Time-Scale Modification of Speech Based on Short-Time Fourier Analysis

MICHAEL R. PORTNOFF, MEMBER, IEEE

Abstract—**This paper develops the theoretical basis for time-scale modification of speech based on short-time Fourier analysis. The goal is the development of a high-quality system for changing the apparent rate of articulation of recorded speech, while at the same time preserving such qualities as naturalness, intelligibility, and speaker-dependent features. The results of the theoretical study were used as the framework for the design of a high-quality speech rate-change system that was simulated on a general-purpose minicomputer.**

I. Introduction

THE objective of this paper is to develop the theory of time-scale modification of speech based on short-time Fourier analysis (Part I) and to report on a high-quality rate-change system for speech based on this theory (Part II). The goal is the development of a high-quality system for changing the apparent rate of articulation of recorded speech, while at the same time preserving such qualities as naturalness, intelligibility, and speaker-dependent features. The system must not introduce such objectionable artifacts as "glitches," "burbles," or reverberation, often present in vocoded speech, and the system should be robust to noise, i.e., the performance of the system should not degrade severely if the source speech is corrupted by noise, as might occur in recordings of meetings or courtroom proceedings.

For the most part, nearly all algorithms for changing the rate of speech have been based on the Fairbanks technique [1]-[3] or its refinement [4]-[7]. Basically, these methods change the rate of speech by periodically repeating or discarding sections of the speech waveform. The duration of each section is chosen to be at least as long as one pitch period, but shorter than the length of a phoneme. Although computationally simple, such time-domain techniques introduce discontinuities at the section boundaries which are perceived as "burbling" distortion and overall signal degradation [6].

While time-scale modification of speech based on classical vocoder methods [8]-[13] is an obvious approach, the fundamental consideration in the formulation of a vocoder is bandwidth reduction. Consequently, the vocoders currently available simply do not provide the high level of speech quality and naturalness desired. For example, a large class of vocoders require voiced-unvoiced decisions and pitch extraction: the resulting detection errors introduce artifacts to which the ear is particularly sensitive and which are not tolerable in a high-quality system.

Of the classical vocoders, the only one that does not require voiced-unvoiced decisions and pitch extraction, yet is flexible enough to change the rate of speech, is the phase vocoder [11]-[13]. The phase vocoder is a speech analysis/synthesis system based on short-time Fourier analysis and, unlike most vocoders, can be formulated as an identity system in the absence of parameter modification. Moreover, there is evidence that the ear is less sensitive to errors in the short-time spectrum of an acoustic signal than to errors in the time-domain waveform [14]. Unfortunately, because the theory of short-time Fourier analysis and its application to speech was not well understood, previous applications of the phase vocoder to changing the rate of speech generally did not achieve the quality potentially attainable from this technique.

The approach, here, will be to formulate the problem of time-scale modification of speech in terms of three successive subproblems. The first of these is to appropriately model the speech signal. The second problem is to formulate a mathematical representation for the speech signal based on this model. This representation must have the property that simple time-scaling (interpolation/decimation) of the parameters of the representation corresponds to the desired rate change of the speech signal. The third problem is to design and implement a high-quality analysis/synthesis system based on this representation. This system provides the means to manipulate the parameters of the speech model and should reduce to an identity system in the absence of any parameter modifications. Moreover, since the rate-change system is based on the underlying speech-production model, in order for the system to be robust, it is desirable that the underlying model rely on as few assumptions about the structure of speech as possible.

The treatment of these problems will be based heavily on the framework developed in a companion paper [15] on short-time Fourier analysis of speech and a previous paper [16] on the more general aspects of short-time Fourier analysis. Further, the development here will assume that the reader is familiar with the material in these two references. Starting with a model for the production of speech, [15] formulates a quasi-stationary representation for the speech waveform, applicable to the problem of time-scale modification, and develops the relationship between the parameters of this representation and the short-time Fourier transform (STFT) of the speech. Based on these results, a representation for rate-changed speech will be formulated here, and a modification for the STFT will be derived that can be used to synthesize rate-changed speech.

Manuscript received September 5, 1979; revised December 11, 1980.

The author was with the Department of Electrical Engineering and Computer Science, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139. He is now with the Engineering Research Division, University of California, Lawrence Livermore Laboratory, Livermore, CA 94550.

0096-3518/81/0600-0374$00.75

PART I
II. REPRESENTATION OF RATE-CHANGED SPEECH SIGNALS

Let $x(n)$ denote samples of a given speech signal; the superscript notation $x^{\beta}(n)$ will be used to denote samples of the rate-changed signal with apparent rate of articulation multiplied by the factor β. Thus, $\beta > 1$ corresponds to time-scale compression, and $\beta < 1$ corresponds to time-scale expansion. Such rate changes can be defined in terms of the representation formulated in [15]; however, it is necessary, first, to clarify the notion of linear time scaling of discrete-time signals.

A. Representation of Linearly Time-Scaled Sequences

The representation of a discrete-time signal obtained by linearly time compressing or expanding a given discrete-time signal is fundamental to the development of the theory of time compression and expansion of speech signals. The discrete-time signals that will be of interest fall into two categories: those corresponding to samples of band-limited continuous-time waveforms and those obtained by nonlinear transformations of such signals. A sequence $x(n)$, corresponding to samples of a continuous-time waveform, will be interpreted as samples of the continuous-time waveform $x(t)$, with the sampling interval normalized to unity. Based on this interpretation, it is meaningful to define the sequence $x(\beta n)$, corresponding to samples of $x(\beta t)$ with unity sampling interval, as $x(n)$ linearly time scaled by β. Moreover, if β is rational, so that it can expressed as the quotient of two integers

$$\beta = D/I,$$

and if $x(t)$ is appropriately band limited, then the sequence $x(\beta n)$ can be obtained directly from $x(n)$ by the decimation/interpolation formula [19]

$$x(\beta n) = \sum_{r=-\infty}^{\infty} f(nD - rI)\, x(r) \tag{1}$$

where $f(n)$ is a $1:I$ interpolating filter. A sequence $y(n)$, obtained by a nonlinear transformation of $x(n)$, above, will be linearly time-scaled by β to obtain $y(\beta n)$, by linearly time scaling the underlying band-limited sequence according to (1).

B. Representation of Rate-Changed Voiced Speech

Let $x(n)$ represent samples of a voiced-speech signal modeled, according to [15], as the response of a linear time-varying system $t(n, m)$ excited by a quasi-periodic impulse train $v(n)$ with local pitch period $P(n)$. It was shown that the excitation, $v(n)$, could be modeled as the sum of harmonically related complex exponentials

$$v(n) = \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} \exp\left[jk(\phi(n) + \phi_0)\right] \tag{2}$$

where the phase angle $\phi(n)$ satisfies the recursion relation

$$\phi(n) = \phi(n-1) + \Omega(n) \tag{3}$$

with the initial condition that

$$\phi(0) = 0.$$

$\Omega(n) = 2\pi/P(n)$ is the (instantaneous) fundamental frequency corresponding to $P(n)$ and, from (3), can be obtained from $\phi(n)$ as its first (backward) difference:

$$\Omega(n) = \phi(n) - \phi(n-1). \tag{4}$$

The constant phase angle ϕ_0 was introduced as a convenience so that $\phi(0) = 0$ and so that the time origin will be preserved under rate-change modifications. Since the speech is modeled as quasi-periodic, the pitch is slowly time varying so that

$$P(n_0 + \tau) \approx P(n_0) \qquad \text{for small } |\tau|; \tag{5}$$

and the phase, $\phi(n)$, has a slowly varying slope, $\Omega(n)$, so that

$$\phi(n_0 + \tau) \approx \phi(n_0) + \Omega(n_0)\,\tau \qquad \text{for small } |\tau|. \tag{6}$$

It followed that the voiced speech signal, $x(n)$, could be represented as the linear combination of harmonically related complex exponentials

$$x(n) = \sum_{k=0}^{P(n)-1} c_k(n) \exp\left[jk\phi(n)\right] \tag{7a}$$

where

$$c_k(n) = \frac{1}{P(n)}\, T_2(n, k\Omega(n)) \exp\left[jk\phi_0\right] \tag{7b}$$

and $T_2(n, k\Omega(n))$ denotes the frequency response of the linear time-varying filter $t(n, m)$, evaluated for the kth pitch harmonic.

Based on this harmonic representation, the rate-changed signal, $x^{\beta}(n)$, is modeled by linearly time scaling the time-varying parameters of the filter, $t(n, m)$, and the pitch contour, $\Omega(n)$, by the factor β. Thus, $x^{\beta}(n)$ corresponds to the output of the filter

$$t^{\beta}(n, m) = t(\beta n, m) \tag{8a}$$

with time-varying frequency response

$$T_2^{\beta}(n, \omega) = T_2(\beta n, \omega) \tag{8b}$$

driven by the quasi-periodic unit-sample train $v^{\beta}(n)$, with time-scaled pitch $\Omega(\beta n)$. Moreover, it will be shown that the excitation, $v^{\beta}(n)$, can be expressed as

$$v^{\beta}(n) = \frac{1}{P(\beta n)} \sum_{k=0}^{P(\beta n)-1} \exp\left[jk(\phi(\beta n)/\beta + \phi_0)\right]. \tag{9}$$

To show that (9) represents the desired "rate-changed" unit-sample train, we must show that its pitch is the time-scaled instantaneous frequency $\Omega(\beta n)$, and the phase of its kth harmonic at $n = 0$ is $k\phi_0$. Since $\phi(0) = 0$, the latter condition is obvious. To show the former, linearly time scale (6) in both n_0 and τ to obtain

$$\phi(\beta(n_0 + \tau)) \approx \phi(\beta n_0) + \Omega(\beta n_0)\,\beta\tau$$

or

$$\phi(\beta(n_0 + \tau))/\beta \approx \phi(\beta n_0)/\beta + \Omega(\beta n_0)\,\tau. \tag{10}$$

Setting $\tau = -1$ gives

$$\Omega(\beta n) \approx \phi(\beta n)/\beta - \phi(\beta(n-1))/\beta \tag{11}$$

which shows that the fundamental frequency of $v^{\beta}(n)$ is indeed $\Omega(\beta n)$.

From (9) and (10), $v^{\beta}(n)$ can be represented locally as

$$v^{\beta}(n_0+\tau) \approx \frac{1}{P(\beta n_0)} \sum_{k=0}^{P(\beta n_0)-1} \cdot \exp\left[jk(\phi(\beta n_0)/\beta + \Omega(\beta n_0)\,\tau + \phi_0)\right]. \quad (12)$$

Using (12) in the superposition sum and paralleling the development of [15] yields the harmonic representation for rate-changed voiced speech

$$x^{\beta}(n) = \sum_{k=0}^{P(\beta n)-1} c_k(\beta n) \exp\left[jk\phi(\beta n)/\beta\right]. \quad (13)$$

It is important to note that the instantaneous phase, $\phi(n)$, in the above equations is the "unwrapped phase" angle [15] that satisfies the recursion (6), not its principal value. This distinction is important because $1/\beta$ is not, in general, an integer; thus, while an integer multiple of 2π added to the argument of the exponential in (13) is invisible, an integer multiple of $2\pi/\beta$ is not.

C. Representation of Rate-Changed Unvoiced Speech

If $x(n)$ now represents samples of an unvoiced-speech signal, then $x^{\beta}(n)$ is modeled as the output of the filter

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

driven by the white-noise process $u(n)$. Consequently, if the original speech $x(n)$ is characterized by its time-varying power spectrum [15]

$$S_x(n,\omega) = \sigma_u^2 \left|T_2(n,\omega)\right|^2 \quad (14)$$

and autocorrelation function [15]

$$R_x(n,\tau) = E\{x(n+\tau)\,x^*(n)\} \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(n,\omega)\exp\left[j\omega\tau\right]\,d\omega \quad (15)$$

then the rate-changed speech is characterized by the time-varying power spectrum

$$S_x^{\beta}(n,\omega) = \sigma_u^2\left|T_2(\beta n,\omega)\right|^2 = S_x(\beta n,\omega) \quad (16)$$

and autocorrelation function

$$R_x^{\beta}(n,\tau) \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x^{\beta}(n,\omega)\exp\left[j\omega\tau\right]\,d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(\beta n,\omega)\exp\left[j\omega\tau\right]\,d\omega \approx R_x(\beta n,\tau). \quad (17)$$

Thus, changing the rate of an unvoiced-speech signal preserves the local statistics of the signal, but linearly time scales the time-varying parameters of the statistics.

III. Theory of Time-Scale Modification of Speech Based on Short-Time Fourier Analysis

Now that the representation for a rate-changed speech signal has been formulated, in order to actually change the rate of a given speech signal, the parameters of this representation must be estimated, appropriately modified, and used to synthesize the rate-changed signal. We now consider short-time Fourier analysis as a method of accomplishing this objective.

The approach is to modify the STFT of the given speech signal such that the signal synthesized from the modified transform corresponds to the desired rate-changed speech. The modification will be formulated, first, for voiced speech. Then, the same modification will be shown to produce rate changes of unvoiced speech as well. Consequently, the problem of distinguishing voiced from unvoiced speech is avoided.

Heuristically, the use of short-time Fourier analysis for time-scale modification of speech is based on the idea of mapping the one-dimensional speech signal to a two-dimensional signal that is a function of time and frequency, such that "temporal features" of the speech appear as functions of the time variable and "spectral features" appear as functions of the frequency variable. The STFT is then appropriately modified such that it is compressed or expanded along the time axis but not the frequency axis. This modification will require decimation and interpolation of the STFT in the time direction *and* a nonlinear modification, affecting its phase.

The distinction between temporal and spectral features of an audio signal is, in general, a fuzzy notion, based on the nature of the signal processing performed by the auditory system. Here, the distinction between temporal and spectral features for the class of speech signals is defined, based on the speech-production model and the corresponding signal representation. Specifically, the dependence on n of both the time-varying frequency response $T_2(n,\omega)$ and the pitch $\Omega(n)$, in the speech-production model, is assumed to represent the temporal characteristics of the speech, and the dependence on ω of $T_2(n,\omega)$ and the value of $\Omega(n)$ are assumed to represent the spectral characteristics of the speech. Unfortunately, the functions $T_2(n,\omega)$ and $\Omega(n)$ cannot, in general, be exactly determined by observing the speech waveform.

The technique of short-time Fourier analysis provides a means for estimating and modifying these functions. Expressing the STFT of $x(n)$ as the convolution in time

$$X_2(n,\omega) = \sum_{m=-\infty}^{\infty} h(n-m)\,x(m)\exp\left[-j\omega m\right] \quad (18)$$

suggests that features of $x(n)$ that change slowly as a function of time over the duration of $h(n)$ appear in the STFT as functions of the time variable, whereas features that change rapidly as a function of time over the duration of $h(n)$ appear as a function of the frequency variable.

A. Short-Time Fourier Analysis of Voiced Speech

According to the harmonic representation for voiced speech, a voiced-speech signal is modeled as a linear combination of harmonically related complex exponentials; on a short-time basis, each is a narrow-band signal with slowly varying complex amplitude and instantaneous frequency. If the STFT is interpreted as the output of a filter-bank spectrum analyzer, and if the analysis filter, $h(n)$, is designed for narrow-band analysis so that its bandwidth is sufficiently wide to pass any one of the individual harmonic components of the speech, yet narrow enough to pass at most one such component, then the STFT of a voiced-speech signal, $x(n)$, can be expressed in terms of the parameters of the harmonic representation (7) for $x(n)$. Specifically, the (narrow-band) STFT of a voiced-speech signal was shown to be [15]

$$X_2(n,\omega) = \begin{cases} c_k(n)\, H(k\Omega(n) - \omega) \exp\left[j(k\phi(n) - \omega n)\right] \\ \qquad \text{for } \left|\omega - k\Omega(n)\right| < \omega_h \\ 0 \qquad \text{otherwise} \end{cases} \tag{19}$$

where ω_h denotes the cutoff frequency of the analysis filter $H(\omega)$.

Expressing $X_2(n, w)$ in polar form as

$$X_2(n,\omega) = A(n,\omega) \exp\left[j\theta(n,\omega)\right] \tag{20}$$

where

$$A(n,\omega) = \left|X_2(n,\omega)\right|$$

and

$$\theta(n,\omega) = \arg\left[X_2(n,\omega)\right]$$

and investigating the structure of $A(n,\omega)$ and $\theta(n,\omega)$ will suggest how to modify $X_2(n,\omega)$ in order to obtain rate-changed speech. The magnitude of the STFT, obtained from (19) as

$$A(n,\omega) = \left|c_k(n)\right| \left|H(k\Omega(n) - \omega)\right| \quad \text{for } \left|\omega - k\Omega(n)\right| < \omega_h, \tag{21}$$

is a slowly varying function of n because both $c_k(n)$ and $\Omega(n)$ are slowly varying functions of n. Referring again to (19), the phase of the STFT can be expressed as the sum of two components

$$\theta(n,\omega) = \alpha(n,\omega) + \vartheta(n,\omega) \tag{22a}$$

where

$$\alpha(n,\omega) = \arg\left[c_k(n)\right] + \arg\left[H(k\Omega(n) - \omega)\right] \tag{22b}$$

and

$$\vartheta(n,\omega) = k\phi(n) - \omega n \tag{22c}$$

for

$$\left|\omega - k\Omega(n)\right| < \omega_h.$$

The component $\alpha(n,\omega)$ contributes a slowly time-varying phase and is called the "phase-modulation" component. The other component, $\vartheta(n,\omega)$, can be expressed as [15]

$$\vartheta(n,\omega) = \begin{cases} \sum_{r=1}^{n} \{k\Omega(r) - \omega\} & \text{for } n > 0 \\ 0 & \text{for } n = 0 \\ \sum_{r=0}^{n+1} -\{k\Omega(r) - \omega\} & \text{for } n < 0 \end{cases}$$

$$\text{where } \left|\omega - k\Omega(n)\right| < \omega_h \tag{23}$$

and satisfies the recursion relation

$$\vartheta(n,\omega) = \vartheta(n-1,\omega) + k\Omega(n) - \omega. \tag{24}$$

Equivalently,

$$\vartheta(n,\omega) = \vartheta(n-1,\omega) + \Omega(n,\omega) \tag{25}$$

where

$$\Omega(n,\omega) = k\Omega(n) - \omega \tag{26}$$

with k such that

$$\left|\Omega(n,\omega)\right| < \omega_h. \tag{27}$$

Because

$$\Omega(n,\omega) = \vartheta(n,\omega) - \vartheta(n-1,\omega), \tag{28}$$

$\Omega(n,\omega)$ is referred to as the "instantaneous frequency" of the STFT. Furthermore, because $\Omega(n)$ is a slowly varying function of n, $\Omega(n,\omega)$ is also a slowly varying function of n. Thus, $\vartheta(n,\omega)$ can be expressed locally as

$$\vartheta(n_0 + \tau, \omega) \approx \vartheta(n_0,\omega) + \Omega(n_0,\omega)\,\tau \tag{29}$$

for $(n_0 + \tau)$ in the neighborhood about n_0 for which the speech is modeled as periodic. Therefore, $\vartheta(n,\omega)$ will be referred to as the "linear-phase," or "frequency-modulation," component of the STFT.

B. Synthesis of Rate-Changed Voiced Speech

The previous section provides the framework necessary to define the modified STFT from which rate-changed speech can be synthesized. Let $x(n)$ denote a given voiced-speech signal, and let $Y_2(n,\omega)$ denote the modified STFT from which the rate-changed speech, $x^{\beta}(n) = y(n)$, is to be synthesized. It will be shown that $Y_2(n,\omega)$ is given by

$$Y_2(n,\omega) = A(\beta n,\omega) \exp\left[j(\alpha(\beta n,\omega) + \vartheta(\beta n,\omega)/\beta)\right], \tag{30}$$

i.e., both the magnitude and phase of the short-time Fourier transform are linearly time scaled by β, and the frequency-modulation component of the phase is divided by β. We will see that the combined effect of these modifications is to linearly time-scale the magnitude and instantaneous frequency of the STFT. Remember that $\vartheta(n,\omega)$ is the "unwrapped" phase,

given by (23), and not its principal value. This distinction is important whenever $1/\beta$ is not an integer.

To show that $y(n)$ is the desired rate-changed speech $x^{\beta}(n)$, the definitions (21) and (22) of the magnitude and phase of the STFT are substituted into the modified STFT (30) to obtain

$$Y_2(n,\omega)=\begin{cases} c_k(\beta n)\,H(k\Omega(\beta n)-\omega)\exp\left[j(k\phi(\beta n)/\beta-\omega n)\right] \\ \qquad \text{for } \left|\omega-k\Omega(\beta n)\right|<\omega_h \\ 0 \qquad \text{otherwise.} \end{cases} \tag{31}$$

Because the shifted and weighted images of $H(\omega)$ are assumed to be nonoverlapping, $Y_2(n,\omega)$ can be written as the sum of these images.

$$Y_2(n,\omega)=\sum_{k=0}^{P(\beta n)-1} c_k(\beta n)\,H(k\Omega(\beta n)-\omega)\cdot\exp\left[j(k\phi(\beta n)/\beta-\omega n)\right]. \tag{32}$$

The synthesized signal, $y(n)$, is generated according to the short-time Fourier synthesis formula

$$y(n)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{r=-\infty}^{\infty} f(n-r)\,Y_2(r,\omega)\exp\left[j\omega n\right]d\omega \tag{33}$$

where $f(n)$ denotes the synthesis filter [16]. Thus, substituting $Y_2(n,\omega)$, given by (32), into (33) gives

$$\begin{aligned} y(n)&=\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{r=-\infty}^{\infty}\sum_{k=0}^{P(\beta r)-1} f(n-r)\,c_k(\beta r)\,H(k\Omega(\beta r)-\omega) \\ &\quad\cdot\exp\left[j(k\phi(\beta r)/\beta-\omega r)\right]\exp\left[j\omega n\right]d\omega \\ &=\sum_{r=-\infty}^{\infty}\sum_{k=0}^{P(\beta r)-1} f(n-r)\,c_k(\beta r)\exp\left[jk\phi(\beta r)/\beta\right] \\ &\quad\cdot\frac{1}{2\pi}\int_{-\pi}^{\pi} H(k\Omega(\beta r)-\omega)\exp\left[j\omega(n-r)\right]d\omega. \end{aligned}$$

Integrating over ω

$$y(n)=\sum_{r=-\infty}^{\infty}\sum_{k=0}^{P(\beta r)-1} f(n-r)\,h(r-n)\,c_k(\beta r)\cdot\exp\left[jk\{\phi(\beta r)/\beta+\Omega(\beta r)\,(n-r)\}\right]$$

and using the local representation (10) for $\phi(\beta n)/\beta$ gives

$$y(n)=\sum_{r=-\infty}^{\infty}\sum_{k=0}^{P(\beta r)-1} f(n-r)\,h(r-n)\,c_k(\beta r)\cdot\exp\left[jk\phi(\beta n)/\beta\right]. \tag{34}$$

Assuming that the time-scaled pitch, $\Omega(\beta n)$, is constant over any interval less than the duration of $\{f(n)\,h(-n)\}$ so that

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

or equivalently (with $r=n-\tau$)

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

gives

$$y(n)=\sum_{k=0}^{P(\beta n)-1}\sum_{r=-\infty}^{\infty} f(n-r)\,h(r-n)\,c_k(\beta r)\exp\left[jk\phi(\beta n)/\beta\right]. \tag{35}$$

The summation over r in (35) is just the convolution of the time-scaled harmonic amplitudes, $c_k(\beta n)$, with the filter $\{f(n)\,h(-n)\}$. The frequency response of this composite filter,

$$\sum_{n=-\infty}^{\infty} f(n)\,h(-n)\exp\left[-j\omega n\right]=\frac{1}{2\pi}\int_{-\pi}^{\pi} F(\omega+\psi)\,H(\psi)\,d\psi,$$

has a bandwidth on the order of the sum of the bandwidths of $F(\omega)$ and $H(\omega)$. If this bandwidth is wider than the bandwidth of each of the $c_k(\beta n)$'s, then the $c_k(\beta n)$'s are passed by the composite filter with negligible distortion. Consequently

$$y(n)=\sum_{k=0}^{P(\beta n)-1} c_k(\beta n)\exp\left[jk\phi(\beta n)/\beta\right] \tag{36}$$

which is the desired rate-transformed speech signal, $x^{\beta}(n)$, given by (13).

To conclude this section, the assumptions used (both here and in [15]) in formulating the preceding theory of time-scale modification of voiced speech based on short-time Fourier analysis will now be summarized. First, for the analysis of voiced speech, the analysis window, $h(n)$, is assumed to be sufficiently narrow in time to resolve the temporal features of the speech and also sufficiently narrow in frequency to resolve the spectral features of the speech. On the one hand, to resolve temporal changes in the pitch, the duration of $h(n)$ is assumed to be short enough that the pitch can be regarded as constant over any time interval with duration less than that of $h(n)$. Furthermore, to resolve temporal changes in the speech due to changes in the vocal-tract geometry, the bandwidth of $H(\omega)$ is assumed to be much greater than each of the bandwidths of the individual speech harmonics, $C_k(\omega)$. This assumption is roughly equivalent to the assumption that the harmonic amplitudes $c_k(n)$ can be regarded as constant over any interval shorter than the duration of $h(n)$. On the other hand, to resolve spectral features of the speech, i.e., the value of the pitch and the shape of the spectral envelope, the bandwidth of $H(\omega)$ must be less than the value of the pitch, $\Omega(n)$.

For the synthesis of rate-changed speech, the composite window $f(n)\,h(-n)$ is assumed to be sufficiently narrow in time to resolve the temporal features of the rate-changed speech. Specifically, the duration of the composite window $f(n)\,h(-n)$ is assumed to be short enough that the time-scaled pitch, $\Omega(\beta n)$, can be regarded as constant over any interval with duration less than $f(n)\,h(-n)$, and the bandwidth of $f(n)\,h(-n)$ is assumed to be much greater than each of the bandwidths of the time-scaled harmonic amplitudes $c_k(\beta n)$.

C. Short-Time Fourier Analysis of Unvoiced Speech

Unvoiced speech is modeled as a quasi-stationary random process characterized by its second moments. It was shown that a convenient characterization for the STFT of unvoiced speech was the modified correlation function [15] defined by

$$K_x(n,\omega,\tau,\epsilon) = E\left\{X_2\left(n+\tau,\omega-\frac{\epsilon}{2}\right)X_2^*\left(n,\omega+\frac{\epsilon}{2}\right)\right\} \cdot \exp\left[-j\left(n+\frac{\tau}{2}\right)\epsilon\right] \tag{37}$$

which can be expressed in terms of the time-varying power spectrum of $x(n)$ and the analysis filter $H(\omega)$ as

$$K_x(n,\omega,\tau,\epsilon) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(n,\omega+\varphi)\,H\left(\varphi+\frac{\epsilon}{2}\right)H^*\left(\varphi-\frac{\epsilon}{2}\right) \cdot \exp\left[j\varphi\tau\right]\,d\varphi. \tag{38}$$

Because the analysis filter $h(n)$ is chosen to be narrow in both time and frequency, $K_x(n,\omega,\tau,\epsilon)$ can be approximated by the first few terms of a two-dimensional power series in τ and ϵ:

$$K_x(n,\omega,\tau,\epsilon) = J_x(n,\omega)\left\{1-\tfrac{1}{2}\left[D_h^2\tau^2+2\mu_h\tau\epsilon+d_h^2\epsilon^2\right]+\cdots\right\} \tag{39}$$

where $J_x(n,\omega)$ is the smoothed spectrum

$$J_x(n,\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(n,\omega+\varphi)\,|H(\varphi)|^2\,d\varphi, \tag{40}$$

D_h is the rms bandwidth of $H(\omega)$, d_h is the rms duration of $h(n)$, and μ_h is a real number that vanishes if $h(n)$, or $H(\omega)$, or both are real [15]. The expansion (39) also assumes that the analysis filter has been defined with suitable shifts in time and frequency so that its first moments in both time and frequency vanish. This representation will be useful for determining the power spectrum of unvoiced speech synthesized from the nonlinearly modified STFT (30), originally defined to effect rate changes of voiced speech.

Because the bandwidth of the analysis filter is narrow compared with the sampling frequency of the speech, $X_2(n,\omega)$ is a narrow-band low-pass random process in n for each value of ω. Consequently, the STFT of unvoiced speech, expressed in polar form, exhibits slowly-varying amplitude and instantaneous frequency similar to the STFT of voiced speech; however, it possesses no underlying harmonic structure, as does voiced speech. The phase components $\alpha(n,\omega)$ and $\vartheta(n,\omega)$ for unvoiced speech will be defined as the values calculated by the same estimator used to calculate these quantities for voiced speech (one such estimator will be discussed in Part II).

D. The Time-Varying Power Spectrum of the Synthesized Signal

Let $Y_2(n,\omega)$ denote a short-time or modified short-time Fourier transform, and let $y(n)$ denote the signal synthesized from $Y_2(n,\omega)$ according to the short-time Fourier synthesis formula (33). If $K_y(n,\omega,\tau,\epsilon)$ denotes the modified correlation function for $Y_2(n,\omega)$, and if the bandwidth of the synthesis filter is wider than the bandwidth of $K_y(n,\omega,\tau,\epsilon)$ in the n direction, then according to the derivation given in Appendix I, the time-varying power spectrum of $y(n)$ is given by

$$S_y(n,\omega) = \left(\frac{1}{2\pi}\right)^2\int_{-\pi}^{\pi}\int_{-\epsilon_0}^{\epsilon_0}|F(\omega-\varphi)|^2 \cdot \sum_{r=-\infty}^{\infty} K_y(n,\varphi,r,\epsilon)\exp\left[-j(\omega-\varphi)r\right]\,d\epsilon\,d\varphi \tag{41}$$

where the limits on the inner integral are given in terms of $\epsilon_0(\omega) = 2(\pi - |\omega|)$.

IV. Synthesis of Rate-Changed Unvoiced Speech

Suppose that the STFT of a given unvoiced speech signal, $x(n)$, has been expressed in the form

$$X_2(n,\omega) = A(n,\omega)\exp\left[j(\alpha(n,\omega)+\vartheta(n,\omega))\right] \tag{42}$$

where $A(n,\omega)$, $\alpha(n,\omega)$, and $\Omega(n,\omega) = \vartheta(n,\omega) - \vartheta(n-1,\omega)$ are slowly varying functions of n. Define the modified STFT $Y_2(n,\omega)$, just as in the voiced-speech case, according to (30), by linearly time-scaling the magnitude and phase of the STFT by β, and dividing the linear-phase component by β. This section argues that the time-varying power spectrum of the synthesized signal $y(n)$ is approximately the same as the time-varying power spectrum of the desired rate-changed signal $x^{\beta}(n)$.

The determination of the time-varying power spectrum of the synthesized signal $y(n)$ is a nonlinear stochastic-processes problem that, in general, has no closed form solution in terms of elementary functions. In order to make the problem more tractable and to gain some insight into the effects of the nonlinear modification (30) of the STFT, several simplifying assumptions will be made.

1) The underlying random process, $u(n)$, in the speech production model is Gaussian.

2) The spectral resolution of the filter-bank analyzer is sufficient to resolve the unvoiced-speech spectrum, $S_x(n,\omega)$, for the purpose of calculating the moments of $K_x(n,\omega,\tau,\epsilon)$. This assumption means, for example, that the average bandwidths and center frequencies of the filter-bank output signals are determined principally by the shape of the analysis filter, $H(\omega)$, rather than by fine structure in the spectrum $S_x(n,\omega)$. This assumption is reasonable because of the smoothness of the unvoiced-speech spectrum [22].

3) The slowly varying phase component $\alpha(n,\omega)$ of the STFT will be neglected in the computation of the moments of

$K_y(n, \omega, \tau, \epsilon)$. Therefore, the approximation

$$K_y(n, \omega, \tau, \epsilon) \approx E\left\{A\left(\beta(n+\tau), \omega - \frac{\epsilon}{2}\right)A\left(\beta n, \omega + \frac{\epsilon}{2}\right) \cdot \exp\left[j\left(\theta\left(\beta(n+\tau), \omega - \frac{\epsilon}{2}\right) - \theta\left(\beta n, \omega + \frac{\epsilon}{2}\right)\right)\Big/\beta\right]\right\} \cdot \exp\left[-j\left(n + \frac{\tau}{2}\right)\epsilon\right] \quad (43)$$

will be used to calculate the moments of $K_y(n, \omega, \tau, \epsilon)$.

Based on these assumptions, the modified autocorrelation function (43) can be expanded in a power series in τ and ϵ with the coefficients of this series expressed in terms of the moments of $K_x(n, \omega, \tau, \epsilon)$. This calculation is sketched in Appendix II to give

$$K_y(n, \omega, \tau, \epsilon) = J_x(\beta n, \omega) \cdot \left\{1 - \frac{\gamma^2}{2}\left[D_h^2\tau^2 + 2\mu_h\tau(\epsilon/\beta) + d_h^2(\epsilon/\beta)^2\right] + \cdots\right\} \quad (44a)$$

where

$$\gamma^2 = \tfrac{1}{2}(1 + \beta^2). \quad (44b)$$

Comparing (39) and (44) shows that the modified correlation function for $Y_2(n, \omega)$ is given, to second order in τ and ϵ, by

$$K_y(n, \omega, \tau, \epsilon) \approx K_x(\beta n, \omega, \gamma\tau, \gamma\epsilon/\beta). \quad (45)$$

The time-varying power spectrum of the synthesized rate-changed speech is given according to (41) as

$$S_y(n, \omega) = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\epsilon_0}^{\epsilon_0} |F(\omega - \varphi)|^2 \cdot \sum_r K_x(\beta n, \varphi, \gamma r, \gamma\epsilon/\beta) \cdot \exp\left[-jr(\omega - \varphi)\right] d\epsilon\, d\varphi. \quad (46)$$

Now, substituting the expression (38) for $K_x(n, \omega, \tau, \epsilon)$ into the expression (46) gives

$$S_y(n, \omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(\beta n, \omega + \varphi)\, G_\beta(\varphi)\, d\varphi \quad (47a)$$

where

$$G_\beta(\omega) = [\beta/\gamma(1-\gamma)] \cdot |F(\omega\gamma/(1-\gamma))|^2\, H_1(\omega/(1-\gamma)) \quad (47b)$$

and

$$H_1(\omega) = \frac{1}{2\pi}\int_{-\epsilon_0}^{\epsilon_0} H\left(\omega + \frac{\epsilon}{2}\right)H^*\left(\omega - \frac{\epsilon}{2}\right)d\epsilon = \frac{1}{\pi}\int_{-\omega_h}^{\omega_h} H(2\omega - \varphi)\, H^*(\varphi)\, d\varphi. \quad (47c)$$

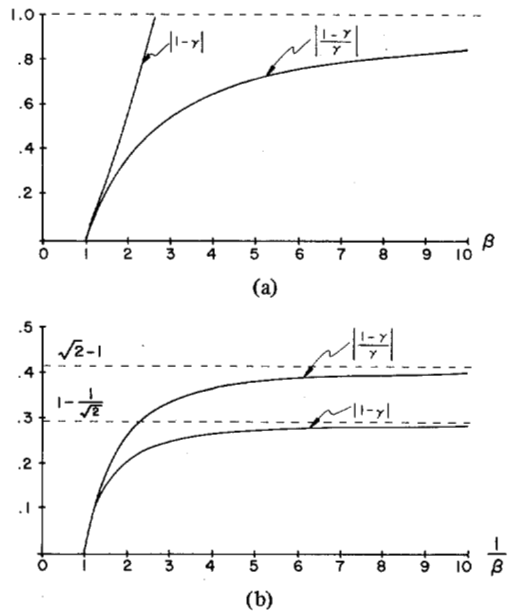

Fig. 1. (a) Scale factors for width of spectral-smoothing function for time-scale compression of unvoiced speech. (b) Scale factors for width of spectral-smoothing function for time-scale expansion of unvoiced speech.

Thus, $S_y(n, \omega)$ corresponds to the ideal spectrum of the rate-changed speech, $S_x(\beta n, \omega)$, smoothed by the function $G_\beta(\omega)$, whose width depends on the rate-changed scale factor β.

The width of $G_\beta(\omega)$ can be shown to be sufficiently narrow that the spectrum, $S_y(n, \omega)$ for the synthesized speech is an acceptably close approximation to the ideal spectrum, $S_x(\beta n, \omega)$, for the rate-changed speech. The width of $G_\beta(\omega)$ is approximately the smaller of the widths of $|F(\omega\gamma/(1-\gamma))|^2$ and $H_1(\omega/(1-\gamma))$. In practice, the bandwidth of $F(\omega)$ is chosen to be approximately equal to D_h, the bandwidth of $H(\omega)$. Furthermore, from (47c), the bandwidth of $H_1(\omega)$ is seen to be approximately equal to D_h, also. Consequently, the width of the spectral smoothing function $G_\beta(\omega)$ is on the order of the smaller of $|(1-\gamma)/\gamma|\, D_h$ and $|(1-\gamma)|\, D_h$.

Fig. 1(a) shows plots of $|(1-\gamma)/\gamma|$ and $|1-\gamma|$ versus β for $\beta \geq 1$, corresponding to time-scale compression. Here,

$$|(1-\gamma)/\gamma| \leq |1-\gamma|$$

and

$$|(1-\gamma)/\gamma| < 1.$$

Thus, for time-scale compression, the width of $G_\beta(\omega)$ is less than $|(1-\gamma)/\gamma|\, D_h$, which is less than D_h.

Fig. 1(b) shows plots of $|(1-\gamma)/\gamma|$ and $|1-\gamma|$ versus $1/\beta$ for $\beta \leq 1$, corresponding to time-scale expansion. Here,

$$|1-\gamma| \leq |(1-\gamma)/\gamma|$$

and

$$|1-\gamma| < 1 - 2^{-1/2} \approx 0.3.$$

Thus, for time-scale expansion, the width of $G_\beta(\omega)$ is less than $|1-\gamma|\, D_h$, which is less than $0.3\, D_h$. In practice, the bandwidth D_h of the analysis filter is on the order of 100 Hz. Thus, the spectral smearing of the synthesized speech is on the order of 100 Hz for time-scale compression and 30 Hz for time-scale

expansion. The quality of real speech processed by a system based on this formulation confirms that the degree of spectral smearing of the unvoiced portions of the speech is acceptable.

PART II

V. DESIGN AND SIMULATION OF A DIGITAL RATE-CHANGE SYSTEM FOR SPEECH

The theoretical results of Part I will now be applied to the design of a rate-change system for speech that was simulated on a general-purpose minicomputer. The system consists of an analysis/synthesis system that represents a given sampled speech signal in terms of its discrete (sampled) STFT [16] and a parameter modification system that appropriately modifies the STFT to effect the desired rate change in the synthesized speech. In the absence of any parameter modification the analysis/synthesis system is designed to be an identity system.

In order for the time-scale modification system to be realizable on a digital processor, the STFT must be represented by a finite number of frequency samples. Moreover, to make the amount of computation tractable, the STFT must be decimated in time (down-sampled) as well. Consequently, in addition to the design of the analysis and synthesis filters based on the requirements of the previous discussions, a number of new issues arise as a result of the sampled-transform representation.

The remainder of this paper addresses the issues related to implementing a time-scale modification system for speech based on discrete short-time Fourier analysis and also introduces a procedure for estimating and modifying the FM component of the phase of the STFT. Section VI raises the issue of how finely the STFT must be sampled to be adequately represented for the application of time-scale modification of speech. Section VII discusses the analysis/synthesis system and, in particular, the considerations for designing the analysis and synthesis filters and choosing the temporal and spectral sampling intervals. Because the modification of the STFT used to effect rate changes of speech does not, in general, preserve the structure of the STFT of an *arbitrary* signal, the results of [16] concerning the design of the analysis and synthesis filters and the sampling-rate requirements for the discrete STFT cannot be applied directly. These issues must, therefore, be reconsidered for the particular problem of time-scale modification of speech. Section VIII discusses the modification of the STFT to effect rate changes in the synthesized speech. This modification consists of two operations: linearly time-scaling the STFT by the rate-change scale factor β and dividing the FM component of the phase of the STFT by the factor β. Because these two operations each affect the bandwidth of the STFT (considered as a time sequence), the order in which they are implemented becomes important to avoid aliasing of the sampled STFT. The linear time-scaling operation will, generally, be implemented using conventional digital interpolation/decimation techniques [19]. Under certain circumstances, however, it may be computationally more efficient to effect the linear time-scaling implicitly by performing the short-time Fourier analysis and synthesis assuming different temporal sampling rates. This issue will be discussed in Section VIII-D. The phase modification will require removing jumps of π and 2π in the phase curve (as a function of time) and will introduce a further consideration in choosing the temporal sampling rate for the STFT; namely, that unwrapping the phase of the STFT requires samples of the principal value of the phase of the STFT at a rate of at least twice that of the Nyquist rate for the STFT. The paper concludes with a discussion of the simulation on a DEC PDP-11/50 minicomputer.

VI. TIME-SCALE MODIFICATION OF SPEECH BASED ON DISCRETE SHORT-TIME FOURIER ANALYSIS

Because the modification of the STFT required to effect rate changes of speech is nonlinear, its application to samples of the STFT will result in some degree of time-domain aliasing in the time signal synthesized from those samples. Consequently, the result of the rate-change modification applied to the discrete STFT is not, in general, equivalent to the result of that modification applied to the (nonsampled) STFT. However, since the discrete STFT becomes the STFT, in the limit as the temporal and spectral sampling intervals approach unity and zero, respectively, the question arises: how fast must the discrete STFT be sampled in time and frequency so that the result of processing the samples of the STFT is "sufficiently close" to the result of processing the STFT itself? The nonlinear nature of the processing makes this question difficult to answer in general. This section proposes an answer by obtaining conditions such that 1) with no parameter modification, the overall system is an identity system; 2) for the particular case of quasi-periodic voiced speech, the results of modifying the STFT and the discrete STFT are the same; and 3) as the speech deviates from this model, either because the speech itself does not fit the quasi-periodic model, or because it has been corrupted by noise, the system is required to behave gracefully, i.e., the system is to be robust.

In order for the short-time Fourier analysis/synthesis system to be an identity system, in the absence of parameter modification, the analysis and synthesis filters must satisfy the condition derived in [16]. This condition is that the effective filter $w(n, m)$, defined by

$$w(n, m) = \sum_{s=-\infty}^{\infty} f(n - sR)\, h(sR + m - n) \tag{48}$$

must have the property

$$w(n, pM) = 0 \qquad \text{for all } n, \tag{49}$$

where R is the temporal sampling interval and $\Omega_M = 2\pi/M$ is the spectral sampling interval.

For a voiced speech signal, $x(n)$, under quasi-stationary analysis, the modified STFT, $Y_2(n, \omega)$, given by (32), is the STFT of the desired rate-changed signal, $x^\beta(n)$, obtained using the same analysis filter, $h(n)$. Thus, for quasi-periodic voiced speech, condition (49) guarantees that the speech synthesized from the samples $Y_2(sR, k\Omega_M)$ will be the desired rate-changed speech $x^\beta(n)$.

If condition (49) is the sole criterion for designing the analysis/synthesis system, we find that, in the absence of parameter modification, the simulated system is, indeed, an identity system, and for quasi-periodic speech (i.e., steady-state vowels) the rate-changed speech is high quality. As the speech deviates from the quasi-periodic model, however, an objectionable amount of reverberation becomes apparent in

the rate-changed speech. This reverberation is, in fact, time-domain aliasing, resulting because the nonlinear operation required to effect the desired rate change of the speech is performed on the *samples* of the STFT, $X_2(sR, k\Omega_M)$, rather than on the STFT, $X_2(n, \omega)$, itself. Thus, the reverberation can be reduced by increasing the number of frequency samples, M. In particular, if the effective window, $w(n, m)$, defined by (48) has finite duration in m, then the reverberation in the processed speech is effectively eliminated by choosing the number of frequency samples, M, to be equal to or greater than this duration.

The modification of the STFT requires an estimate of certain parameters of the speech model embedded in the STFT. In order to extract these parameters from the sampled STFT, $X_2(sR, k\Omega_M)$, the temporal sampling interval, R, must be small enough to prevent frequency aliasing of $X(\psi, k\Omega_M)$ in ψ [16]. Since the analysis filter $h(n)$ is chosen to have low-pass characteristics, the interpretation of the STFT as the output of $h(n)$ requires that the sampling frequency, $2\pi/R$, be greater than twice the cutoff frequency of $H(\omega)$ in order to prevent frequency aliasing.

VII. Design of the Analysis/Synthesis System

The analysis/synthesis system represents a given speech signal in terms of its complex discrete STFT. Such a system, when applied to speech coding, is generally referred to as a phase vocoder [11]. With appropriate analysis and synthesis filters and proper sampling rates, the samples of the STFT are related to the parameters of the harmonic representation for voiced speech and the time-varying spectrum for unvoiced speech as discussed in Part I, thus providing a tool for manipulating these parameters.

For a given input signal, $x(n)$, the short-time Fourier analyzer calculates samples of the STFT

$$X_2(sR, k\Omega_M) = \sum_{m=-\infty}^{\infty} h(sR - m)\, x(m) \exp\,[-j\Omega_M km] \tag{50}$$

where $\Omega_M = 2\pi/M$ represents the spacing of the samples in frequency and R the spacing of the samples in time. The analyzer calculates the samples $X_2(sR, k\Omega_M)$ efficiently with the FFT algorithm using the techniques discussed in [10], [12], [17], [18].

The short-time Fourier synthesizer calculates samples of the time sequence, $y(n)$, from samples of the modified STFT, $Y_2(sR, k\Omega_M)$, according to the formula

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} f(n - sR)\, Y_2(sR, k\Omega_M) \exp\,[j\Omega_M nk]. \tag{51}$$

The synthesis is also implemented efficiently using the FFT algorithm as described in [10], [12], [17], [18].

A. Design of the Analysis Filter

As discussed in Part I, the analysis filter, $h(n)$, must have a bandwidth sufficiently narrow to resolve the "spectral features" of the speech, i.e., its bandwidth must be less than the frequency spacing between the harmonics of the voiced portions of the speech, and small enough that the time-varying power spectrum of the unvoiced portions of the speech is smooth over any frequency interval smaller than this bandwidth. In addition, $H(\omega)$ must be sufficiently broad-band to pass the temporal features of the speech. A final requirement on the analysis filter is dictated by the practical consideration of keeping the number of frequency samples small. Because the number of frequency samples, M, must be greater than the duration, in m, of $w(n, m)$, given by (48), to prevent time aliasing of the processed signal, $h(n)$ should have finite duration and be as short as possible.

In order to minimize the effects of time-domain aliasing in the discrete-transform implementation, a finite length Hamming window was chosen as the analysis filter $h(n)$. If N denotes the duration of $h(n)$, then $H(\omega)$ is approximately band limited to $\pm 4\pi/N$ [20], i.e., the width of the mainlobe of $H(\omega)$ is approximately $8\pi/N$ and the peak amplitude of the sidelobes is -41 dB. If $4\pi/N$ is taken as the cutoff frequency of $H(\omega)$, then in order to resolve the pitch harmonics of voiced speech with a fundamental frequency $\Omega(n)$, N must be chosen such that

$$8\pi/N < \Omega(n) \tag{52a}$$

or equivalently,

$$N > 4P(n). \tag{52b}$$

This condition means that the duration of the Hamming window must be at least four pitch periods. If the samples $X_2(sR, k\Omega_M)$ are to represent $X_2(n, k\Omega_M)$ with no aliasing of $X(\psi, k\Omega_M)$ in ψ, then the temporal sampling interval, R, must be chosen less than $N/4$. Calculating $X_2(sR, k\Omega_M)$, therefore, corresponds to sliding $h(n)$ by no more than one fourth of its length for each FFT frame. In order to unwrap the phase of the STFT, however, it will be shown that it is necessary to sample at twice the Nyquist rate for $X_2(n, k\Omega_M)$. Thus, for the simulation, we choose

$$R \leqslant N/8. \tag{53}$$

B. Design of the Synthesis Filter

Because the nonlinear rate-changed modification does not preserve the structure of the STFT of an arbitrary signal, the synthesized signal, in general, depends on the design of $f(n)$. The interpretation of the discrete STFT as samples of the output of the band-limited analysis filter, $h(n)$, implies that $X_2(n, k\Omega_M)$ can be reconstructed from its samples $X_2(sR, k\Omega_M)$ by $1:R$ band-limited interpolation. This result is valid for the modified STFT, $Y_2(sR, k\Omega_M)$, as well, provided that the modification does not cause frequency aliasing of the sampled transform. The synthesis filter, $f(n)$, is, therefore, chosen as a $1:R$ optimal band-limited interpolating filter in order to guarantee proper interpolation of the speech parameters embedded in the modified STFT.

A procedure that is particularly well suited to designing interpolating filters for the short-time Fourier synthesizer is the algorithm proposed by Oetken *et al.* [21] for designing opti-

mal FIR digital interpolating filters. This procedure is attractive because it is a simple and efficient technique for designing filters of very high order. Furthermore, if the data to be interpolated is oversampled, then the design algorithm can exploit this property to improve the performance of the filter.

An additional benefit of choosing $f(n)$ as an interpolating filter is that the duration of the effective filter $w(n, m)$ is kept small. Assuming $h(n)$ to be band-limited and R to be less than the Nyquist interval for $h(n)$, the design of $f(n)$ as a $1:R$ band-limited interpolating filter gives

$$w(n,m) = \sum_{s=-\infty}^{\infty} f(n - sR)\, h(sR + m - n)$$
$$= h(m) \tag{54}$$

and the effective filter length is just the length of the analysis filter $h(n)$. Other designs for $f(n)$ will, in general, result in an effective filter length equal to the sum of the durations of $f(n)$ and $h(n)$. Since the number of frequency samples, M, required to prevent time aliasing of the synthesized signal increases with the duration of $w(n, m)$, filter designs other than a band-limited interpolating filter will generally require a larger number of frequency samples to represent the STFT.

VIII. Design of the Parameter Modification System

The modification of the STFT used to effect rate changes of speech consists of two basic procedures: linear time-scaling and phase modification. The details of these two procedures will be considered individually, then their combination and incorporation into the overall processing scheme will be considered.

A. Linear Time Scaling

In order to effect a rate change by the factor β in the synthesized signal, the magnitude and phase of the STFT must be linearly time-scaled by the factor β. Because the magnitude and phase are obtained by nonlinear operations on the STFT, these quantities have greater bandwidths than the real and imaginary parts of the STFT. Consequently, the linear time-scaling is implemented by applying band-limited decimation/interpolation techniques [19] to the real and imaginary parts of the STFT. The interpolating filter, denoted $f_M(n)$, is again designed by Oetken's technique [21].

Assuming the rate-changed scale factor β is a rational number, expressed as

$$\beta = D/I, \tag{55}$$

band-limited decimation/interpolation amounts to the direct evaluation of the sum

$$X_2(\beta sR, k\Omega_M) = \sum_{r=-\infty}^{\infty} f_M(sD - rI)\, X_2(rR, k\Omega_M) \tag{56a}$$

for the STFT, and

$$Y_2(sR, k\Omega_M) = \sum_{r=-\infty}^{\infty} f_M(sD - rI)\, Y_2(rR/\beta, k\Omega_M) \tag{56b}$$

for the modified STFT. This procedure is efficient because $f_M(n)$ has finite duration and the sum need only be evaluated once to compute each value of $X_2(\beta sR, k\Omega_M)$ or $Y_2(sR, k\Omega_M)$.

B. Phase Modification

The second procedure in modifying the phase of the STFT is dividing the FM component of the phase by the factor β. Specifically, we are given samples of the STFT expressed, according to the discussion in Part I, as

$$X_2(sR', k\Omega_M) = a(sR', k\Omega_M) \exp\,[j\vartheta(sR', k\Omega_M)] \tag{57}$$

where $\vartheta(n, \omega)$ denotes the FM component of the phase of the STFT and $a(n, \omega)$ is a complex function that varies slowly in n. The temporal sampling interval R' is either $R' = R$ or $R' = \beta \cdot R$ depending upon whether the phase modification is performed before or after the linear time-scaling operation. From these samples we wish to calculate the samples

$$Y_2(sR'/\beta, k\Omega_M) = a(sR', k\Omega_M) \exp\,[j\vartheta(sR', k\Omega_M)/\beta]\,. \tag{58}$$

Implementing this modification requires first estimating samples of the unwrapped FM phase component $\vartheta(n, \omega)$ from samples of the principal value of the phase of the STFT. If

$$X_2(n, \omega) = a(n, \omega) \exp\,[j\vartheta(n, \omega)]$$
$$= A(n, \omega) \exp\,[j(\alpha(n, \omega) + \vartheta(n, \omega))]$$
$$= A(n, \omega) \exp\,[j\theta(n, \omega)]$$

where

$$a(n, \omega) = A(n, \omega) \exp\,[j\alpha(n, \omega)]$$

and

$$A(n, \omega) = |a(n, \omega)| = |X_2(n, \omega)|$$
$$\alpha(n, \omega) = \arg\,[a(n, \omega)]$$
$$\theta(n, \omega) = \alpha(n, \omega) + \vartheta(n, \omega) = \arg\,[X_2(n, \omega)]\,, \tag{59}$$

then the samples of the FM phase component, $\vartheta(n, \omega)$, are to be estimated from samples of the principal value of the phase, denoted $PV[\theta(n, \omega)]$. As a function of n, $PV[\theta(n, \omega)]$ contains jumps of 2π due to the principal-value operator, and jumps of π due to sign changes in the real and imaginary parts of $a(n, \omega)$. Although the real and imaginary parts of $a(n, \omega)$ are slowly varying functions of n, the phase $\alpha(n, \omega)$ can jump by π when the real and imaginary parts of $a(n, \omega)$ simultaneously change sign. Except for these jumps of π, $\alpha(n, \omega)$ is also a slowly varying function of n. In order to estimate $\vartheta(n, \omega)$ from $PV[\theta(n, \omega)]$, it will be convenient to define the phase functions $\theta_\pi(n, \omega)$ and $\alpha_\pi(n, \omega)$, corresponding to $\theta(n, \omega)$ and $\alpha(n, \omega)$ with jumps of integer multiples of π removed. Thus,

$$\theta_\pi(n, \omega) = \alpha_\pi(n, \omega) + \vartheta(n, \omega). \tag{60}$$

Once $\theta_\pi(n, \omega)$ has been determined, $\vartheta(n, \omega)$ is estimated from $\theta_\pi(n, \omega)$ using the property that $\alpha_\pi(n, \omega)$ and the first difference of $\vartheta(n, \omega)$ are both slowly varying functions of n.

A procedure will now be developed for determining $\theta_\pi(n, \omega)$ from $PV[\theta(n, \omega)]$, based on the interpretation of the STFT as

the output of the low-pass filter $h(n)$. Denoting the first backward difference operator, with respect to n, by ∇_n, the first difference of the principal value of the phase of the STFT is

$$\begin{aligned}\nabla_n PV[\theta(n,\omega)] &= PV[\theta(n,\omega)] - PV[\theta(n-1,\omega)] \\ &= \nabla_n\theta_\pi(n,\omega) + \pi I_1(n,\omega) + 2\pi I_2(n,\omega)\end{aligned} \tag{61}$$

where I_1 and I_2 are integer-valued functions of n and ω representing the jumps of π and 2π in $PV[\theta(n,\omega)]$. Therefore, $\nabla_n\theta_\pi(n,\omega)$ differs from $\nabla_n PV[\theta(n,\omega)]$ only by an unknown integer multiple of π. Now, $X_2(n,\omega)$ is the output of the (low-pass) analysis filter, $h(n)$: since $\nabla_n\theta_\pi(n,\omega)$ is the instantaneous frequency of $X_2(n,\omega)$, and since both the instantaneous frequency and amplitude of $X_2(n,\omega)$ are modeled as slowly time-varying, $|\nabla_n\theta_\pi(n,\omega)|$ is assumed to be less than the cutoff frequency of $h(n)$. If ω_h denotes the cutoff frequency of $h(n)$, then

$$|\nabla_n\theta_\pi(n,\omega)| < \omega_h. \tag{62}$$

Furthermore, if $\omega_h < \pi/2$, then

$$|\nabla_n\theta_\pi(n,\omega)| < \frac{\pi}{2} \tag{63}$$

and $\nabla_n\theta_\pi(n,\omega)$ can be determined from $\nabla_n PV[\theta(n,\omega)]$ simply by adding integer multiples of π to $\nabla_n PV[\theta(n,\omega)]$ until the result satisfies condition (63). $\theta_\pi(n,\omega)$ can then be reconstructed by the running sum

$$\theta_\pi(n,\omega) = \sum_{r=n_0+1}^{n} \nabla_r\theta_\pi(r,\omega) + \theta_\pi(n_0,\omega) \tag{64}$$

where n_0 is an initial time at which $\theta_\pi(n_0,\omega)$ is assumed to be

$$\theta_\pi(n_0,\omega) = PV[\theta(n_0,\omega)].$$

For the sampled transform implementation,

$$\nabla_s\theta_\pi(sR', k\Omega_M) \approx R'\nabla_n\theta_\pi(n, k\Omega_M)\big|_{n=sR'} \tag{65}$$

because $\nabla_n\theta(n, k\Omega_M)$ is a slowly varying function of n. Multiplying both sides of (62) by R' and substituting (65) gives

$$|\nabla_s\theta_\pi(sR', k\Omega_M)| < \omega_h R'. \tag{66}$$

Now, if the sampling interval R' is chosen such that $\omega_h R' < \pi/2$, which corresponds to the sampling frequency $2\pi/R'$ being greater than $4\omega_h$, or twice the frequency required by the sampling theorem for sampling the output of $h(n)$, then (61) and condition (63) become

$$\begin{aligned}\nabla_s PV[\theta(sR', k\Omega_M)] &= \nabla_s\theta_\pi(sR', k\Omega_M) \\ &\quad + \pi I_1'(s,k) + 2\pi I_2'(s,k)\end{aligned} \tag{67}$$

and

$$|\nabla_s\theta_\pi(sR', k\Omega_M)| < \frac{\pi}{2}, \tag{68}$$

respectively. $\nabla_s\theta_\pi(sR', k\Omega_M)$ can, therefore, be determined from $\nabla_s PV[\theta(sR', k\Omega_M)]$ by adding integer multiples of π until (68) is satisfied.

The problem of estimating the FM component of the phase, $\vartheta(n,\omega)$, from $\theta_\pi(n,\omega)$ is basically a problem of curve fitting. Since

$$\theta_\pi(n,\omega) = \alpha_\pi(n,\omega) + \vartheta(n,\omega) \tag{69}$$

where $\alpha_\pi(n,\omega)$ is a slowly varying function of n with

$$\nabla_n\alpha_\pi(n,\omega) \approx 0$$

and $\vartheta(n,\omega)$ has a first difference

$$\nabla_n\vartheta_\pi(n,\omega) = \Omega(n,\omega)$$

that is also a slowly varying function of n, $\Omega(n,\omega)$ is the slope of a first-order polynomial (in n) that locally fits $\theta_\pi(n,\omega)$. While there are a variety of approaches to this problem, the technique to be described here was chosen for its simplicity and good performance in the actual simulation.

The first backward difference of (69) is given by

$$\begin{aligned}\nabla_n\theta_\pi(n,\omega) &= \nabla_n\alpha_\pi(n,\omega) + \nabla_n\vartheta(n,\omega) \\ &= \nabla_n\alpha_\pi(n,\omega) + \Omega(n,\omega) \\ &\approx \Omega(n,\omega)\end{aligned} \tag{70}$$

and the first forward difference by

$$\begin{aligned}\Delta_n\theta_\pi(n,\omega) &= \theta_\pi(n+1,\omega) - \theta_\pi(n,\omega) \\ &= \Delta_n\alpha_\pi(n,\omega) + \Delta_n\vartheta(n,\omega) \\ &= \Delta_n\alpha_\pi(n,\omega) + \Omega(n+1,\omega) \\ &\approx \Omega(n,\omega).\end{aligned} \tag{71}$$

A reasonable estimate for $\Omega(n,\omega)$, denoted $\tilde{\Omega}(n,\omega)$, is therefore the average of (70) and (71), given by

$$\begin{aligned}\tilde{\Omega}(n,\omega) &= \tfrac{1}{2}\,[\Delta_n + \nabla_n]\,\theta_\pi(n,\omega) \\ &= \tfrac{1}{2}\,[\theta_\pi(n+1,\omega) - \theta_\pi(n-1,\omega)] \\ &= \mu\delta_n\theta_\pi(n,\omega)\end{aligned}$$

where $\mu\delta_n = \frac{1}{2}\,[\Delta_n + \nabla_n]$ is known as the mean central-difference operator [23]. In practice, since

$$\Delta_n\theta_\pi(n,\omega) = \nabla_n\theta_\pi(n+1,\omega),$$

$\theta_\pi(n,\omega)$ is never actually computed from (64), but $\tilde{\Omega}(n,\omega)$ is computed directly as the average of the forward and backward differences

$$\begin{aligned}\tilde{\Omega}(n,\omega) &= \tfrac{1}{2}\,[\Delta_n + \nabla_n]\,\theta_\pi(n,\omega) \\ &= \tfrac{1}{2}\,[\nabla_n\theta_\pi(n+1,\omega) + \nabla_n\theta_\pi(n,\omega)].\end{aligned} \tag{72}$$

The estimate $\tilde{\vartheta}(n,\omega)$ for the FM phase component is constructed from the running sum

$$\tilde{\vartheta}(n,\omega) = \sum_{r=1}^{n} \tilde{\Omega}(r,\omega) \tag{73}$$

where $\tilde{\vartheta}(0,\omega) = \vartheta(0,\omega) = 0$.

For the sampled transform implementation, the estimate for $\Omega(sR', k\Omega_M)$ becomes

$$\begin{aligned}&\tilde{\Omega}(sR', k\Omega_M)\,R' \\ &= \mu\delta_s\theta_\pi(sR', k\Omega_M) \\ &= \tfrac{1}{2}\,[\Delta_s + \nabla_s]\,\theta_\pi(sR', \omega) \\ &= \tfrac{1}{2}\,[\nabla_s\theta_\pi(sR'+R', k\Omega_M) + \nabla_s\theta_\pi(sR', k\Omega_M)]\end{aligned} \tag{74}$$

and

$$\tilde{\vartheta}(sR', k\Omega_M) = \sum_{r=1}^{s} \tilde{\Omega}(sR', k\Omega_M)\, R'. \tag{75}$$

Once the estimate $\tilde{\vartheta}(sR', k\Omega_M)$ of the sampled FM phase component $\vartheta(sR', k\Omega_M)$ is calculated from (75), samples of the phase-modified STFT are calculated by adding $(1/\beta - 1) \cdot \tilde{\vartheta}(sR', k\Omega_M)$ to the phase of $X_2(sR', \Omega_M)$, or equivalently, multiplying $X_2(sR', \Omega_M)$ by $\exp[j(1/\beta - 1)\,\tilde{\vartheta}(sR', k\Omega_M)]$, to obtain

$$\begin{aligned} Y_2(sR'/\beta, k\Omega_M) &\approx a(sR', k\Omega_M) \exp[j\vartheta(sR', k\Omega_M)] \\ &\quad \cdot \exp\left[j\left(\frac{1}{\beta} - 1\right)\tilde{\vartheta}(sR', k\Omega_M)\right] \\ &= a(sR', k\Omega_M) \exp\left[j\left\{\vartheta(sR', k\Omega_M) + \left(\frac{1}{\beta} - 1\right)\tilde{\vartheta}(sR', k\Omega_M)\right\}\right] \\ &\approx a(sR', k\Omega_M) \exp[j\vartheta(sR', k\Omega_M)/\beta]. \end{aligned} \tag{76}$$

C. The Overall Modification System

The linear time-scaling and phase modification of the STFT each affect the bandwidth of $X_2(n, \omega)$, considered as a sequence in n for each ω. If the STFT is down-sampled, close to its Nyquist rate, to obtain $X_2(sR, \omega)$, then the order in which the linear time-scaling and phase modification are implemented becomes important to prevent frequency aliasing of $X(\psi, \omega)$ in ψ. In particular, the bandwidth (in ψ) of the time-scaled STFT

$$X_2(\beta n, \omega) = a(\beta n, \omega) \exp[j\vartheta(\beta n, \omega)]$$

is β times the bandwidth of the original STFT $X_2(n, \omega)$. In contrast, the bandwidth (in ψ) of the phase-modified STFT

$$Y_2(n/\beta, \omega) = a(n, \omega) \exp[j\vartheta(n, \omega)/\beta]$$

is approximately $1/\beta$ times the bandwidth of $X_2(n, \omega)$. The bandwidth of the modified STFT

$$Y_2(n, \omega) = a(\beta n, \omega) \exp[j\vartheta(\beta n, \omega)/\beta]$$

obtained as the result of both linear time-scaling and phase modification, is approximately the same as the bandwidth of $X_2(n, \omega)$. Thus, for time-scale expansion ($0 < \beta < 1$), the linear time-scaling operation decreases the bandwidth of the STFT in ψ, while the phase modification operation increases it. Conversely, for time-scale compression ($\beta > 1$), the linear time-scaling operation increases the bandwidth of the STFT, while the phase modification decreases it.

Suppose $X_2(n, \omega)$ is represented by its samples $X_2(sR, k\Omega_M)$, where R is close to the Nyquist interval for sampling $X_2(n, \omega)$ in n. More precisely, suppose that either $\beta \cdot R$ or $1/\beta \cdot R$ is greater than the Nyquist interval for $X_2(n, \omega)$. Then, in order to prevent frequency aliasing (in ψ) when implementing time-scale expansion, the linear time-scaling must be implemented first, followed by the phase modification; conversely, when implementing time-scale compression, the phase modification must be implemented first, followed by the linear time scaling.

D. Implicit Time-Scaling

Under certain circumstances, computational savings may be gained by incorporating the linear time-scaling of the STFT into the analysis and synthesis procedures. This method will be called the *implicit* method for linearly time-scaling the STFT, in contrast to the previously described *explicit* method, and is effected by assuming different temporal sampling rates for the short-time Fourier analysis and synthesis. Let R_A denote the sampling interval at the output of the short-time Fourier analyzer, and R_S the sampling interval assumed by the synthesizer. The output of the analyzer is, therefore, given by

$$\begin{aligned} X_2(sR_A, k\Omega_M) &= \sum_{m=-\infty}^{\infty} h(sR_A - m)\, x(m) \exp[-j\Omega_M km] \\ &= a(sR_A, k\Omega_M) \exp[j\vartheta(sR_A, k\Omega_M)] \end{aligned} \tag{77}$$

and the output of the synthesizer given by

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} f(n - sR_S)\, Y_2(sR_S, k\Omega_M) \cdot \exp[j\Omega_M nk]. \tag{78}$$

If $Y_2(sR_S, k\Omega_M) = Y_2(sR_A/\beta, k\Omega_M)$ is defined as the result of dividing the FM component of the phase of STFT by β (with no time-scaling), i.e.,

$$Y_2(sR_S, k\Omega_M) = a(sR_A, k\Omega_M) \exp[j\vartheta(sR_A, k\Omega_M)/\beta], \tag{79}$$

then substituting (79) into (78) gives the expression for the synthesized signal

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} f(n - sR_S)\, a(sR_A, k\Omega_M) \cdot \exp[j\vartheta(sR_A, k\Omega_M)/\beta] \exp[j\Omega_M nk]. \tag{80}$$

Assuming $\beta = D/I$, let $R_A = R/I$ and $R_S = R/D$, where R is an integer parameter that specifies the actual sampling intervals such that R_A and R_S are integers. Thus, (80) becomes

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} f(n - sR/D)\, a(sR/I, k\Omega_M) \cdot \exp[j\vartheta(sR/I, k\Omega_M)/\beta] \exp[j\Omega_M nk]. \tag{81}$$

If the sampling parameter R is chosen with careful attention to the issues discussed in the preceding sections, namely, that $X_2(sR_A, k\Omega_M) = X_2(sR/I, k\Omega_M)$ corresponds to samples of $X_2(n, \omega)$ sampled often enough to estimate the FM component of its phase, and $Y_2(sR_S, k\Omega_M) = Y_2(sR/I, k\Omega_M)$ is sampled often enough to prevent aliasing, then, if $f(n)$ is a $1 : R_S$

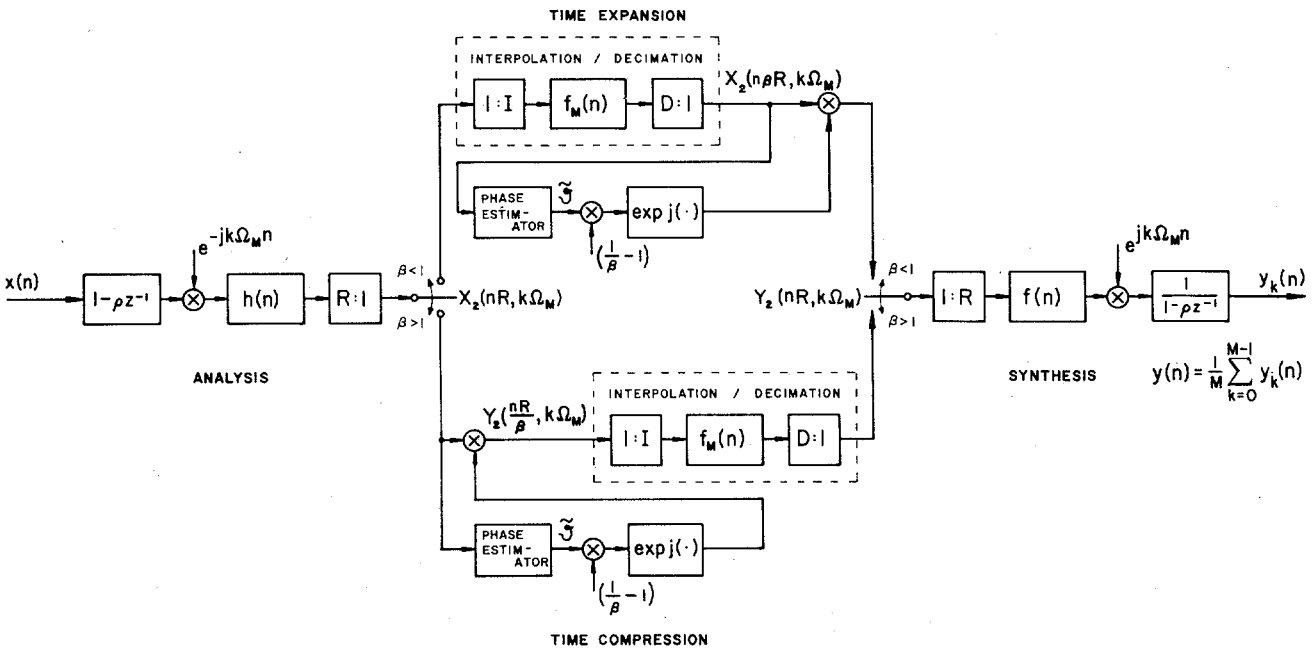

Fig. 2. Block diagram for one channel of a speech rate-change system.

band-limited interpolating filter, (81) becomes

$$y(n) = \frac{1}{M}\sum_{k=0}^{M-1}\sum_{s=-\infty}^{\infty} a(nD/I, k\Omega_M) \cdot \exp\left[j\vartheta(nD/I, k\Omega_M)/\beta\right] \exp\left[j\Omega_M nk\right]$$
$$= \frac{1}{M}\sum_{k=0}^{M-1}\sum_{s=-\infty}^{\infty} a(\beta n, k\Omega_M) \cdot \exp\left[j\vartheta(\beta n, k\Omega_M)/\beta\right] \exp\left[j\Omega_M nk\right] \quad (82)$$

which is the desired rate-changed signal.

Whether or not the implicit method of linearly time-scaling the STFT is more efficient than the explicit method depends primarily on the factor β. For example, if either D or I is unity, then the implicit method is more efficient, whereas if D and I are both large integers, then the explicit method is generally more efficient. Another consideration in choosing between these two methods is the mode of operation of the system. For example, if the same speech passage is to be processed several times with different rate-changed scale factors, i.e., for perceptual studies, then it may be most efficient to perform the analysis once at a relatively high sampling rate, store the samples of the STFT, and perform the synthesis assuming different sampling rates as described above. Another useful aspect of the implicit method is that it permits exploiting the requirement that the output of the short-time Fourier analyzer is sampled at twice the minimum rate required to avoid frequency aliasing. Thus, for example, time-scale expansion by the factor of 2:1 ($\beta = \frac{1}{2}$) can be implemented simply by modifying the phase of the discrete STFT and performing the synthesis assuming $R_S = 2 \cdot R_A$. Finally, the implicit and explicit methods are easily combined, in practice, to realize the advantages of both.

IX. Simulation of a Time-Scale Modification System

The concepts of the previous sections were employed in a general-purpose minicomputer simulation of a complete time-scale modification system for speech. This section discusses some of the details and results of the simulation.

The complete time-scale modification system is depicted in Fig. 2. The figure is segmented into three sections: a short-time Fourier analyzer, a system for modifying the samples of the short-time Fourier transform, and a short-time Fourier synthesizer. Note that the two realizations of the modification system, one for time-scale expansion ($0 < \beta < 1$) and the other for time-scale compression ($\beta > 1$), are shown explicitly.

For the simulation, the input sequence, $x(n)$, was obtained by sampling 5 kHz low-pass filtered speech at the sampling rate of 10 kHz. Then, $x(n)$ was preemphasized using the first-order system

$$H_p(z) = 1 - \rho z^{-1} \quad (83)$$

with $\rho = 0.95$. The analysis window was chosen as an N-point Hamming window with N in the range $256 < N < 512$, corresponding to analog filter bandwidths of 156 Hz to 78 Hz. The output of the short-time Fourier analyzer was sampled every $R_A \leq N/8$ samples, and the number of frequency samples M was fixed at 512. The synthesis was performed assuming the sampling interval R_S less than $N/4$ and, whenever possible, using the implicit time-scaling method, as described in Section VIII-D, to exploit the oversampling of the transform. The synthesis filter was an optimal $1:R_S$ FIR digital interpolating fil-

ter designed using Oetken's method, taking into account the cutoff frequency of the data. For the explicit method of time-scaling the STFT, the interpolating filter $f_M(n)$ was designed with the same technique. The output sequence $y(n)$ was post-emphasized with the system

$$H_p^{-1}(z) = 1/(1 - \rho z^{-1}), \tag{84}$$

desampled, assuming the sampling rate of 10 kHz, and low-pass filtered at 5 kHz.

The results of the simulation demonstrate that this system is capable of producing high-quality rate-changed speech for reasonable values of β, i.e., for compression ratios as high as 3:1 and expansion ratios as high as 4:1 [24]. Informal listening indicates that the processed speech retained its natural quality and speaker-dependent features and was free from artifacts such as glitches, burbles, and reverberation. For time-scale expansion by greater than 4:1, the processed speech began to exhibit small amounts of reverberant coloration, due to time-domain aliasing, which one should be able to eliminate by increasing the number of frequency samples, M. For time-scale compression of 4:1, or greater, the processed speech often became rough. Note, however, that while arbitrarily high expansion ratios are physically reasonable, arbitrarily high compression ratios are not. Consider, for example, a voiced phoneme containing four pitch periods. Greater than 4:1 compression reduces this phoneme to less than one pitch period, destroying the periodic character of the phoneme. Thus, one might expect speech, time-compressed with high compression ratios, to have a rough quality and low intelligibility.

In addition to producing high-quality rate-changed speech when the original speech was high quality, the system also performed well processing degraded speech. Specifically, speech corrupted by additive Gaussian white noise with 0 dB signal-to-noise ratio was processed with good results [24]. Informal listening indicated that while the processed speech was, of course, noisy, the system was robust in the sense that the noise in the processed speech was not perceived to increase in intensity, nor was the noise perceived as correlated with the speech.

Appendix I

The purpose of this Appendix is to derive the time-varying power spectrum, (41), for a synthetic speech signal generated by short-time Fourier synthesis.

Let $Y_2(n, \omega)$ denote a modified short-time Fourier transform and let $y(n)$ denote the signal synthesized according to the short-time Fourier synthesis formula

$$y(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{r=-\infty}^{\infty} f(n - r)\, Y_2(r, \omega) \exp [j\omega n]\, d\omega. \tag{A1}$$

The autocorrelation function $R_y(n, \tau)$, defined as

$$R_y(n, \tau) = E\{y(n + \tau)\, y^*(n)\} \tag{A2}$$

can be expressed in terms of the modified autocorrelation function $K_y(n, \omega, \tau, \epsilon)$ for $Y_2(n, \omega)$, and the synthesis filter, $f(n)$, by replacing $y(n)$ in (A2) with (A1):

$$\begin{aligned} R_y(n, \tau) &= E\left\{\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \sum_{p=-\infty}^{\infty} f(n + \tau - p)\, Y_2(p, \varphi) \cdot \exp [j\varphi(n + \tau)]\, d\varphi\right) \cdot \frac{1}{2\pi}\int_{-\pi}^{\pi} \sum_{q=-\infty}^{\infty} f(n - q)\, Y_2(q, \zeta) \exp [j\zeta n]\, d\zeta\right)^*\right\} \\ &= \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \sum_p \sum_q f(n + \tau - p) f^*(n - q) \cdot E\{Y_2(p, \varphi)\, Y_2^*(q, \zeta)\} \exp [j\{\varphi - \zeta) n + \varphi\tau\}]\, d\varphi\, d\zeta. \end{aligned}$$

Making the change of variables $\varphi = \omega - \epsilon/2$, $\zeta = \omega + \epsilon/2$, and $p = q + r$ gives

$$\begin{aligned} R_y(n, \tau) &= \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\epsilon_0}^{\epsilon_0} \sum_r \sum_q f(n + \tau - q - r) f^*(n - q) \cdot E\left\{Y_2\left(q + r, \omega - \frac{\epsilon}{2}\right) Y_2^*\left(q, \omega + \frac{\epsilon}{2}\right)\right\} \cdot \exp\left[-j\left(n + \frac{\tau}{2}\right)\epsilon\right] \exp [j\omega\tau]\, d\epsilon\, d\omega \\ &= \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\epsilon_0}^{\epsilon_0} \sum_r \sum_q f(n + \tau - q - r) \cdot f^*(n - q)\, K_y(q, \omega, r, \epsilon) \exp [j\omega\tau]\, d\epsilon\, d\omega \end{aligned} \tag{A3}$$

where the limits on the inner integral are given in terms of

$$\epsilon_0 = \epsilon_0(\omega) = 2(\pi - |\omega|).$$

Equivalently, letting $l = n - q$ and $m = \tau - r$ gives the result that the autocorrelation function for the synthetic speech is

$$\begin{aligned} R_y(n, \tau) &= \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\epsilon_0}^{\epsilon_0} \sum_l \sum_m f(l + m) f^*(l) \cdot K_y(n - l, \omega, \tau - m, \epsilon) \exp [j\omega\tau]\, d\epsilon\, d\omega. \end{aligned} \tag{A4}$$

The time-varying power spectrum $S_y(n, \omega)$ for $y(n)$ is the partial Fourier transform of $R_y(n, \tau)$ with respect to τ, i.e.,

$$S_y(n, \omega) \doteq \sum_{\tau=-\infty}^{\infty} R_y(n, \tau) \exp [-j\omega\tau]. \tag{A5}$$

Replacing $R_y(n, \tau)$ by (A4) gives

$$\begin{aligned} S_y(n, \omega) = \sum_\tau \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\epsilon_0}^{\epsilon_0} \sum_q \sum_r & \cdot f(n + \tau - q - r) f^*(n - q)\, K_y(q, \varphi, r, \epsilon) \\ & \cdot \exp [j\varphi\tau] \exp [-j\omega\tau]\, d\epsilon\, d\varphi \end{aligned}$$

and, letting $s = n + \tau - q - r$,

$$
\begin{aligned}
&S_y(n,\omega) \\
&= \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\epsilon_0}^{\epsilon_0} \sum_q \sum_r \sum_s f(s) f^*(n-q) K_y(q,\varphi,r,\epsilon) \\
&\quad \cdot \exp\left[j(\omega-\varphi)(n-q-r-s)\right] d\epsilon\, d\varphi \\
&= \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\epsilon_0}^{\epsilon_0} \sum_q \sum_r F(\omega-\varphi) f^*(n-q) K_y(q,\varphi,r,\epsilon) \\
&\quad \cdot \exp\left[j(\omega-\varphi)(n-q-r)\right] d\epsilon\, d\varphi \\
&= \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\epsilon_0}^{\epsilon_0} \sum_r F(\omega-\varphi) \exp\left[j(\omega-\varphi)(n-r)\right] \\
&\quad \cdot \left\{\sum_q f(n-q) K_y^*(q,\varphi,r,\epsilon) \right. \\
&\quad \left. \cdot \exp\left[-j(\varphi-\omega) q\right]\right\}^* d\epsilon\, d\varphi. \qquad \text{(A6)}
\end{aligned}
$$

The summation over q is recognized as the short-time Fourier transform of $K_y(q,\varphi,r,\epsilon)$, with respect to q, using $f(n)$ as the *analysis* window. Assuming that $K_y(n,\omega,\tau,\epsilon)$ varies slowly in n, so that it is low pass and narrow band compared to $F(\omega)$, the technique presented in the Appendix of [15] for approximating the short-time Fourier transform of a narrow-band signal can be applied to (A6) to give

$$
\begin{aligned}
S_y(n,\omega) = \left(\frac{1}{2\pi}\right)^2 &\int_{-\pi}^{\pi}\int_{-\epsilon_0}^{\epsilon_0} \sum_{r=-\infty}^{\infty} F(\omega-\varphi) \\
&\cdot \exp\left[j(\omega-\varphi)(n-r)\right] \\
&\cdot \{F(-(\varphi-\omega)) K_y^*(n,\varphi,r,\epsilon) \\
&\cdot \exp\left[-j(\varphi-\omega) n\right]\}^* d\epsilon\, d\varphi. \qquad \text{(A7)}
\end{aligned}
$$

Thus, the desired expression for the time-varying power spectrum of the synthetic signal is

$$
\begin{aligned}
S_y(n,\omega) = \left(\frac{1}{2\pi}\right)^2 &\int_{-\pi}^{\pi}\int_{-\epsilon_0}^{\epsilon_0} |F(\omega-\varphi)|^2 \\
&\cdot \sum_{r=-\infty}^{\infty} K_y(n,\varphi,r,\epsilon) \exp\left[-j(\omega-\varphi) r\right] d\epsilon\, d\varphi.
\end{aligned} \qquad \text{(A8)}
$$

Appendix II

The purpose of this Appendix is to derive an approximation to $K_y(n,\omega,\tau,\epsilon)$ for a synthetic time-scale-modified signal $y(n)$ in terms of $K_x(n,\omega,\tau,\epsilon)$ for the original signal $x(n)$.

The modified short-time Fourier transform $Y_2(n,\omega)$ for synthesizing rate-changed speech is

$$
\begin{aligned}
Y_2(n,\omega) &= a(\beta n,\omega) \exp\left[j\vartheta(\beta n,\omega)/\beta\right] \\
&= A(\beta n,\omega) \exp\left[j(\alpha(\beta n,\omega) + \vartheta(\beta n,\omega)/\beta)\right]
\end{aligned} \qquad \text{(A9)}
$$

where

$$
\begin{aligned}
X_2(n,\omega) &= a(n,\omega) \exp\left[j\vartheta(n,\omega)\right] \\
&= A(n,\omega) \exp\left[j(\alpha(n,\omega) + \vartheta(n,\omega))\right] \\
&= A(n,\omega) \exp\left[j\theta(n,\omega)\right]
\end{aligned} \qquad \text{(A10)}
$$

and the functions $a(n,\omega)$, $A(n,\omega)$, $\alpha(n,\omega)$, $\vartheta(n,\omega)$, and $\theta(n,\omega)$ are defined in Section VIII. In approximating $K_y(n,\omega,\tau,\epsilon)$, the slowly varying phase angle $\alpha(n,\omega)$ will be neglected so that

$$
Y_2(n,\omega) = A(\beta n,\omega) \exp\left[j\theta(\beta n,\omega)/\beta\right] \qquad \text{(A11)}
$$

will be taken as the modified short-time Fourier transform, and $K_y(n,\omega,\tau,\epsilon)$ becomes

$$
\begin{aligned}
&K_y(n,\omega,\tau,\epsilon) \\
&= E\left\{A\left(\beta(n+\tau), \omega - \frac{\epsilon}{2}\right) A\left(\beta n, \omega + \frac{\epsilon}{2}\right)\right. \\
&\quad \left. \cdot \exp\left[j\left(\theta\left(\beta(n+\tau), \omega - \frac{\epsilon}{2}\right) - \theta\left(\beta n, \omega + \frac{\epsilon}{2}\right)\right)/\beta\right]\right\} \\
&\quad \cdot \exp\left[-j\left(n + \frac{\tau}{2}\right)\epsilon\right]. \qquad \text{(A12)}
\end{aligned}
$$

$K_y(n,\omega,\tau,\epsilon)$ can be approximated by expanding the right-hand side of (A12) in a two-dimensional Taylor series about $\tau = \epsilon = 0$, and the coefficients approximated in terms of the moments of the analysis filter, $h(n)$. To obtain the coefficients in the expansion, assume that $A(n,\omega)$ and $\theta(n,\omega)$ correspond to samples of $A(t,\omega)$ and $\theta(t,\omega)$, defined as the magnitude and phase of $X_2(t,\omega)$, where $X_2(t,\omega)$ is defined by band-limited interpolation of $X_2(n,\omega)$. Thus

$$
\begin{aligned}
&K_y(t,\omega,\tau,\epsilon) \\
&= E\left\{A\left(\beta(t+\tau), \omega - \frac{\epsilon}{2}\right) A\left(\beta t, \omega + \frac{\epsilon}{2}\right)\right. \\
&\quad \left. \cdot \exp\left[j\left(\theta\left(\beta(t+\tau), \omega - \frac{\epsilon}{2}\right) - \theta\left(\beta t, \omega + \frac{\epsilon}{2}\right)\right)/\beta\right]\right\} \\
&\quad \cdot \exp\left[-j\left(t + \frac{\tau}{2}\right)\epsilon\right]. \qquad \text{(A13)}
\end{aligned}
$$

For quasi-stationary speech, assume that $K_y(t,\omega,\tau,\epsilon)$ is a slowly-varying function of t, so that

$$
\begin{aligned}
&K_y(t,\omega,\tau,\epsilon) \\
&\approx K_y\left(t - \frac{\tau}{2}, \omega, \tau, \epsilon\right) \\
&= E\left\{A\left(\beta\left(t + \frac{\tau}{2}\right), \omega - \frac{\epsilon}{2}\right) A\left(\beta\left(t - \frac{\tau}{2}\right), \omega + \frac{\epsilon}{2}\right)\right. \\
&\quad \cdot \exp\left[j\left(\theta\left(\beta\left(t + \frac{\tau}{2}\right), \omega - \frac{\epsilon}{2}\right)\right.\right. \\
&\quad \left.\left. - \theta\left(\beta\left(t - \frac{\tau}{2}\right), \omega + \frac{\epsilon}{2}\right)\right)/\beta\right]\Bigg\} \exp\left[-j\epsilon t\right]. \qquad \text{(A14)}
\end{aligned}
$$

Although this approximation is not necessary, it does simplify the mathematics. For convenience, the time-scaled function

$$K_y(t/\beta, \omega, \tau/\beta, \epsilon) = E\left\{A\left(t+\frac{\tau}{2}, \omega-\frac{\epsilon}{2}\right) A\left(t-\frac{\tau}{2}, \omega+\frac{\epsilon}{2}\right) \cdot \exp\left[j\left(\theta\left(t+\frac{\tau}{2}, \omega-\frac{\epsilon}{2}\right) - \theta\left(t-\frac{\tau}{2}, \omega+\frac{\epsilon}{2}\right)\right)/\beta\right]\right\} \cdot \exp\left[-j\epsilon t/\beta\right] \quad \text{(A15)}$$

will be expanded, rather than (A14).

Expanding the right-hand side of (A15) in a two-dimensional Taylor series in τ and ϵ and retaining terms up to second order gives

$$\begin{aligned} K_y(t/\beta, \omega, \tau/\beta, \epsilon) &= E\{A^2\} + j\left[E\{\theta_t A^2\}\,\tau/\beta - E\{(\theta_\omega + t)A^2\}\,\epsilon/\beta\right] \\ &\quad + \tfrac{1}{2}\left[\tfrac{1}{2}\cdot E\{\hat{A}_{tt}A^2\} - E\{\theta_t^2 A^2\}\right]\tau^2/\beta^2 \\ &\quad - \left[\tfrac{1}{2}\cdot E\{\hat{A}_{t\omega}A^2\} - E\{\theta_t(\theta_\omega + t)A^2\}\right]\tau\epsilon/\beta^2 \\ &\quad + \tfrac{1}{2}\left[\tfrac{1}{2}\cdot E\{\hat{A}_{\omega\omega}A^2\} - E\{(\theta_\omega + t)^2 A^2\}\right]\epsilon^2/\beta^2 \\ &\quad + \cdots \end{aligned} \quad \text{(A16)}$$

where $\hat{A} = \log A(t, \omega)$ and the subscripts t and ω denote partial differentiation of the subscripted quantity with respect to the subscript, e.g., $\theta_t = \partial_t \theta(t, \omega)$. Similarly, expanding

$$K_x(t, \omega, \tau, \epsilon) = E\left\{A\left(t+\frac{\tau}{2}, \omega-\frac{\epsilon}{2}\right) A\left(t-\frac{\tau}{2}, \omega+\frac{\epsilon}{2}\right) \cdot \exp\left[j\left(\theta\left(t+\frac{\tau}{2}, \omega-\frac{\epsilon}{2}\right) - \theta\left(t-\frac{\tau}{2}, \omega+\frac{\epsilon}{2}\right)\right)\right]\right\} \cdot \exp\left[-j\epsilon t\right]$$

gives

$$\begin{aligned} K_x(t, \omega, \tau, \epsilon) &= E\{A^2\} + j\left[E\{\theta_t A^2\}\,\tau - E\{(\theta_\omega + t)A^2\}\,\epsilon\right] \\ &\quad + \tfrac{1}{2}\left[\tfrac{1}{2}\cdot E\{\hat{A}_{tt}A^2\} - E\{\theta_t^2 A^2\}\right]\tau^2 \\ &\quad - \left[\tfrac{1}{2}\cdot E\{\hat{A}_{t\omega}A^2\} - E\{\theta_t(\theta_\omega + t)^2 A^2\}\right]\tau\epsilon \\ &\quad + \tfrac{1}{2}\left[\tfrac{1}{2}\cdot E\{\hat{A}_{\omega\omega}A^2\} - E\{(\theta_\omega + t)^2 A^2\}\right]\epsilon^2 \\ &\quad + \cdots \end{aligned} \quad \text{(A17)}$$

which corresponds to the expansion (A16) with $\beta = 1$. Thus,

$$\begin{aligned} K_y(t/\beta, \omega, \tau/\beta, \epsilon) &= K_x(t, \omega, \tau, \epsilon) + \left[K_y(t/\beta, \omega, \tau/\beta, \epsilon) - K_x(t, \omega, \tau, \epsilon)\right] \\ &= K_x(t, \omega, \tau, \epsilon) \\ &\quad + j(1/\beta - 1)\left[E\{\theta_t A^2\}\,\tau - E\{(\theta_\omega + t)A^2\}\,\epsilon\right] \\ &\quad - \tfrac{1}{2}(1/\beta^2 - 1)\left[E\{\theta_t^2 A^2\}\tau^2 - 2E\{\theta_t(\theta_\omega + t)A^2\}\,\tau\epsilon\right. \\ &\quad \left. + E\{(\theta_\omega + t)^2 A^2\}\,\epsilon^2\right] + \cdots. \end{aligned} \quad \text{(A18)}$$

Since the expansion for $K_x(t, \omega, \tau, \epsilon)$ to second order has already been determined, the expansion for $K_y(t/\beta, \omega, \tau/\beta, \epsilon)$ requires evaluating the expectations

$$E\{\theta_t A^2\}, \quad E\{(\theta_\omega + t)A^2\}$$
$$E\{\theta_t^2 A^2\}, \quad E\{\theta_t(\theta_\omega + t)A^2\}, \quad E\{(\theta_\omega + t)^2 A^2\}. \quad \text{(A19)}$$

Assuming Gaussian statistics for the unvoiced speech signal, $x(n)$, a tedious calculation yields

$$E\{\theta_t A^2\} \approx J_x(t, \omega)\, M_h \quad \text{(A20a)}$$
$$E\{(\theta_\omega + t)A^2\} \approx J_x(t, \omega)\, m_h \quad \text{(A20b)}$$
$$E\{(\theta_t - M_h)^2 A^2\} \approx \tfrac{1}{2} J_x(t, \omega)\, D_h^2 \quad \text{(A20c)}$$
$$E\{(\theta_t - M_h)(\theta_\omega + t - m_h)A^2\} \approx -\tfrac{1}{2} J_x(t, \omega)\, \mu_h^2 \quad \text{(A20d)}$$
$$E\{(\theta_\omega + t - m_h)^2\} A^2 \approx \tfrac{1}{2} J_x(t, \omega)\, d_h^2 \quad \text{(A20e)}$$

where the functions on the right-hand side are defined in [15].

Briefly, the procedure for calculating the expectations (A20) is the following. Define $U(t, \omega)$ and $V(t, \omega)$ as the real and imaginary parts of $X_2(t, \omega)$, so that

$$X_2(t, \omega) = U(t, \omega) + jV(t, \omega)$$

and

$$A^2(t, \omega) = U^2(t, \omega) + V^2(t, \omega)$$

and let the subscripts μ and ν denote partial differentiation with respect to either t or ω. The quantity $\theta_\mu(t, \omega)$ can be expressed as

$$\theta_\mu(t, \omega) = \operatorname{Im}\{\partial_\mu \log X_2(t, \omega)\} = \{(UV_\mu - U_\mu V)/A^2\} \quad \text{for} \quad \mu = t, \omega.$$

Furthermore, it follows that

$$E\{\theta_\mu A^2\} = E\{UV_\mu - U_\mu V\} \quad \text{(A21)}$$

and

$$E\{\theta_\mu \theta_\nu A^2\} = E\{(UV_\mu - U_\mu V)\cdot(UV_\nu - U_\nu V)/A^2\} \quad \text{for} \quad \mu, \nu = t, \omega. \quad \text{(A22)}$$

The expectations (A21) and (A22) can be evaluated from the conditional means

$$E\{U_\mu | U, V\} \quad \text{and} \quad E\{V_\mu | U, V\} \quad \text{(A23)}$$

and the conditional correlations

$$E\{U_\mu U_\nu | U, V\}, E\{U_\mu V_\nu | U, V\}, \quad \text{and} \quad E\{V_\mu V_\nu | U, V\}. \quad \text{(A24)}$$

Assuming Gaussian statistics for the unvoiced speech signal, the functions U, V and their partial derivatives are jointly Gaussian. Therefore, the conditional means (A23) and the conditional correlations (A24) can be expressed in terms of the (unconditional) means and correlations of U, V and their partial derivatives (see, for example, [25]).

The correlation functions

$$E\{U(t_1, \omega_1)\, U(t_2, \omega_2)\}, E\{U(t_1, \omega_1)\, V(t_2, \omega_2)\},$$
$$\text{and} \quad E\{V(t_1, \omega_1)\, V(t_2, \omega_2)\}$$

can be calculated by interpreting the short-time Fourier transform (for each value of ω) as a demodulated narrow-band random process and paralleling the discussion in [26] for narrow-band Gaussian random processes. The correlation functions for the derivatives are then obtained by appropriately differentiating the correlation functions for U and V.

The desired approximation to $K_y(n, \omega, \tau, \epsilon)$ is now obtained by replacing the expectations in the series (A18) with the values (A20) and setting $M_h = m_h = 0$ to give

$$\begin{aligned} K_y(t/\beta, \omega, \tau/\beta, \epsilon) &= K_x(t, \omega, \tau, \epsilon) \\ &\quad - \tfrac{1}{4}(1/\beta^2 - 1) J_x(t, \omega) [D_h^2\tau^2 + 2\mu_h\tau\epsilon + d_h^2\epsilon^2] \\ &\quad + \cdots . \end{aligned} \tag{A25}$$

Expanding $K_x(t, \omega, \tau, \epsilon)$ in a two-dimensional power series

$$\begin{aligned} K_x(n, \omega, \tau, \epsilon) &= J_x(n, \omega) \\ &\quad \cdot \{1 - \tfrac{1}{2} [D_h^2\tau^2 + 2\mu_h\tau\epsilon + d_h^2\epsilon^2] + \cdots\} \end{aligned} \tag{A26}$$

and collecting like terms yields

$$\begin{aligned} K_y(t/\beta, \omega, \tau/\beta, \epsilon) &= J_x(t, \omega) \\ &\quad \cdot \{1 - \tfrac{1}{4}(1/\beta^2 + 1) [D_h^2\tau^2 + 2\mu_h\tau\epsilon + d_h^2\epsilon^2] + \cdots\}. \end{aligned} \tag{A27}$$

Scaling t and τ by β, (A27) becomes

$$\begin{aligned} K_y(t, \omega, \tau, \epsilon) &= J_x(\beta t, \omega) \\ &\quad \cdot \{1 - \tfrac{1}{4}(1 + \beta^2) [D_h^2\tau^2 + 2\mu_h\tau(\epsilon/\beta) + d_h^2(\epsilon/\beta)^2] \\ &\quad + \cdots\} \end{aligned} \tag{A28}$$

and defining the parameter γ such that

$$\gamma^2 = \tfrac{1}{2}(1 + \beta^2) \tag{A29}$$

gives

$$\begin{aligned} K_y(t, \omega, \tau, \epsilon) &= J_x(\beta t, \omega) \\ &\quad \cdot \left\{1 - \frac{\gamma^2}{2} [D_h^2\tau^2 + 2\mu_h\tau(\epsilon/\beta) + d_h^2(\epsilon/\beta)^2] + \cdots\right\}. \end{aligned} \tag{A30}$$

Finally, replacing t with n and comparing the resulting expansion for $K_y(n, \omega, \tau, \epsilon)$ with the expansion (A26), for $K_x(n, \omega, \tau, \epsilon)$ shows that

$$K_y(n, \omega, \tau, \epsilon) \approx K_x(\beta n, \omega, \gamma\tau, \gamma\epsilon/\beta). \tag{A31}$$

References

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

[2] —, "Recording device," U.S. Patent 2 886 650, May 12, 1959.

[3] F. F. Lee, "Time compression and expansion of speech by the sampling method," *J. Audio Eng. Soc.*, vol. 20, pp. 738–742, Nov. 1972.

[4] R. J. Scott and S. E. Gerber, "Pitch synchronous time compression of speech," in *Proc. Conf. Speech Comm. Processing*, Apr. 1972, pp. 63–65.

[5] A. W. F. Huggins, "More temporally segmented speech: Is duration or speech content the critical variable in its loss of intelligibility?" Research Laboratory of Electronics, M.I.T., Cambridge, MA, Quarterly Progress Rep. 114, July 15, 1974, pp. 185–193.

[6] H. D. Toong, "A study of time-compressed speech," Ph.D. dissertation, Dep. Elec. Eng. Comput. Sci., M.I.T., Cambridge, 1974.

[7] E. P. Neuburg, "Simple pitch-dependent algorithm for high-quality speech rate-change," presented at the 93rd Meeting Acoustic Soc. Amer., June 1977; *J. Acoust. Soc. Amer.* (abstract), vol. 61, suppl. 1, Spring 1977.

[8] M. R. Schroeder, "Vocoders: Analysis and synthesis of speech," *Proc. IEEE*, vol. 54, pp. 720–734, May 1966; reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel, Eds. New York: IEEE Press, 1979.

[9] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd Ed. Berlin, Germany: Springer, 1972.

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

[11] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, pp. 1493–1509, Nov. 1966; reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel, Eds. New York: IEEE Press, 1979.

[12] M. R. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 243–248, June 1976; reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel, Eds. New York: IEEE Press, 1979.

[13] J. A. Moorer, "The use of the phase vocoder in computer music applications," presented at the 55th Conv. Audio Engineering Soc., preprint 1146 (E-1), Oct. 1976.

[14] M. W. Callahan, "Acoustic signal processing based on the short-time spectrum," Ph.D. dissertation, Dep. Comput. Sci., Univ. Utah, Salt Lake City, Tech. Rep. UTEC-CSc-76-209, Mar. 1976.

[15] M. R. Portnoff, "Short-time Fourier analysis of speech," this issue, pp. 364–373.

[16] —, "Representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 55–69, Feb. 1980.

[17] —, "Time-scale modification of speech based on short-time Fourier analysis," Sc.D. dissertation, Dep. Elec. Eng. Comput. Sci., M.I.T., Cambridge, Apr. 1978.

[18] R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 99–102, Feb. 1980.

[19] R. W. Schafer and L. R. Rabiner, "A digital signal processing approach to interpolation," *Proc. IEEE*, vol. 61, pp. 692–702, June 1973.

[20] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[21] G. Oetken, T. W. Parks, and H. W. Scheussler, "New results in the design of digital interpolators," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 301–309, June 1975.

[22] J. M. Heinz and K. N. Stevens, "On the properties of voiceless fricative constants," *J. Acoust. Soc. Amer.*, vol. 33, pp. 589–596, May 1961; reprinted in *Readings in Acoustic Phonetics*, I. Lehiste, Ed. Cambridge, MA: M.I.T. Press, 1967.

[23] F. B. Hildebrand, *Finite Difference Equations and Simulations*. Englewood Cliffs, NJ: Prentice-Hall, 1968.

[24] M. R. Portnoff, "A mathematical framework for time-scale modification of speech," presented at the 93rd Meeting Acoustic Soc. Amer., Pennsylvania State Univ., University Park, June 1977; *J. Acoust. Soc. Amer.* (abstract), vol. 61, suppl. 1, Spring 1977.

[25] F. C. Schweppe, *Uncertain Dynamical Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

[26] W. B. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*. New York: McGraw-Hill, 1958.

Michael R. Portnoff (S'69–M'77–M'78), for a photograph and biography, see this issue, p. 373.