3.6), the time mapping (3.4) can then be written for $t'_k < t' \leq t'_{k+1}$ as (t'_k being the beginning of the k th frame):

$$\begin{aligned} t' &= \int_0^{t'} \rho(\tau) d\tau \\ &= \int_0^{t'_k} \rho(\tau) d\tau + \int_{t'_k}^{t'} \rho(\tau) d\tau \\ &= t'_k + \rho(t_k)(t' - t'_k) \end{aligned} \quad (3.7)$$

where t_k and t'_k are the times at the beginning of the k th original frame and modified frame, respectively. The inverse time t and thus the inverse mapping $W^{-1}(\cdot)$ is then given by

$$t = W^{-1}(t') = t_k + \rho^{-1}(t_k)(t' - t'_k) \quad (3.8)$$

Since the parameters of the sinusoidal components are available as continuous functions of time [i.e., equations (2.11) through (2.18)] over each analysis frame in the original time scale, they can always be found at the required inverse time of (3.8). Note that having modified the speech waveform up to time t_k , the time into the k th analysis frame can be written as

$$t = W^{-1}(t') = \rho^{-1}(t_k)(t') \quad (3.9)$$

where t' is interpreted here as the time into the k th modified frame.

With this piecewise constant $\rho(t)$, the excitation phase function can be also considerably simplified. Suppose that the excitation phase estimate $\hat{\Omega}'_2(t')$ has been evaluated up to time t'_k which when inverted corresponds to $t_k = kR$, the beginning of the k th frame in the original time frame. Then with the substitution of variables, $\sigma = \tau - t'_k$, (3.5d) can be written for $t'_k < t' \leq t'_{k+1}$ in terms of a constant and time-varying component:

$$\begin{aligned} \hat{\Omega}'_2(t') &= \hat{\Omega}'_2(t'_k) + \int_{t'_k}^{t'} \hat{\omega}[W^{-1}(\tau)] d\tau \\ &= \hat{\Omega}'_2(t'_k) + \int_0^{t'-t'_k} \hat{\omega}[W^{-1}(\sigma + t'_k)] d\sigma \end{aligned} \quad (3.10)$$

where $\hat{\Omega}'_2(t'_k)$ represents the excitation phase estimate computed up to time t'_k in the new time scale. Now using the recursion in (3.8), then (3.10) can be written as

$$\hat{\Omega}'_2(t') = \hat{\Omega}'_2(t'_k) + \int_0^{t'-t'_k} \hat{\omega}[t_k + \rho^{-1}(t_k)\sigma]d\sigma \quad (3.11)$$

With some further manipulation, the expression in (3.11) can now be written in terms of modified versions of the constant and time-varying excitation phase components of (2.17). Making the substitution of variables $\tau = \rho^{-1}(t_k)\sigma$, (3.11) is given by

$$\hat{\Omega}'_2(t') = \hat{\Omega}'_2(t'_k) + \int_0^{(t'-t'_k)\rho^{-1}(t_k)} \hat{\omega}[t_k + \tau]d\tau/\rho^{-1}(t_k) \quad (3.12)$$

The integral expression in (3.12) is just a time-warped and scaled version of the time-varying excitation phase component $\hat{V}_2(t)$ of (2.17b) over the k th modified frame. Thus for the k th frame (3.12) can be written as

$$\hat{\Omega}'_2(t') = \hat{\Omega}'_2(t'_k) + \hat{V}_2[\rho^{-1}(t_k)(t' - t'_k)]/\rho^{-1}(t_k) \quad (3.13)$$

Likewise, the constant term in (3.13) is similar to the constant term in (2.18) but with scaled phase components. Specifically, (3.13) can be expressed as

$$\hat{\Omega}'_2(t') = \hat{V}_2[\rho^{-1}(t_k)(t' - t'_k)]/\rho^{-1}(t_k) + (\Sigma_2^k)' + \hat{\phi}_2 \quad (3.14a)$$

where $(\Sigma_2^k)'$ can be written recursively as

$$(\Sigma_2^{k+1})' = (\Sigma_2^k)' + \hat{V}_2(R)/\rho^{-1}(kR) \quad (3.14b)$$

with $(\Sigma_2^k)'$ representing the accumulated scaled time-varying excitation phase component. Note that in (3.14) $t' - t'_k$ is the time into the k th modified frame and $\rho^{-1}(t_k)(t' - t'_k)$ is the time into the k th original frame.

With the inverse time recursion (3.8) and the excitation phase recursion in (3.14), a digital implementation of a time-varying rate change system can be realized. In this implementation the inverted time is given by

$$t_{n'} = kR + \rho^{-1}(kR)(n' - n'_k) \quad (3.15)$$

where $t_{n'}$ denotes the inverse to the discrete time index n' and where the sampling interval is assumed normalized to unity. As in the uniform time-scale case, the estimates (2.11) through (2.18) provide functional forms for the parameters in (3.5). Since these functional estimates were derived on a frame-by-frame basis, it is natural to view the inverted time $t_{n'}$ in (3.15) as the *time into* the k th frame within the original time scale. Thus as before this time is computed *modulo* R which is denoted by $(t_{n'})_R$, R being the number of samples over the original frame duration. The discretized version of (3.5) can then be written as

$$\hat{s}'(n') = \sum_{\ell=1}^{L(n')} \hat{A}'_{\ell}(n') \cos[\hat{\Omega}'_{\ell}(n') + \hat{\Phi}'_{\ell}(n')] \quad (3.16a)$$

where

$$\hat{A}'_{\ell}(n') = \hat{A}_{\ell} [(t_{n'})_R] \quad (3.16b)$$

and

$$\hat{\Phi}'_{\ell}(n') = \hat{\Phi}_{\ell} [(t_{n'})_R] \quad (3.16c)$$

and

$$\hat{\Omega}'_{\ell}(n') = \hat{V}_{\ell} [(t_{n'})_R] / \rho^{-1}(kR) + (\Sigma_{\ell}^k)' + \hat{\phi}_{\ell} \quad (3.16d)$$

with $(\Sigma_{\ell}^k)'$ updated as

$$(\Sigma_{\ell}^{k+1})' = (\Sigma_{\ell}^k)' + \hat{V}_{\ell}(R) / \rho^{-1}(kR) \quad (3.16e)$$

and the inverse time recursion is given by

$$t_{n'} = kR + \rho^{-1}(kR)(n' - n'_k) \quad (3.16f)$$

The phase recursion in (3.16e) is initialized with $(\Sigma_{\ell}^k)' = 0$ at the onset of each sine wave. Note also that this recursion requires scaling the time-varying excitation phase by $\rho^{-1}(kR)$ which changes on each frame. This has the effect of keeping the pitch as close to the original as possible.

As with uniform time-scale modification, the discrete-time implementation given in (3.16) can be performed in block fashion where storage of only two consecutive sets of amplitude and phase parameters is required. In fact, for a particular frame, the operations represented by (3.3) and (3.16) are essentially identical. A flow diagram of the process is illustrated in Figure 3-14 and is similar to that of the uniform case given in Figure 3-6.

The generalized time-scale modification system (3.16) was demonstrated using two long speech passages (25-30 s), for a male and for a female speaker, with various time-varying rate changes. For this experiment $\rho(\tau)$ was held constant over the duration of each 5 ms analysis frame. Decreasing the frame duration (and hence the time over which the rate change is held constant) or increasing the frame duration, but bounding it by 10 ms, did not noticeably change the quality. Both linear and oscillatory rate changes were performed. In the linear case the scaling factor changed from unity to 0.5 and to 2. In the oscillatory case, the scaling factor was modulated between 0.5 and 2, one oscillation being about 5 s in duration. The synthetic speech was generally natural-sounding and free of artifacts.

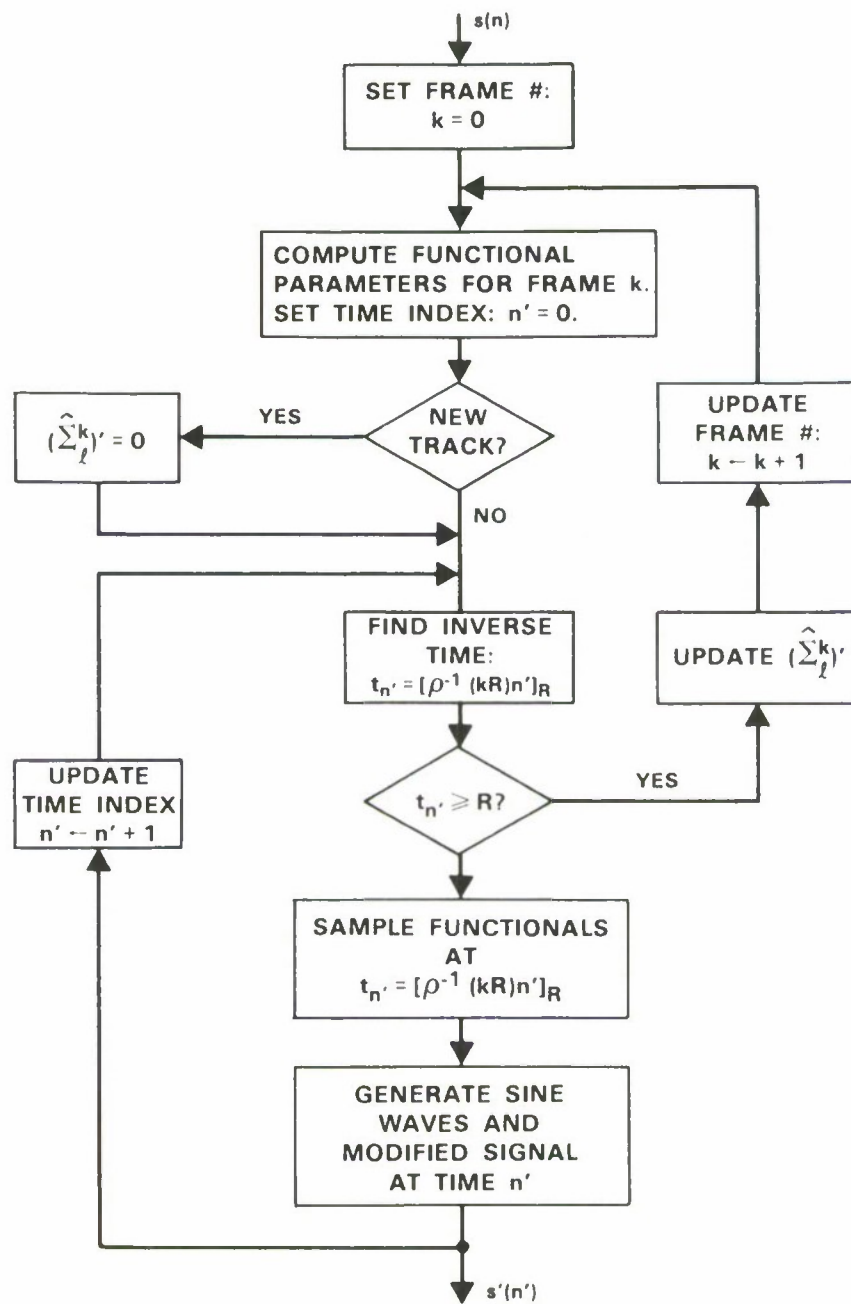


Figure 3-14. Flow diagram of computer implementation of nonuniform rate-change system.

This section has developed just one functional form for $\rho(t)$, i.e., the piecewise-constant case. Clearly, this procedure can be generalized with higher order representations for $\rho(t)$. For example, a piecewise-linear $\rho(t)$ can be used in (3.7) which leads to a time-inversion formula somewhat more complicated than (3.8). Furthermore, since this inversion formula requires a computation of the form $(t')^{1/2}$, and since the frequency track estimate of each sine wave is a quadratic function of time, then the integral expression in (3.10) is easily evaluated. Consequently, given any desired smoothly changing rate-change function $\rho(t)$, a closer fit can be obtained than offered by the piecewise-constant case. More generally, it is possible to obtain even higher-order approximations to an arbitrary $\rho(t)$ over each frame. However, since the piecewise-constant case appears to yield satisfactory quality, these higher-order approximations may result in negligible improvements

3.3 Feature-Based Time-Scale Modification

The accuracy of the representation of time-scale modification of Section 3.1 is subject to a number of conditions. First, as with the sine-wave model for unmodified speech, the vocal tract and vocal cord parameters must be slowly-varying relative to the duration of the vocal tract impulse response. This condition is generally satisfied as is evident from the accurate reconstruction of unmodified speech. A second condition is that the rate change of the actual vocal tract and vocal cord articulators be fixed as a function of time. Generally, this condition will not be satisfied. For example, when the rate of articulation is reduced in natural speech, consonants and fricatives are generally slowed down less than vowels and other more steady-state sounds. Although, as seen in Section 3.1, uniform rate change results in generally high-quality synthesis, the excessive slowing down of rapid speech sounds can render a "drunken man" quality to the synthetic speech. In an attempt to reduce this effect, an adaptive rate-change system was implemented. This system was designed to generate more natural sounding speech by continuously adapting the rate change to the temporal characteristics of speech.

The degree of time-compression or expansion in an adaptive system should become a function of the rate at which speech events change. Transient events should be slowed down less. This requires that some sort of detector be developed which can measure significant speech activity. One such detector is the spectral derivative²² which in this report is defined using normalized squared differences computed from the available matched peak magnitudes. Specifically, the spectral derivative over two consecutive frame boundaries, is given by

$$D(k) = \frac{[\sum_{\ell} (\hat{A}_{\ell}^k - \hat{A}_{\ell}^{k-1})^2]^{1/2}}{[\sum_{\ell} (\hat{A}_{\ell}^k)^2]^{1/2}} \quad (3.17)$$

where k refers to the frame number and \hat{A}_{ℓ}^k is defined in (2.8a). The spectral derivative as defined in (3.17) tends to increase at voiced/unvoiced boundaries, during unvoiced speech and during consonant transitions. This norm has the additional advantage that it is bounded; in particular, it is straightforward to show with a simple geometric argument that

$$0 \leq D(k) \leq 2 \quad (3.18)$$

Thus when $\rho(t)$ is made a function of $D(k)$, the range of $\rho(t)$ can be determined from (3.18). In slowing down the rate of articulation, when the spectral derivative is high due to rapid speech activity, the time scale should be expanded less than when the spectral derivative is low. Thus $\rho(t)$ is a monotonically decreasing function of $D(k)$. In compressing speech, the time scale should be compressed less when the spectral derivative is large, and so in this case $\rho(t)$ is a monotonically increasing function of $D(k)$. The functional relation should be chosen to lead to some desired average modified time scale. For example, one functional relation between $D(k)$ and $\rho(t)$ is given by

$$\rho(kR) = 3 - D(k) \text{ for } 0 \leq D(k) \leq 2 \quad (3.19)$$

where $\rho(t)$ is specified only at frame boundaries and is linearly interpolated across each frame. Under the assumption that $D(k)$ is equally probable over its range of values, it can be shown that (3.19) results in an average rate change factor of two.

An example of the spectral derivative defined in (3.17) using sine-wave amplitudes, is given in Figure 3-15 where it is seen that the spectral derivative increases for regions of frication and other regions of "high activity". The top segment represents the fricative sound "sh" in the word

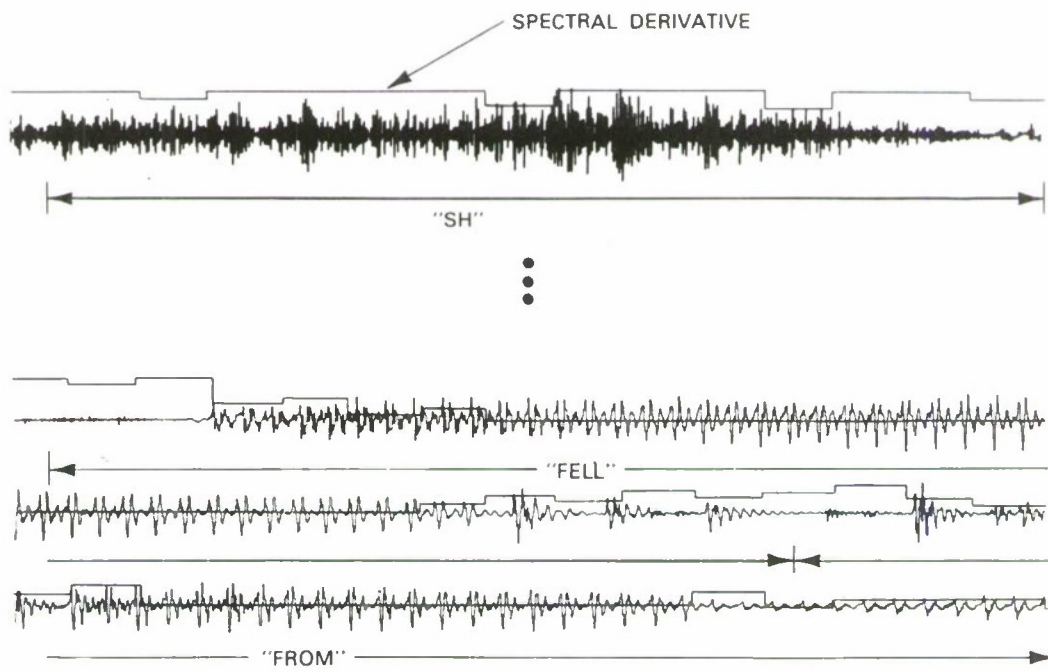


Figure 3-15. Speech segments from the passage, "she feel from the car," with superimposed spectral derivative. The spectral derivative has been normalized to lie between zero and unity and was assumed constant over an analysis frame.

149383-R-01

“she”, and it can be seen that the spectral derivative is high. In the lower segment, the spectral derivative is high during the plosive at the onset of the “f” in word “fell”, while the spectral derivative decreases in traversing to the voiced sound “e”.

In some preliminary experiments, speech passages of two seconds in duration, slowed down by a fixed factor of two, were compared with the same speech passages slowed down by the adaptive system. In the adaptive system, the overall average rate was kept approximately equal to the fixed rate by applying the functional relation of (3.19). Informal listening indicated that the fricative and consonant regions sounded more natural when modified with the adaptive system. Moreover, the “drunken man” effect that occurred for the fixed rate-change system seems to be reduced. Additional more extensive listening tests, however, are required before definitive conclusions can be drawn.

The broad goal in this work was to generate modified speech that sounded more like that spoken at a slower or faster rate. Although the rapid transitions in speech can be identified with rapid transitions in the magnitude of its short-time spectrum^{22,23}, it is not clear that such a speech “feature” is the only indicator or the most accurate indicator of when to change the rate. More accurate indicators may require linguistic knowledge of changing events that occur in slowly and rapidly spoken speech. Furthermore, additional work would need to be done relating the speech activity and the rate of time-scale modification.

4. FREQUENCY TRANSFORMATIONS

Since the synthesis procedure consists of adding up the sinusoidal waveforms for each of the measured frequencies, the procedure is ideally suited for performing frequency transformations. In this section, two such transformations are presented. The first transformation, often referred to as frequency scaling,¹⁷ compresses or expands the short-time Fourier transform in frequency such that both the spectral envelope and the spacing between harmonic components in the spectrum (and thus the pitch contour) are scaled. The second transformation scales the pitch contour but preserves the short-time spectral envelope. As with time-scale modification, it is shown that pitch modification can be performed with a time-varying scale factor.

4.1 Frequency Scaling

Frequency compression or expansion of the short-time Fourier transform can be represented with a slight modification of the excitation phase given in (2.2). In this procedure, each frequency track $\omega_\ell(t)$ is scaled by a desired factor β . This results in the modified excitation phase:

$$\begin{aligned}\Omega'_\ell(t) &= \int_{t_\ell}^t \beta \omega_\ell(\tau) d\tau + \phi_\ell \\ &= \beta V_\ell(t) + \phi_\ell\end{aligned}\quad (4.1)$$

The original composite amplitude $A_\ell(t)$ and system phase $\Phi_\ell(t)$ estimates are simply shifted to the new location of the frequency track $\beta\omega_\ell(t)$. These operations are equivalent to shifting the excitation function to the new frequency track locations and scaling of the frequency argument of the vocal tract system function to form $H(\beta\omega, t)$. An illustration of frequency compression is given in the time-frequency domain in Figure 4-1.

Using (4.1) a discrete-time implementation of a frequency-scaling system can then be realized as a simple extension of (2.20). The resulting modified waveform over the k th frame is given (in discrete time) by

$$\hat{s}'(n) = \sum_{\ell=1}^{L(n)} \hat{A}_\ell(n) \cos [\hat{\Omega}'_\ell(n) + \hat{\Phi}_\ell(n)] \quad (4.2a)$$

where

$$\Omega'_\ell(n) = \beta V_\ell(n) + (\Sigma_\ell^k)' + \phi_\ell \quad (4.2b)$$

with $(\Sigma_\ell^k)'$ computed recursively as

$$(\Sigma_\ell^{k+1})' = (\Sigma_\ell^k)' + \beta V_\ell(R) \quad (4.2c)$$

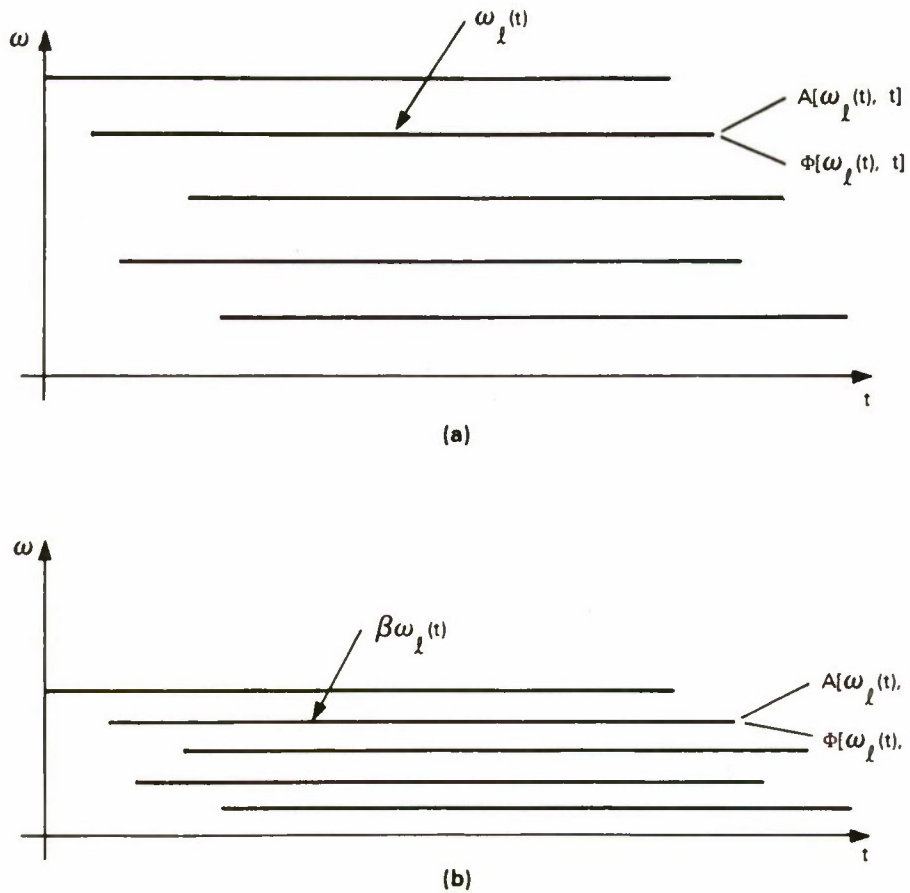


Figure 4.1. Time-frequency illustration of frequency compression. (a) Original. (b) Frequency compressed.

where the index $n = 0, 1, 2 \dots R - 1$ is interpreted as the discrete time into the k th frame and where the amplitude and phase functions are sampled versions of their continuous counterparts in (2.11) to (2.18). This waveform modification corresponds to an expansion or compression of the spectral envelope and a change in pitch. In some applications, such as frequency-scaling for the hearing impaired⁴, it may be required to scale the spectrum nonlinearly in frequency. This transformation can also be easily realized with (4.2) by simply making the scale factor β a function of frequency (i.e., a function of the frequency track index ℓ). Note that as in the previous sections, the computer implementation of (4.2) can be performed on a frame-by-frame basis, thus requiring storage of only two sets of amplitude and phase parameters.

In one experiment, the spectrum of a speech signal was mapped from the frequency range 0-5 kHz down to 0-4 kHz and thus the pitch was also lowered by 20%. An example of this spectral transformation is given in Figure 4-2 where it is seen that the spectral envelope has been compressed and the pitch lowered (i.e., the harmonic spacing is decreased). In a similar experiment, the speech spectrum was scaled from the range 0-4 kHz up to 0-5 kHz.

152711-N-02

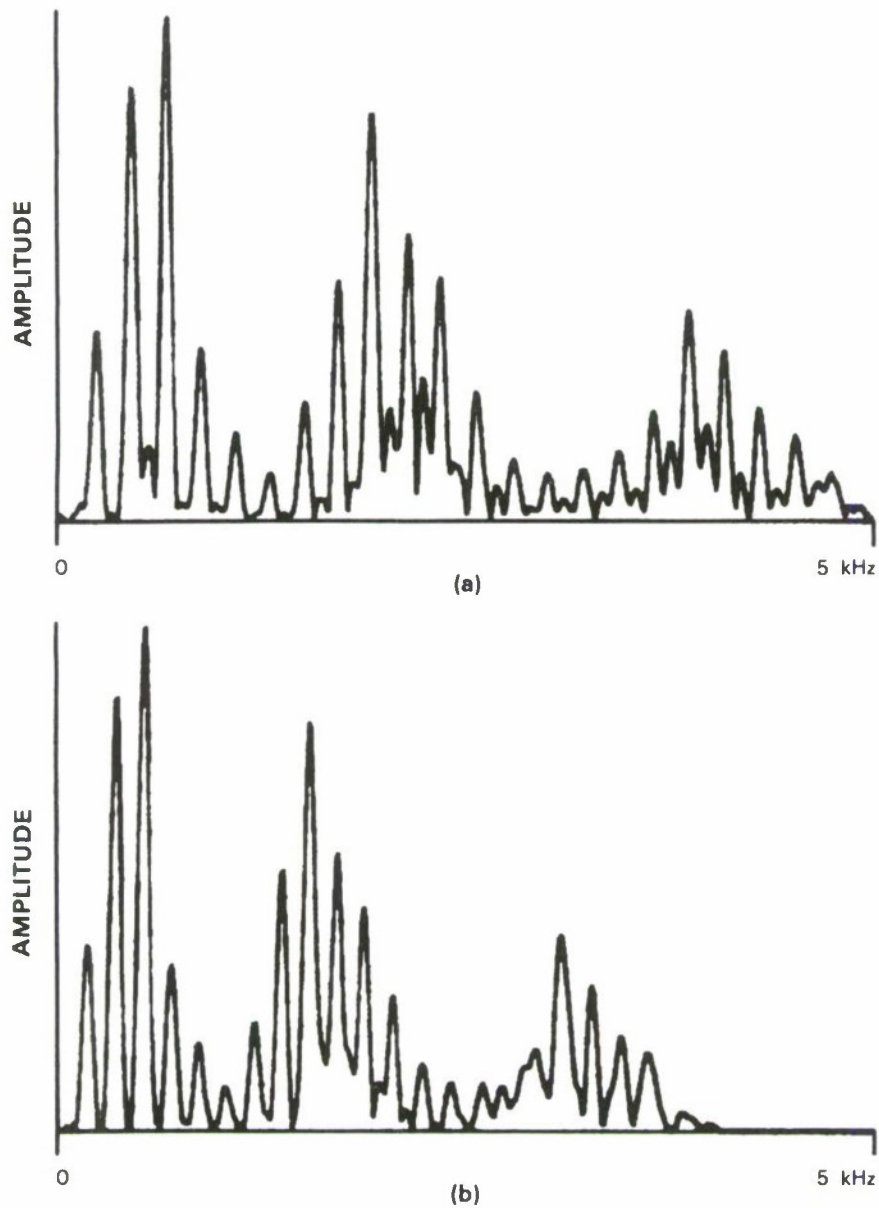


Figure 4.2. Frequency compression of the spectral magnitude. (a) Original. (b) Compression.

4.2 Pitch Modification

In a simplified model for pitch modification, the excitation function of the speaker is modified as in the previous section, while the spectral envelope of the speaker's vocal tract system function is unchanged. This corresponds to voice quality which is similar to the original, but

which is characterized by a change in pitch. The model for pitch-modification is depicted in Figure 4-3. This modification is easily simulated with a system based on the sinusoidal analysis/synthesis structure in Figure 2-2 and 2-4.

The first step in frequency scaling of the excitation function requires that the estimated frequency track of each sine-wave component be scaled by a desired factor β to generate a new frequency track $\beta\hat{\omega}_\ell(t)$. This results (in discrete time) in an excitation phase given in (4.2b) and 4.2c). The next step in frequency scaling the excitation function requires that the excitation amplitude estimate $\hat{a}_\ell(t)$ be shifted to the new frequency track locations. It is assumed that only the frequencies within the resulting bandwidth of the modified excitation function are used in synthesis (i.e., to maintain the original speech bandwidth with frequency compression would require high-frequency regeneration).

To preserve the shape of the short-time spectral envelope, the system amplitudes and phases must be computed at the new frequency track locations, $\beta\hat{\omega}_\ell(t)$. These system amplitude and phase functions are obtained by first sampling (in frequency) the smooth system amplitude and phase estimates, derived from the homomorphic analyzer, at the modified frequencies $\beta\hat{\omega}_\ell^k$. These values are then linearly interpolated across successive frame boundaries to generate the amplitude estimate, $\hat{M}[\beta\hat{\omega}_\ell(t),t]$, and the phase estimate, $\hat{\Phi}[\beta\hat{\omega}_\ell(t),t]$, along the new frequency track locations. A time-frequency illustration of the excitation and system modifications required in the model for pitch modification is given in Figure 4-4.

With the above modified excitation and system components, the resulting modified discrete-time waveform over the k th frame is given by

$$\hat{s}'(n) = \sum_{\ell=1}^{L(n)} \hat{a}_\ell(n) \hat{M}'_\ell(n) \cos [\hat{\Omega}'_\ell(n) + \hat{\Phi}'_\ell(n)] \quad (4.3a)$$

where

$$\hat{M}'_\ell(n) = \hat{M}_\ell[\beta\hat{\omega}_\ell(n),n] \quad (4.3b)$$

and

$$\hat{\Phi}'_\ell(n) = \hat{\Phi}_\ell[\beta\hat{\omega}_\ell(n),n] \quad (4.3c)$$

and

$$\hat{\Omega}'_\ell(n) = \beta\hat{V}_\ell(n) + (\Sigma_\ell^k) + \hat{\phi}_\ell \quad (4.3d)$$

with $(\Sigma_\ell^k)'$ computed recursively as

$$(\Sigma_\ell^{k+1})' = (\Sigma_\ell^k)' + \beta\hat{V}_\ell(R) \quad (4.3e)$$

where (4.3b) and (4.3c) are the discrete-time forms of the above continuous-time magnitude estimate $\hat{M}[\beta\hat{\omega}_\ell(t),t]$ and phase estimate $\hat{\Phi}[\beta\hat{\omega}_\ell(t),t]$. The discrete-time function $\hat{a}_\ell(n)$ is a sampled version of the continuous function (2.11b). As before, the discrete-time index $n = 0, 1, 2 \dots R - 1$ is viewed as the time into each frame.

152709-N-02

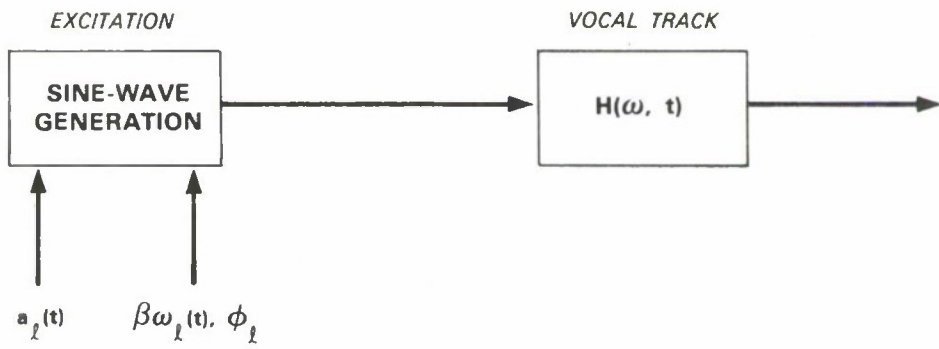


Figure 4.3. Sinusoidal model for pitch modification.

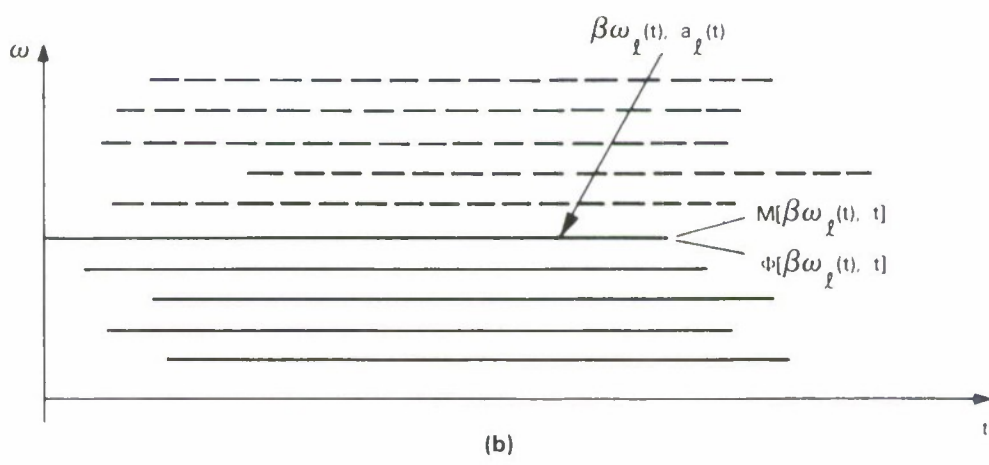
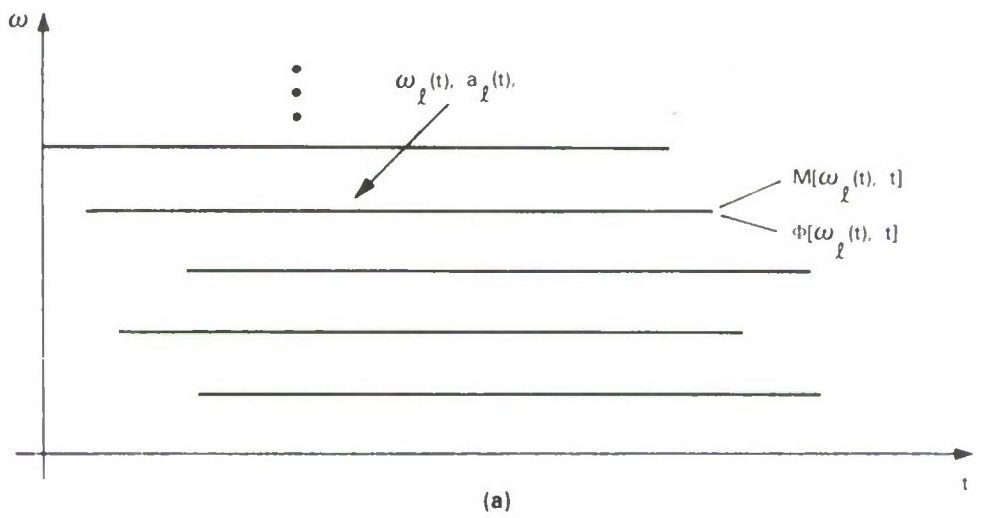


Figure 4.4. Time-frequency illustration of pitch modification. (a) Original. (b) Pitch-scaled (lowered).

152708-N-02

The system was first evaluated using the synthetic waveform of Figure 4-5. In this example the pitch was altered with factors of $\beta = 1.5$ and $\beta = 0.5$ with reconstructions using 40 peaks over 4 kHz. With real speech, the system was demonstrated by scaling the pitch over a range from $\beta = 0.5$ to $\beta = 2$ for a variety of male and female speakers. An example of a pitch increase of 50% (i.e., $\beta = 1.5$) is illustrated in Figure 4-6 where it can be seen that the harmonic line spacing has decreased while the spectral shape is maintained. Figure 4-7 shows in the time domain the result of pitch lowering by 20% where the pitch period has increased. The synthetic speech resulting from this system is smooth and without artifacts such as glitches or reverberation. Except for a lack of formant shaping, the reconstruction takes on the characteristics of higher or lower pitched speakers. With a scaling factor less than unity, comparison with the original speech was performed over a frequency range lower than the original frequency band to avoid the requirement of high-frequency regeneration.

It is possible to generalize Equation (4.1) used as the basis for pitch scaling by making the scaling factor β a function of time. The modified frequency tracks are given by $\beta(t)\omega_{\ell}(t)$ and the resulting excitation phase can then be written as

$$\Omega'_{\ell}(t) = \int_{t_{\ell}}^t \beta(\tau)\omega_{\ell}(\tau)d\tau + \hat{\phi}_{\ell} \quad (4.4)$$

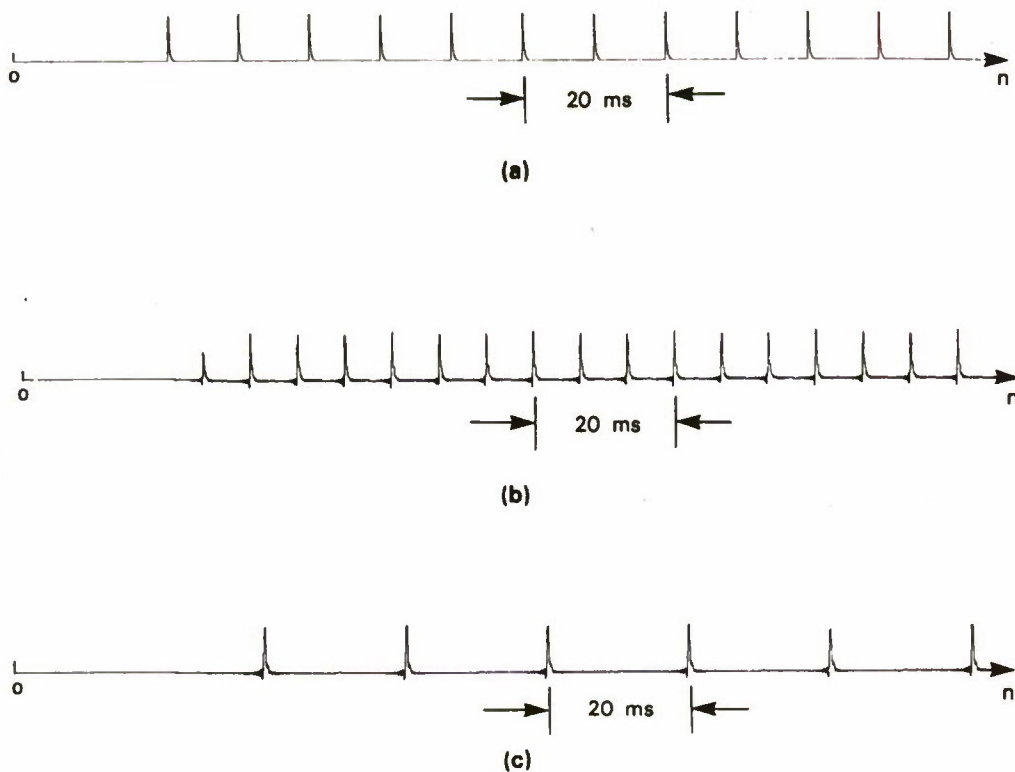


Figure 4.5. Pitch modification of synthetic waveform. (a) Original. (b) Increase in pitch ($\beta = 1.5$). (c) Decrease in pitch ($\beta = 0.5$).

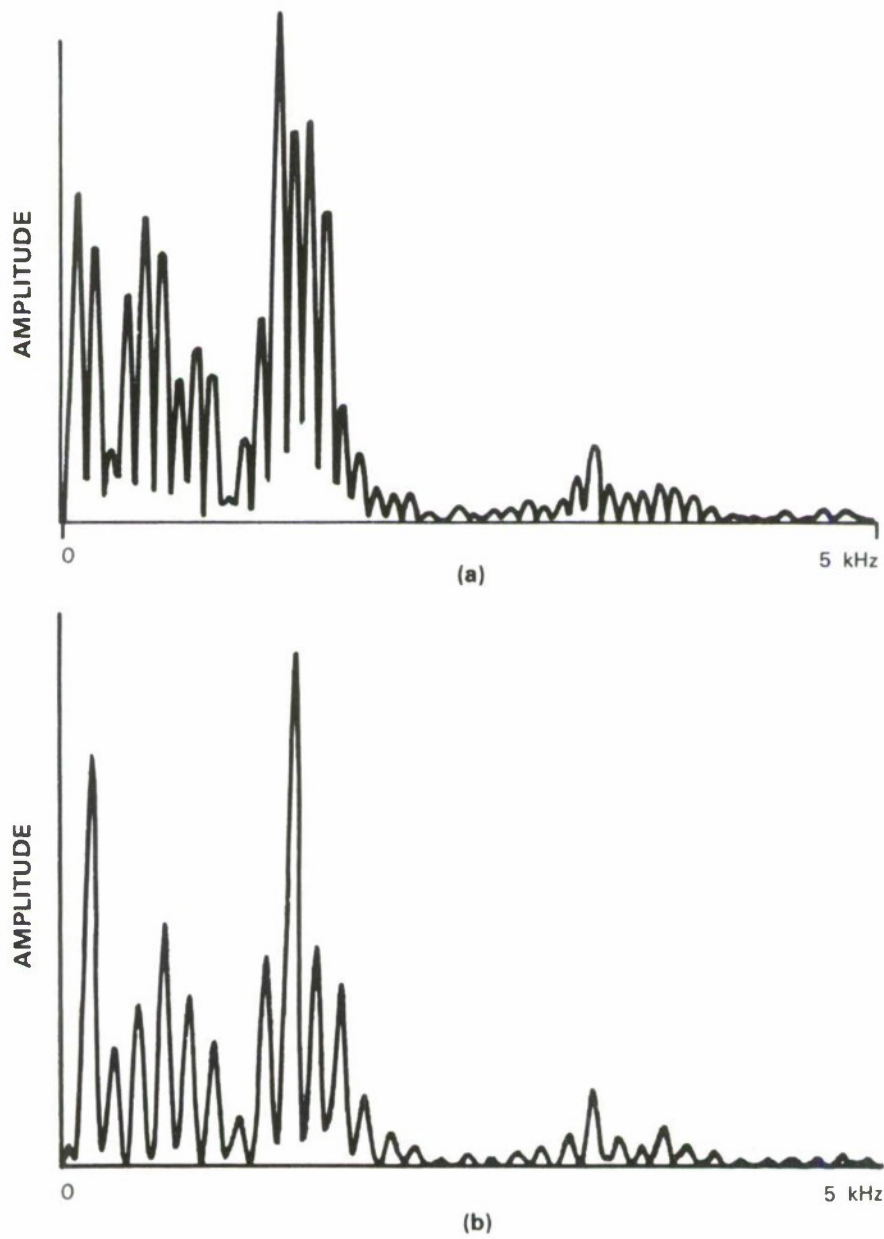


Figure 4.6. Pitch modification of speech in the frequency domain. (a) Original. (b) Pitch-scaled spectral magnitude ($\beta = 1.5$).

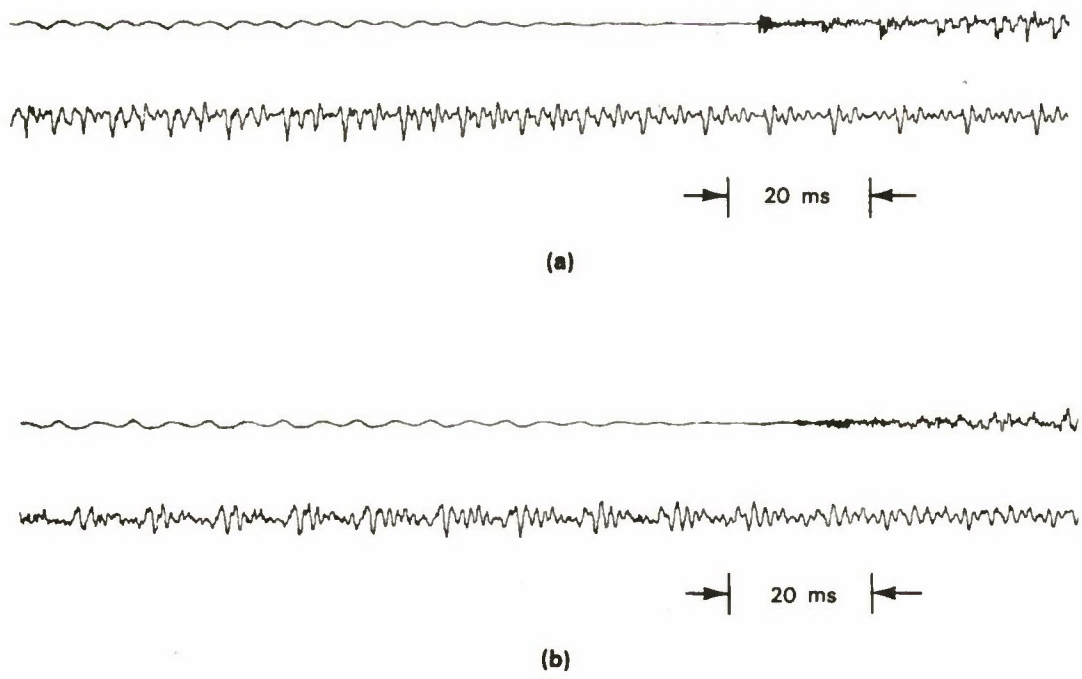


Figure 4.7. Pitch modification of speech in the time domain. (a) Original. (b) Pitch-scaled ($\beta = 0.8$).

One approach to digitally implementing (4.4) is to constrain $\beta(t)$ to be piecewise-constant (as was done for time-scale modification) and solve the integral exactly over each frame. This results in an implementation which is identical to the uniform pitch-scale case over each frame. However, the sudden jumps in the resulting frequency contours are not desirable for high-quality speech synthesis. Consequently, other approaches must be taken. In one method for a discrete-time implementation of (4.4), the pitch-scale factor $\beta(t)$ is assumed to vary slowly with respect to the sampling time interval. Under this condition, the integral in (4.4) can be approximated by the recursion

$$\hat{\Omega}'_k(n) = \hat{\Omega}'_k(n-1) + \beta(n)\hat{\omega}_k(n) \tag{4.5}$$

where here the discrete-time index n is thought of as the time into the k th frame and where the recursion is initialized with the phase resulting from the previous $k-1$ st frame. Another approach, similar to that discussed at the end of Section 3.2, allows $\beta(t)$ to take on a higher-order piecewise-functional form (e.g., linear or quadratic), thus assuring the continuity of frequency. Since $\hat{\omega}_k(\tau)$ in (4.4) is a quadratic function of time, then the integral in (4.4) can be evaluated exactly over each frame. Such an approach avoids having to approximate the desired frequency trajectory and may also avoid phase drift which can occur through the recursion (4.5).

In demonstrating this system, a time-varying scale factor $\beta(t)$ was applied to two long passages (25-30 s), one for a male and one for a female speaker. The excitation phase was computed using the recursive approximation of (4.5). A piecewise-functional representation of

152711-N-01

$\beta(t)$ has not yet been simulated. For the female speaker, the pitch was modulated by $\pm 20\%$ corresponding to an oscillation in $\beta(t)$ over the range 0.8 to 1.2. For the male speaker, the pitch was modulated with a 20% lowering to a 50% increase corresponding to an oscillation in $\beta(t)$ over the range 0.8 to 1.5. In both cases, the frequency range for analysis was adapted to the pitch change such that the resulting synthesized speech fell over a 4 kHz range. As with a uniform pitch change, this time-varying modification resulted in natural-sounding synthetic speech which was free of artifacts.

Note that the proposed model for pitch modification has neglected changes in the system spectral characteristics which may take place during human pitch modification. For example, to convert a female voice to a male-like voice, the vocal tract spectral envelope may need to be compressed by as much as 20% in frequency.⁹ The transformation procedure given here preserves the system spectral envelope while altering the fundamental frequency of the excitation function. Hence it may be desirable to develop a further generalization which allows modification of the system spectral amplitude and phase as well as the excitation.

5. JOINT TIME-FREQUENCY MODIFICATIONS

The speech transformation systems of the previous sections can be further generalized to perform simultaneous time-scale modification, frequency-scaling and pitch-scaling. These joint operations can be carried out by simultaneously stretching and shifting frequency tracks. They can also be performed with a continuously adjustable rate change $\rho(t)$ and frequency scaling $\beta(t)$. By combining the modified excitation phases (3.5d) and (4.1), the resulting generalized excitation phase is given in continuous time by

$$\Omega'_\ell(t') = \int_{t'_\ell}^{t'} \beta(\tau)\omega_\ell[W^{-1}(\tau)]d\tau + \phi_\ell \quad (5.1)$$

where $W^{-1}(t)$ is the inverse time mapping of Section 3.2 which gives time values in the original time scale required for time-varying rate change. When pitch modification is one of the desired operations, it follows from the previous section that the system amplitude and phase will be computed along the modified frequency tracks $\beta(t)\omega_\ell [W^{-1}(t)]$.

Evaluation of (5.1) in a discrete-time implementation requires a procedure similar to those of Sections 3.2 and 4.2 when applying time-varying, time-scale, and pitch modifications. One frame-based method which invokes an approximation, but which is easy to implement, uses the time-inversion recursion similar to (4.5) resulting in the recursion for the excitation phase

$$\hat{\Omega}'_\ell(n') = \hat{\Omega}'_\ell(n' - 1) + \beta[(t_{n'})_R]\hat{\omega}_\ell [(t_{n'})_R] \quad (5.2)$$

where the inverse time $t_{n'}$ given by (3.16f) is computed *modulo* R and where the recursion is initialized with the phase resulting from the $k - 1$ st frame. Another approach uses functional representations of $\beta(t)$ and $\rho(t)$ resulting in an exact evaluation of the integral expression in (5.1). This approach allows for a closer approximation to arbitrary time-varying functions.

In first demonstrating the capability of the system to perform joint operations, frequency-compression and time-scale expansion were performed simultaneously, both by fixed factors where $\beta = 0.5$ and $\rho = 0.5$. The original values of the excitation amplitude and system amplitude and phase are expanded in time and shifted to the new frequency tracks $0.5\hat{\omega}_\ell(t/2)$. Figure 5-1 a,b shows the original and modified speech from a female speaker. The time scale has been expanded and the pitch period increased. The inverse to these joint operations was also performed; i.e., simultaneous time-scale compression and frequency-expansion, illustrated in Figure 5-1c, was performed on the modified waveform of Figure 5-1b with fixed factors of $\rho = \beta = 2$. These operations effectively invert the original modifications thus resulting in an estimate of the original speech waveform. Here the pitch period duration is increased by a factor of two. Thus in order to maintain the original frequency resolution, an analysis window of twice the original length (i.e., from 20 ms to 40 ms) was used in implementing the inverse operations. Since the time-scale was expanded, little time resolution was sacrificed with this increased analysis window length. Although the perceptual difference of the reconstruction from the original is nearly unnoticeable

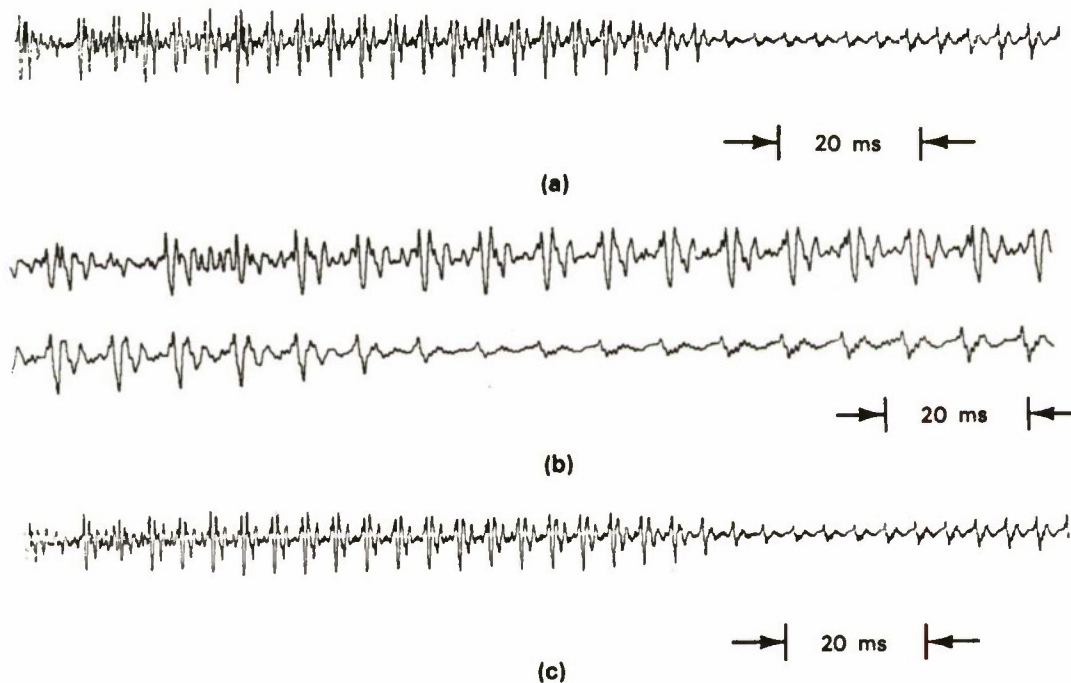


Figure 5.1 Joint frequency scaling and time-scale modification. (a) Original, (b) Frequency compression and time-scale expansion, (c) Inversion of Figure 5.1 (b).

in this case, some slight degradation in rapid transitions was occasionally heard, probably due to the smearing by sequential operations. In particular, very low-pitch speakers, requiring long analysis windows, appear to be most sensitive to sequential operations. The effect of the concatenation of such transformations needs further investigation.

The flexibility of the system was also demonstrated with simultaneous time-varying time-scale and pitch-scale modification. Here the excitation phase was computed according to (5.2), and the system amplitude and phase are sampled along the modified frequency trajectories $\beta(t')\hat{\omega}_q[W^{-1}(t')]$ (in discrete-time by the phase correction term $\beta[(t_n)_R]\hat{\omega}_q[(t_n)_R]$ of (5.2)). In one set of experiments, the operations of linearly increasing and decreasing pitch and the time-scale were jointly applied to 3 s male and female passages. The time scale changed by 50% and the pitch changed by 20% in both directions. In experiments with a 30 s male passage and a 25 s female passage, these joint operations were successfully demonstrated for an oscillatory pitch change with a $\pm 20\%$ modulation and an oscillatory time-scale change with a $\pm 50\%$ modulation.

Finally, similar joint operations were used to enhance a passage which suffered from pitch disjunctions, background noise and a lack of clarity due to the rapidity of the spoken speech. The problem of pitch disjunctions occurs in butting segments of speech together without pitch contouring or smoothing. In this particular problem, three concatenated passages of the same

speaker were analyzed. An average pitch was computed within each passage. A constant pitch-scaling was then performed over each passage so that all three passages took on about the same average pitch. The pitch disjunctions were reduced in the synthesized speech. Simultaneously with this transformation, the waveform was slowed down by a factor of $\rho = 1.5$. The pitch changes were more evident and the clarity of the passage improved.

6. DISCUSSION

In this report, a sinusoidal representation of the speech production mechanism was used as the basis for an analysis/synthesis technique which requires specification of amplitudes, frequencies, and phases of vocal tract and excitation contributions of the component sine waves. This system was successfully applied to a variety of speech transformations including time-scale modification, frequency scaling and pitch scaling. Both fixed and continuously adjustable changes were possible. These transformations do not require either an explicit pitch estimate or voiced/unvoiced decisions. Although this waveform representation was originally designed for single-speaker signals, it is equally capable of reconstructing and modifying nonspeech signals such as music, multiple speakers, marine biologic sounds and speech recorded in the presence of noise and musical backgrounds. Background interferences are modified along with the speech, and were found to be synthesized naturally and without artifacts.

The main computational load of this system is in the evaluation of a high-resolution spectrum using a 512-point FFT, homomorphic filtering for vocal tract parameter estimation, and the generation of the sinewave components. Other operations such as frequency estimation by peak-picking, frequency matching, and phase interpolation and unwrapping add an insignificant amount to the overall computational load. A study of the computational complexity and feasibility of this system is being made by way of a 16-bit real-time (fixed-point) implementation on Lincoln Laboratory's Digital Signal Processors (LDSP)²⁴. The current status of the implementation indicates that the method, with an analysis frame rate of 20 ms, can be realized in real-time using a few commercially available signal processing chips. A number of proposed multiple-processor architecture designs appear to be feasible both in terms of cost and size. More detailed design studies of processor architectures for the sinusoidal analysis/synthesis system are in progress.

It should be noted that an earlier "magnitude-only" version²⁵ of the sine-wave-based system provided an important stepping stone to the speech transformation system in this report. The baseline analysis/synthesis for this system did not rely on a measurement of phase nor did it use the speech production model; rather the sine-wave phase function was obtained by integrating a frequency trajectory formed by linearly interpolating matched frequencies over consecutive frames. While the transformed speech was very intelligible and free of artifacts, it was perceived as being different in quality from the original; the differences were more pronounced for low-pitched (i.e., pitch less than about 100 Hz) speakers. When the magnitude-only system was used to synthesize noisy speech, the noisy speech took on a tonal quality that was unnatural and annoying. The use of the measured sine-wave phase and the introduction of the speech production model resulted in a much improved quality in the modified speech.

In spite of the initial success of the system in this report, there remain a number of areas of possible improvement and some interesting questions for further exploration. For example, although the modified synthetic waveforms tend to be speech-like in appearance, some structure of the original waveform is lost, due possibly to the minimum phase vocal tract assumption

implicit in the homomorphic analyzer and to dispersion of the excitation function. Thus alternate methods of estimating the system phase and procedures for making the excitation function less dispersive might be sought. One of the more interesting questions yet unanswered involves the good performance of the system in the face of nonspeech signals and speech in interference. Since the model underlying the modification system is based on the speech production mechanism, the robustness of the system to such a signal class is not understood. Finally, this report has only touched on the invertibility of the speech transformations, a property which may have considerable practical importance. For example, bandwidth reduction prior to waveform coding could be achieved using the frequency compression transformation. However, it is required that the coded speech be expanded back to its original bandwidth. Since this inversion process requires that speech analysis take place over a 20-30 s duration on the frequency-compressed waveform, it may be difficult to achieve the frequency resolution required for adequate reconstruction.

REFERENCES

1. L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech* (Prentice Hall, Englewood Cliffs, New Jersey, 1978).
2. M.R. Portnoff, "Time-scale Modification of Speech Based on Short-Time Fourier Analysis," Sc.D. Dissertation, Dep. Elec. Eng. Comput. Sci., MIT, Cambridge, Massachusetts (April 1978).
3. M.R. Schroeder, J.L. Flanagan, and E.A. Lundry, "Bandwidth Compression of Speech by Analytic-Signal Rooting," *Proc. IEEE*, **55**, 396-401 (March 1967).
4. L.D. Braida, N.I. Durlach, R.P. Lippmann, B.L. Hicks, W.M. Rabinowitz, and C.M. Reed, "Matching Speech to Residual Auditory Function — A Review of Past Research," ASHA Monograph (1978).
5. M.A. Mack and B. Gold, "The Discrimination of Pitch in Pulse Trains and Speech," Technical Report 680, Lincoln Laboratory, MIT (April 1984), DTIC AD-A142996/8.
6. S. Roucos and A.M. Wilgus, "The Waveform Segment Vocoder: A New Approach for Very-Low-Rate Speech Coding," International Conf. on Acoustics, Speech, and Signal Processing, Tampa, Florida (March 1985).
7. R.J. McAulay and T.F. Quatieri, "Speech Analysis-Synthesis Based on a Sinusoidal Representation," Technical Report 693, Lincoln Laboratory, MIT (May 1985), DTIC AD-A157023.
8. M.R. Schroeder, "Vocoders: Analysis and Synthesis of Speech," *Proc. IEEE*, **54**, 720-734 (May 1966); reprinted in *Speech Analysis*, R.W. Schafer and J.D. Markel, Eds. (IEEE Press, New York, 1979).
9. S. Seneff, "Speech Transformation System Without Pitch Extraction," Technical Report 541, Lincoln Laboratory, MIT (July 1980), DTIC AD-A097091/3.
10. G. Fairbanks, W.L. Everitt, and R.P. Jaeger, "Method for Time or Frequency Compression-Expansion of Speech," *IRE Trans. Professional Group on Audio*, **AU-2**, 7-12 (January-February, 1954).
11. R.J. Scott and S.E. Gerber, "Pitch Synchronous Time Compression of Speech," *Proc. Conf. Speech Comm. Processing*, pp. 63-65 (April 1972).
12. E.P. Neuberg, "Simple Pitch-Dependent Algorithm for High-Quality Speech Rate Change," presented at the 93rd Meeting Acoust. Soc. Amer. (June 1977); *J. Acoust. Soc. Amer.* (abstract), **61**, suppl. 1 (Spring 1977).
13. S.H. Nawab, T.F. Quatieri and J.S. Lim, "Signal Reconstruction from Short-Time Fourier Transform Magnitude," *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-31**, 986-998 (August 1983).

14. D.W. Griffin and J.S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-32, 236-243 (April 1984).
15. P. Hedelin, "A Tone-oriented Voice-excited Vocoder," *International Conf. Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, (March 1981) p. 205.
16. M.R. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-30, 374 (June 1981).
17. D. Malah, "Time-Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27, 121 (April 1979).
18. L.B. Almeida and F.M. Silva, "Variable-frequency Synthesis: An Improved Harmonic Coding Scheme," *International Conf. Acoustics, Speech, and Signal Processing*, San Diego, California (March 1984) p. 27.6.1.
19. J. Makoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE*, 63, 561 (1975).
20. A.V. Oppenheim and R.W. Schaffer, *Digital Signal Processing*, (Prentice-Hall, Englewood Cliffs, New Jersey, 1983).
21. T.F. Quatieri, "Minimum and Mixed Phase Speech Analysis-Synthesis by Adaptive Homomorphic Deconvolution," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27, 328 (August 1979).
22. S.H. Dantus, "Non-uniform Time-scale Modification of Speech," MS and EE Thesis, MIT, Cambridge, Massachusetts, (1980).
23. K.N. Stevens, "The Role of Rapid Spectrum Changes in the Production and Perception of Speech," *Form and Substance*, (Akademisk Forlag, Copenhagen, 1971).
24. P. Blankenship *et al*, "The Lincoln Digital Voice Terminal System," Technical Note 1975-53, Lincoln Laboratory, MIT (August 1975), DDC AD-017569/5.
25. R.J. McAulay and T.F. Quatieri, "Magnitude-Only Reconstruction Using a Sinusoidal Speech Model," *International Conf. on Acoustics, Speech and Signal Processing*, San Diego, California, 27.6.1 (1984).

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|--|---------------------------|---|
| 1. REPORT NUMBER ESD-TR-85-313 | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle) Speech Transformations Based on a Sinusoidal Representation | | 5. TYPE OF REPORT & PERIOD COVERED Technical Report |
| | | 6. PERFORMING ORG. REPORT NUMBER Technical Report 717 |
| 7. AUTHOR(s) Thomas F. Quatieri and Robert J. McAulay | | 8. CONTRACT OR GRANT NUMBER(s) F19628-85-C-0002 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS Lincoln Laboratory, MIT P.O. Box 73 Lexington, MA 02173-0073 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Program Element Nos. 33401F and 64754F |
| 11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Systems Command Andrews AFB Washington, DC 20334 | | 12. REPORT DATE 16 May 1986 |
| | | 13. NUMBER OF PAGES 68 |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Electronic Systems Division Hanscom AFB, MA 01731 | | 15. SECURITY CLASS. (of this Report) Unclassified |
| | | 15a. DECLASSIFICATION DOWNGRADING SCHEDULE |
| 16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited. | | |
| 17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) | | |
| 18. SUPPLEMENTARY NOTES None | | |
| 19. KEY WORDS (Continue on reverse side if necessary and identify by block number) | | |
| speech | homomorphic deconvolution | time-varying modifications |
| music | time-scale modification | joint time frequency |
| sinusoidal model | pitch modification | modifications |
| analysis/synthesis | frequency scaling | |
| 20. ABSTRACT (Continue on reverse side if necessary and identify by block number) | | |
| <p>In this report, a new speech analysis/synthesis technique is presented which provides the basis for a general class of speech transformations including time-scale modification, frequency scaling, and pitch modification. These modifications can be performed simultaneously and with a time-varying change. The method is based on a sinusoidal representation of the speech production mechanism that has been shown to produce synthetic speech that preserves the waveform shape and is essentially perceptually indistinguishable from the original. Although the analysis/synthesis system originally was designed for single-speaker signals, it is equally capable of modifying nonspeech signals such as music, multiple speakers, marine biologic sounds, and speakers in the presence of interferences such as noise and musical backgrounds.</p> | | |