Short-Time Fourier Analysis of Sampled Speech

MICHAEL R. PORTNOFF, MEMBER, IEEE

Abstract—**The theoretical basis for the representation of a speech signal by its short-time Fourier transform is developed. A time-frequency representation for linear time-varying systems is applied to the speech-production model to formulate a quasi-stationary representation for the speech waveform. Short-time Fourier analysis of the resulting representation yields the relationship between the short-time Fourier transform of the speech and the speech-production model.**

I. Introduction

SHORT-TIME Fourier analysis not only provides an effective method for speech analysis [1]-[4], but a mathematical framework for a variety of speech-processing systems [3]-[13]. The purpose of this paper is to develop the relationship between the parameters of a speech waveform, based on a speech-production model, and the short-time Fourier transform of the speech.

The first part of this paper formulates a quasi-stationary representation for sampled speech signals. Voiced speech is represented, on a short-time basis, as a linear combination of narrow-band signals with harmonically related instantaneous frequencies. Each of these component signals can be expressed as a complex exponential, modulated in amplitude, phase, and frequency such that the modulating signals are related to a speech production model. Unvoiced speech is represented by its second moments as a quasi-stationary random process with a slowly time-varying power spectrum that corresponds to the time-varying vocal-tract transfer function. An important feature of this representation is that simple time scaling (decimation/interpolation) of the parameters corresponds to changing the rate of the speech [12], [13].

The second part of this paper formulates the relationship between the short-time Fourier transform of a speech signal and the parameters of the underlying speech-production model. The results of this development suggest how the short-time Fourier transform can be used to estimate and, if desired, modify the parameters of the speech waveform.[1]

II. A Quasi-Stationary Representation for Speech Signals

The generally accepted engineering model for the production of speech signals [14] is illustrated in Fig. 1. According to this model, samples of the speech waveform, denoted by $x(n)$, are assumed to be the output of a linear time-varying filter that approximates the transmission characteristics of the vocal tract and the spectral characteristics of the glottal pulse. For voiced speech, the filter is driven by a quasi-periodic train of unit samples, $v(n)$, such that the spacing between the unit samples corresponds to the pitch, or fundamental, period of the speech. For unvoiced speech, the filter is driven by a stationary random sequence, $u(n)$ with a flat power spectrum, i.e., white noise.

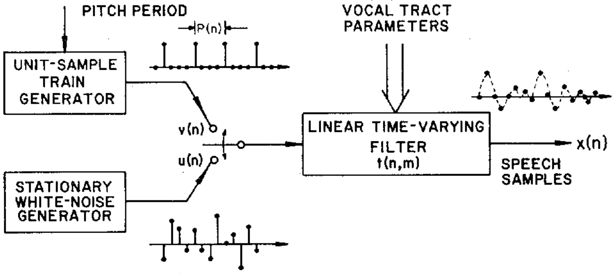

PITCH PERIOD
P(n)
UNIT-SAMPLE TRAIN GENERATOR
VOCAL TRACT PARAMETERS
v(n)
u(n)
LINEAR TIME-VARYING FILTER t(n,m)
x(n)
SPEECH SAMPLES
STATIONARY WHITE-NOISE GENERATOR

Fig. 1. Terminal-analog model of the vocal system (after Schafer and Rabiner [14]).

The input-output behavior of the linear time-varying filter depicted in Fig. 1 is completely characterized by its time-varying unit-sample response $t(n, m)$, defined as the response of the system at time n to a unit sample applied m samples earlier, at time $(n - m)$ [15]-[19]. An equivalent description is given by the time-varying frequency response of the system, $T_2(n, \omega)$, defined such that the response of the system at time sample n to the complex exponential $\exp [j\omega n]$ is $T_2(n, \omega) \cdot \exp [j\omega n]$. It can be shown [19] that $T_2(n, \omega)$ is the Fourier transform of $t(n, m)$ with respect to the second index (m).

$$T_2(n, \omega) = \sum_{m=-\infty}^{\infty} t(n, m) \exp [-j\omega m]. \tag{1}$$

The subscript 2 is used to denote that $T_2(n, \omega)$ is a "partial" Fourier transform with respect to its second argument. Furthermore, the time variation, or nonstationarity, of the system corresponds to the dependence of the functions $t(n, m)$ and $T_2(n, \omega)$ on the index n.

In the speech-production model, the nonstationarity of $t(n,m)$ corresponds to the movement of the physical articulators and is usually relatively slow compared to the time variation of the input and output waveforms that correspond to acoustic disturbances. Because the vocal tract is, in essence, a lossy acoustic resonator, its impulse response can be modeled as finite duration and causal. Moreover, the time variation of the vocal tract is sufficiently slow compared to the decay rate of its impulse response that the system can be modeled as nearly stationary

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.

[1] For the details of implementing short-time Fourier analysis and synthesis the reader is referred to [4], [8], [12], [20], [21].

0096-3518/81/0600-0364$00.75 © 1981 IEEE

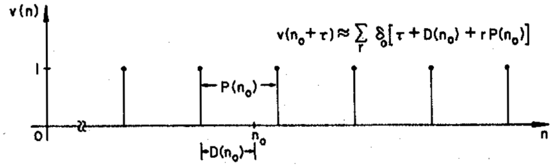


Fig. 2. Quasi-periodic unit-sample train.

for the duration of its memory. Such a system, which is approximately stationary for the duration of its memory, will be termed a "quasi-stationary system."

A. Harmonic Representation of Voiced Speech

According to Fig. 1, voiced speech is modeled as the output of $t(n, m)$ driven by the quasi-periodic unit-sample train $v(n)$. The term "quasi-periodic" is used to indicate that $v(n)$ has a locally periodic behavior. Specifically, if $v(n)$, illustrated in Fig. 2, is expressed relative to a sliding time frame as $v(n_0 + \tau)$, representing, for a fixed value of n_0 and small $|\tau|$, the local behavior of $v(n)$, then $v(n_0 + \tau)$ is periodic in τ for $(n_0 + \tau)$ in the neighborhood about n_0 for which $v(n)$ is modeled as periodic.

A harmonic representation for voiced-speech signals will now be formulated by representing the excitation, $v(n)$, as a sum of harmonically related complex exponentials and using this representation in the superposition sum with $t(n, m)$. Referring to Fig. 2, let $P(n_0)$ denote the local pitch period of $v(n)$ in the neighborhood of n_0, let $D(n_0)$ denote the number of samples to the sample n_0 from the unit sample arriving most recently before, or at the sample n_0, and let $\delta_0(n)$ denote the unit-sample sequence

$$\delta_0(n) = \begin{cases} 1 & \text{for } n = 0 \\ 0 & \text{for } n \neq 0. \end{cases}$$

Therefore, $v(n)$ can be represented locally as

$$v(n_0 + \tau) \approx \sum_{r=-\infty}^{\infty} \delta_0(\tau + D(n_0) + rP(n_0)) \tag{2}$$

for small $|\tau|$. Equivalently, $v(n_0 + \tau)$ can be represented as the sum of harmonically related complex exponentials

$$v(n_0 + \tau) \approx \frac{1}{P(n_0)} \sum_{k=0}^{P(n_0)-1} \exp\,[j2\pi k(D(n_0) + \tau)/P(n_0)] \tag{3}$$

or

$$v(n_0 + \tau) \approx \frac{1}{P(n_0)} \sum_{k=0}^{P(n_0)-1} \exp\,[jk(\phi(n_0) + \Omega(n_0)\tau + \phi_0)] \tag{4}$$

where

$$\Omega(n) = 2\pi/P(n) \tag{5a}$$

$$\phi_0 = \Omega(0)\,D(0) \tag{5b}$$

$$\phi(n) = \Omega(n)\,D(n) + 2\pi I(n) - \phi_0 \tag{5c}$$

and $I(n)$ is an integer whose value depends on n. That (2) and (3) are equivalent follows directly from evaluating the sum in (3), for fixed n_0, as the sum of a finite number of terms in a geometric series.

The quantity $\Omega(n)$, referred to as the "instantaneous frequency" of the fundamental, can assume values in the range

$$0 \leqslant \Omega(n) < 2\pi \tag{6}$$

because the pitch period, $P(n)$, is defined as a positive integer and, for speech, is greater than unity. Furthermore, because $v(n)$ is quasi-periodic, both $P(n)$ and $\Omega(n)$ are slowly varying, so that

$$P(n_0 + \tau) \approx P(n_0) \qquad \text{for small } |\tau| \tag{7a}$$

and

$$\Omega(n_0 + \tau) \approx \Omega(n_0) \qquad \text{for small } |\tau|. \tag{7b}$$

The constant phase angle ϕ_0 is introduced as a convenience so that $\phi(0) = 0$ and to preserve the time origin under rate-change modifications [12], [13]. Note that ϕ_0 is zero if one of the unit samples of $v(n)$ arrives at $n = 0$.

The quantity $\phi(n)$ is referred to as the "instantaneous phase" of the fundamental. The unspecified additive multiple of 2π in (5c) can be specified by requiring that the value of the exponent in (4) be uniquely defined for each value of $n = n_0 + \tau$. Thus, for $n = n_0 + \tau$, we require that

$$\begin{aligned} \phi(n) + \phi_0 &= \phi(n_0 + \tau) + \phi_0 \\ &\approx \phi(n_0) + \Omega(n_0)\,\tau + \phi_0 \end{aligned} \tag{8}$$

or

$$\phi(n_0 + \tau) \approx \phi(n_0) + \Omega(n_0)\,\tau \tag{9}$$

for small $|\tau|$. Further, setting $\tau = -1$ in (9) leads to the property that the instantaneous frequency is (approximately) the first (backwards) difference of the instantaneous phase, i.e.,

$$\Omega(n) \approx \phi(n) - \phi(n-1). \tag{10}$$

Since

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

$\phi(n)$ will now be *defined* explicitly as

$$\phi(n) = \begin{cases} \displaystyle\sum_{r=1}^{n} \Omega(r) & \text{for } n > 0 \\ 0 & \text{for } n = 0 \\ \displaystyle\sum_{r=0}^{n+1} -\Omega(r) & \text{for } n < 0 \end{cases} \tag{12}$$

and (12) will be taken as the definition of $\phi(n)$. Note that, from (10) and (6), $\phi(n)$ has the property

$$0 \leqslant \phi(n) - \phi(n-1) < 2\pi$$

and is, therefore, an "unwrapped phase" angle rather than a principal value. If $v(n)$ is strictly periodic with period P for all time, then $\Omega(n) = 2\pi/P$ is a constant, and $\phi(n)$ becomes $2\pi n/P$.

The excitation, $v(n)$, for voiced speech will, therefore, be modeled as the sum of harmonically related complex exponentials

$$v(n) = \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} \exp\,[jk(\phi(n) + \phi_0)] \tag{13}$$

where $\phi(n)$ is defined by (12),

$$\Omega(n) = 2\pi/P(n) \quad \text{and} \quad \phi_0 = \Omega(0)\,D(0).$$

For a strictly periodic unit-sample train with period P the representation (13) becomes, simply,

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

The harmonic representation (13) is, in fact, more general than the time-domain representation (2). In addition to representing periodic sequences, the harmonic representation also leads to representations for aperiodic sequences with periodic envelopes, such as those obtained by uniformly sampling periodic continuous-time waveforms when the period of the waveform is not an integer multiple of the sampling interval.

A voiced-speech signal $x(n)$, modeled as the output of $t(n,m)$ driven by $v(n)$, is given according to the superposition sum [12], [19] as

$$x(n) = \sum_{m=-\infty}^{\infty} t(n,m)\, v(n-m). \tag{14}$$

For voiced speech, the pitch, $\Omega(n)$, is assumed to be constant for the duration of the memory of $t(n,m)$. Therefore, $v(n-m)$ in (14) can be replaced by the local harmonic representation (4) to obtain

$$\begin{aligned} x(n) &= \sum_{m=-\infty}^{\infty} t(n,m) \\ &\quad \cdot \left\{ \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} \exp\,[jk(\phi(n) - \Omega(n)\,m + \phi_0)] \right\} \\ &= \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} T_2(n, k\Omega(n)) \exp\,[jk(\phi(n) + \phi_0)]. \end{aligned} \tag{15}$$

Thus, the voiced-speech signal, $x(n)$, is represented as the linear combination of harmonically related complex exponentials

$$x(n) = \sum_{k=0}^{P(n)-1} c_k(n) \exp\,[jk\phi(n)] \tag{16a}$$

where

$$c_k(n) = \frac{1}{P(n)} T_2(n, k\Omega(n)) \exp\,[jk\phi_0] \tag{16b}$$

and

$$\phi(n) = \begin{cases} \sum_{r=1}^{n} \Omega(r) & \text{for } n > 0 \\ 0 & \text{for } n = 0 \\ \sum_{r=0}^{n+1} -\Omega(r) & \text{for } n < 0. \end{cases} \tag{16c}$$

The quantities $c_k(n)$, referred to as the "complex harmonic amplitudes" of the speech, are slowly varying functions of n. Because the time variation of the $c_k(n)$'s corresponds to changes in the vocal-tract geometry, the $c_k(n)$'s contain significant Fourier components only up to the order of a few tens of hertz. Thus, on the spectrum of acoustic frequencies, which extends into the tens of kilohertz, the $c_k(n)$'s are narrow-band low-pass sequences. Furthermore, a property that will be exploited in the short-time Fourier analysis of speech is that the bandwidths of the $c_k(n)$'s are much less than the fundamental frequency, $\Omega(n)$, of the speech.

B. Second Moment Representation of Unvoiced Speech

An unvoiced-speech signal, $x(n)$, is modeled as the output of the quasi-stationary filter $t(n,m)$ driven by the real zero-mean stationary white-noise process $u(n)$ with autocorrelation function

$$\begin{aligned} R_u(\tau) &= E\{u(n+\tau)\, u^*(n)\} \\ &= \sigma_u^2 \delta_0(\tau) \end{aligned} \tag{17}$$

(where $E\{\cdot\}$ denotes expected value and * denotes complex conjugate). Clearly, $x(n)$ has zero mean and is nonstationary with its autocorrelation function given by

$$\begin{aligned} &R_x(n,\tau) \\ &= E\{x(n+\tau)\, x^*(n)\} \\ &= E\left\{ \sum_{q=-\infty}^{\infty} t(n+\tau, q)\, u(n+\tau-q) \sum_{m=-\infty}^{\infty} t^*(n,m)\, u^*(n-m) \right\} \\ &= \sum_{q=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} t(n+\tau, q)\, t^*(n,m)\, \sigma_u^2\, \delta_0(\tau - q + m) \\ &= \sum_{m=-\infty}^{\infty} \sigma_u^2 t(n+\tau, m+\tau)\, t^*(n,m). \end{aligned} \tag{18}$$

Since $t(n,m)$ is assumed to be a quasi-stationary system, its time variation (in n) is negligible over the duration of its memory (correlation time in m). Hence, (18) becomes

$$R_x(n,\tau) \approx \sum_{m=-\infty}^{\infty} \sigma_u^2 t(n, m+\tau)\, t^*(n,m)$$

or

$$R_x(n,\tau) \approx \frac{1}{2\pi} \int_{-\pi}^{\pi} \sigma_u^2\, |T_2(n,\omega)|^2 \exp\,[j\omega\tau]\; d\omega. \tag{19}$$

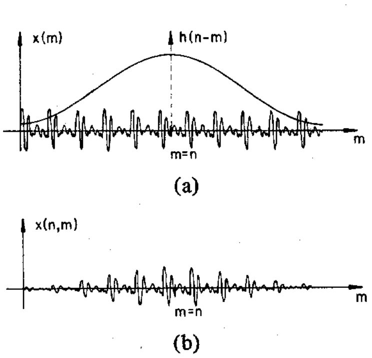


Fig. 3. (a) Time-reversed and shifted analysis window $h(n-m)$ superimposed on data $x(m)$. (b) Short-time sequence $x(n, m) = h(n-m) \cdot x(m)$ for a particular value of n.

A nonstationary random process, such as $x(n)$, that can be modeled as the output of a quasi-stationary linear time-varying system, will be referred to as a "quasi-stationary random process." Furthermore, as a generalization of the power spectrum for a wide-sense stationary random process, the "time-varying power spectrum," $S_x(n, \omega)$, is defined for the quasi-stationary random process, $x(n)$, as

$$S_x(n, \omega) = \sigma_u^2 \, |T_2(n, \omega)|^2. \tag{20}$$

Thus, based on the approximation (19), $R_x(n, \tau)$ and $S_x(n, \omega)$ can be written as the Fourier transform pair

$$R_x(n, \tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(n, \omega) \exp [j\omega\tau] \, d\omega \tag{21}$$

and

$$S_x(n, \omega) = \sum_{\tau=-\infty}^{\infty} R_x(n, \tau) \exp [-j\omega\tau]. \tag{22}$$

An unvoiced-speech signal $x(n)$ will, therefore, be represented as a quasi-stationary random process with zero mean and characterized by its second moments, namely $R_x(n, \tau)$, or equivalently, $S_x(n, \omega)$. An important property of this representation is that it provides a direct relationship between the time-varying statistics of $x(n)$ and the parameters of the filter $t(n, m)$.

III. Short-Time Fourier Analysis

The short-time Fourier transform (STFT) of a discrete-time signal $x(n)$ is defined by the sum

$$X_2(n, \omega) = \sum_{m=-\infty}^{\infty} h(n-m) \, x(m) \exp [-j\omega m] \tag{23}$$

where the subscript 2 is again used to denote that the second argument is the Fourier transform variable [12], [19]; $h(n)$ is referred to as the analysis window and is generally chosen to have the property that it is, in some sense, narrow in time, or frequency, or both, and is normalized such that $h(0) = 1$.

Referring to Fig. 3, $X_2(n, \omega)$ can be interpreted for each value of n as the partial Fourier transform, with respect to m, of the "short-time function"

$$x(n, m) = h(n-m) \, x(m). \tag{24}$$

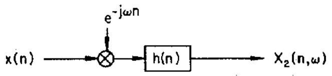


Fig. 4. Short-time Fourier transform as output of demodulator followed by analysis filter.

Equivalently, by considering (23) as the convolution

$$X_2(n, \omega) = h(n) *_n x(n) \exp [-j\omega n] \tag{25}$$

where $*_n$ denotes the convolution operator with respect to n, $X_2(n, \omega)$ can be interpreted as the output of a linear time-invariant filter $h(n)$, excited by the demodulated (frequency-shifted) signal $x(n) \exp [-j\omega n]$, as shown in Fig. 4. For this reason, $h(n)$ is also referred to as the analysis filter.

A. Short-Time Fourier Analysis of Voiced Speech

According to the discussion of Section II-A, a voiced-speech signal is modeled on a short-time basis as a linear combination of harmonically related complex exponentials, each with slowly-varying complex amplitude and instantaneous frequency. The STFT of a voiced-speech signal, $x(n)$, will now be expressed in terms of the parameters of its harmonic representation. Substituting the harmonic representation (16) for $x(n)$ into (23) gives

$$X_2(n, \omega) = \sum_{m=-\infty}^{\infty} \sum_{k=0}^{P(m)-1} h(n-m) \, c_k(m) \exp [jk\phi(m)] \exp [-j\omega m]. \tag{26}$$

Assuming the duration of the analysis filter, $h(n)$, is sufficiently short that the pitch of the speech is constant over the duration of $h(n)$, the local representation for $\phi(n)$ can be used to simplify (26). Specifically, the approximations

$$\begin{aligned} &P(m) \approx P(n) \\ &\phi(m) \approx \phi(n) + \Omega(n)(m-n) \qquad \text{for} \quad h(n-m) \neq 0 \end{aligned} \tag{27}$$

are substituted for $P(m)$ and $\phi(m)$ in (26) to give

$$\begin{aligned} X_2(n, \omega) &= \sum_{m=-\infty}^{\infty} \sum_{k=0}^{P(n)-1} h(n-m) \\ &\quad \cdot c_k(m) \exp [jk(\phi(n) + \Omega(n)(m-n))] \exp [-j\omega m] \\ &= \sum_{k=0}^{P(n)-1} \sum_{m=-\infty}^{\infty} h(n-m) \, c_k(m) \exp [-j(\omega - k\Omega(n)) m] \\ &\quad \cdot \exp [jk(\phi(n) - \Omega(n) n)]. \end{aligned} \tag{28}$$

The summation over m in (28) is recognized as the STFT of $c_k(n)$, evaluated for ω replaced by $[\omega - k\Omega(n)]$. According to the model for voiced speech, discussed in Section II-A, the harmonic amplitudes $c_k(n)$ are narrow-band low-pass sequences. If the analysis filter, $H(\omega)$, is designed with a bandwidth greater than the bandwidth of $C_k(\omega)$, the Fourier transform of $c_k(n)$, then the technique for evaluating the STFT of a narrow-band

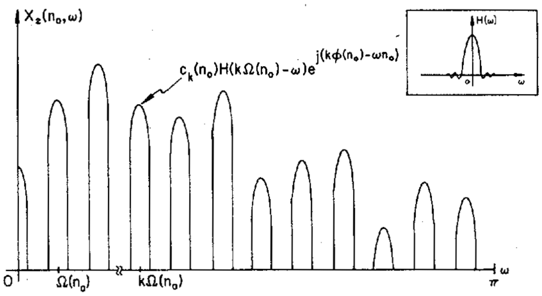


Fig. 5. Short-time Fourier transform of an idealized speech signal for a particular value of $n = n_0$.

signal [see the Appendix] can be applied, for each value of k, to give[2]

$$X_2(n,\omega) = \sum_{k=0}^{P(n)-1} c_k(n)\, H(-(\omega - k\Omega(n)))\exp\left[-j(\omega - k\Omega(n))\,n\right] \cdot \exp\left[jk(\phi(n) - \Omega(n)\,n)\right]$$

or,

$$X_2(n,\omega) = \sum_{k=0}^{P(n)-1} c_k(n)\, H(k\Omega(n) - \omega)\exp\left[j(k\phi(n) - \omega n)\right]. \tag{29}$$

For the fixed value $n = n_0$, (29) expresses the STFT of $x(n)$ as the sum of $P(n_0)$ images of $H(\omega)$ each shifted in frequency by $k\Omega(n_0)$ and weighted by $c_k(n_0)\exp\left[j(k\phi(n_0) - \omega n_0)\right]$, as shown in Fig. 5.

1) Narrow-Band Analysis: For narrow-band (NB) analysis of voiced speech, the bandwidth of $H(\omega)$ is chosen to be less than the instantaneous fundamental frequency, $\Omega(n)$. The shifted and weighted images of $H(\omega)$ shown in Fig. 5 are, therefore, nonoverlapping as illustrated, and $X_2(n,\omega)$ given by (29) reduces to

$$X_2(n,\omega) = \begin{cases} c_k(n)\, H(k\Omega(n) - \omega)\exp\left[j(k\phi(n) - \omega n)\right] \\ \qquad\qquad \text{for } |\omega - k\Omega(n)| < \omega_h \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{30}$$

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

The STFT is a complex quantity, which can be expressed in terms of its magnitude and phase as

$$X_2(n,\omega) = A(n,\omega)\exp\left[j\theta(n,\omega)\right] \tag{31}$$

where

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

and

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

The magnitude of the STFT, obtained from (30) as

$$A(n,\omega) = |c_k(n)|\,|H(k\Omega(n) - \omega)| \qquad \text{for } |\omega - k\Omega(n)| < \omega_h, \tag{32}$$

is a slowly-varying function of n because both $c_k(n)$ and $\Omega(n)$ are slowly-varying functions of n. Referring to (30), the phase of the STFT can be expressed as the sum of two components

$$\theta(n,\omega) = \alpha(n,\omega) + \vartheta(n,\omega) \tag{33a}$$

where

$$\alpha(n,\omega) = \arg\left[c_k(n)\right] + \arg\left[H(k\Omega(n) - \omega)\right] \tag{33b}$$

and

$$\vartheta(n,\omega) = k\phi(n) - \omega n \tag{33c}$$

for

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

The component $\alpha(n,\omega)$ contributes a slowly time-varying phase and will be called the "phase-modulation" component. The other component, $\vartheta(n,\omega)$, can be expressed, using the definition (16c) of $\phi(n)$, as

$$\vartheta(n,\omega) = \begin{cases} \sum_{r=1}^{n} \{k\Omega(r) - \omega\} & \text{for } n > 0 \\ 0 & \text{for } n = 0 \\ \sum_{r=0}^{n+1} -\{k\Omega(r) - \omega\} & \text{for } n < 0 \end{cases}$$
$$\text{where } |\omega - k\Omega(n)| < \omega_h. \tag{34}$$

Consequently, $\vartheta(n,\omega)$ satisfies the recursion relation

$$\vartheta(n,\omega) = \vartheta(n-1,\omega) + k\Omega(n) - \omega \tag{35}$$

or

$$\vartheta(n,\omega) = \vartheta(n-1,\omega) + \Omega(n,\omega) \tag{36}$$

where

$$\Omega(n,\omega) = k\Omega(n) - \omega \tag{37}$$

with k such that

$$|\Omega(n,\omega)| < \omega_h. \tag{38}$$

Because

$$\Omega(n,\omega) = \vartheta(n,\omega) - \vartheta(n-1,\omega), \tag{39}$$

$\Omega(n,\omega)$ will be 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

$$\vartheta(n_0 + \tau, \omega) \approx \vartheta(n_0,\omega) + \Omega(n_0,\omega)\,\tau \tag{40}$$

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.

[2] This same result could be obtaned using an alternative assumption that the vocal-tract geometry be stationary for the duration of the analysis window so that $h(n-m)\,c_k(m) \approx h(n-m)\,c_k(n)$.

2) Wide-Band Analysis: For wide-band (WB) analysis of voiced speech, the bandwidth of the analysis filter $H(\omega)$ is chosen to be several times greater than the fundamental frequency, $\Omega(n)$, of the speech. In general, the goal of WB analysis is two fold. First, because the bandwidth of $H(\omega)$ used for WB analysis is greater than that used for NB analysis, the duration of $h(n)$ is correspondingly shorter. Consequently, the time resolution of the WB-STFT is greater than that of the NB-STFT. Second, because the analysis-filter bandwidth is greater than the frequency spacing of the pitch harmonics, the WB-STFT does not have sufficient frequency resolution to resolve the individual pitch harmonics. The WB-STFT, therefore, provides a means for interpolating between the pitch harmonics of voiced speech to obtain a "smoothed" estimate of the frequency response, $T_2(n, \omega)$, of the speech-production model. In contrast to the NB case where the quasi-periodic nature of the speech was manifested in the magnitude of the STFT as harmonics along the frequency axis, in the WB case it will be shown that the quasi-periodic nature of the speech is manifested in the magnitude of the STFT as quasi-periodic pulses along the time axis.

Because the bandwidth of $H(\omega)$ is greater than the frequency spacing between adjacent pitch harmonics, more than one term in the sum (29) contributes to the STFT for a particular frequency ω. Thus, interpreting (29) for the WB-STFT is more difficult than for the NB-STFT, and it is convenient to introduce the following additional assumptions about the structure of the speech model.

1) The frequency response $T_2(n, \omega)$ is assumed to have a *formant* structure. Specifically, $T_2(n, \omega)$ is assumed to consist of Q poles (and at most Q zeros) so that

$$T_2(n, \omega) = \sum_{i=1}^{Q} \frac{a_i(n)}{1 - \zeta_i(n) \exp[-j\omega]} \tag{41}$$

where

$$\zeta_i(n) = \rho_i(n) \exp[j\omega_i(n)], \quad 0 < \rho_i(n) < 1 \tag{42}$$

are the slowly time-varying poles which, together with the residues $a_i(n)$, are assumed to be approximately constant for the duration of the WB analysis window.

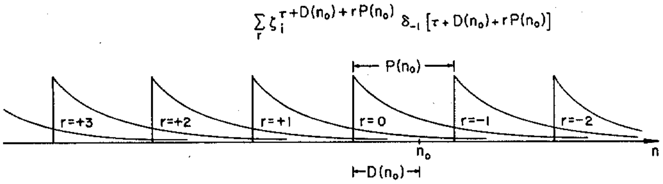
2) The bandwidth of each formant, $\zeta_i(n)$, is assumed to be much less than the bandwidth of the analysis filter, $H(\omega)$.[3]

3) The frequency spacing between formants is assumed to be greater than the bandwidth of $H(\omega)$.

Substituting the assumed formant structure (41) into the expression (29) for the STFT gives

$$X_2(n, \omega) = \sum_{i=1}^{Q} \frac{1}{P} \sum_{k=0}^{P-1} \frac{a_i}{1 - \zeta_i \exp[-jk\Omega]} \cdot H(k\Omega - \omega) \exp[j\{k(\phi(n) + \phi_0) - \omega n\}] \tag{43}$$

[to simplify notation, the time dependence of the slowly varying quantities $a_i(n)$, $\zeta_i(n)$, $\Omega(n)$, and $P(n)$ has not been explicitly shown in (43)].



$\sum_r \zeta_i^{\tau+D(n_0)+rP(n_0)} \delta_{-1}[\tau + D(n_0) + rP(n_0)]$
P(n₀)
r=+3
r=+2
r=+1
r=0
r=-1
r=-2
n₀
n
D(n₀)

Fig. 6. Time response, in the neighborhood of $n = n_0$, of the ith format to the quasi-periodic unit-sample train $v(n)$.

Because the bandwidth of $H(\omega)$ is assumed to be much larger than the bandwidth of any of the formants, an argument analogous to that used for short-time Fourier analysis of narrow-band signals can be made to approximate $H(k\Omega - \omega)$ as constant over the width of each of the formant peaks. Consequently, for each of the Q formant peaks, the approximation

$$\frac{a_i H(k\Omega - \omega)}{1 - \zeta_i \exp[-jk\Omega]} \approx \frac{a_i H(\omega_i - \omega)}{1 - \zeta_i \exp[-jk\Omega]}, \quad i = 1, 2, \cdots, Q \tag{44}$$

can be applied to (43) to obtain

$$X_2(n, \omega) = \sum_{i=1}^{Q} H(\omega_i - \omega) \frac{1}{P} \sum_{k=0}^{P-1} \frac{a_i}{1 - \zeta_i \exp[-jk\Omega]} \cdot \exp[j\{k(\phi(n) + \phi_0) - \omega_n\}]. \tag{45}$$

Substituting the expansion

$$\frac{a_i}{1 - \zeta_i \exp[-jk\Omega]} = a_i \sum_{m=0}^{\infty} \zeta_i^m \exp[-jk\Omega m] \tag{46}$$

in (45), expressing $X_2(n, \omega)$ locally for $n = n_0 + \tau$, and performing the summation over k gives

$$X_2(n_0 + \tau, \omega) = \sum_{i=1}^{Q} H(\omega_i - \omega) \exp[-j\omega(n_0 + \tau)] \cdot a_i \sum_r \zeta_i^{\tau + rP + D} \delta_{-1}(\tau + rP + D) \tag{47}$$

where $\delta_{-1}(n)$ is the unit-step sequence

$$\delta_{-1}(n) = \begin{cases} 1 & \text{for } n \geq 0 \\ 0 & \text{for } n < 0 \end{cases}$$

and $P(n_0)$ and $D(n_0)$ are the local parameters of the voicing excitation $v(n)$, defined in Section II-A. The other slowly varying parameters in (47) are similarly evaluated for $n = n_0$.

Referring to Fig. 6, for each value of i, the summation over r in (47) is the superposition of images of the unit-sample response of the ith formant repeated every $P(n_0)$ samples in the neighborhood of n_0, and is quasi-periodic. Since each image of the unit-sample response is causal, only the "tails" of preceding images contribute to a particular "period." If it is assumed that the time constant of each formant is less than the interval over which $P(n)$ and $t(n, m)$ can be regarded as stationary, then the contributions to a given period from the tails of the preceding periods can be approximated by a one-sided

[3] Average formant bandwidths are on the order of 40–170 Hz [22] and typical WB analysis filter bandwidths are on the order of 300 Hz.

infinite sum of the tails of the preceding periods to give

$$X_2(n_0+\tau,\omega)=\sum_{i=1}^{Q} H(\omega_i-\omega)\exp\left[-j\omega(n_0+\tau)\right]\cdot a_i\sum_{r}\sum_{q=r}^{\infty}\zeta_i^{\tau+qP+D}\,\mathrm{rect}_P\left[\tau+rP+D\right] \tag{48}$$

where

$$\mathrm{rect}_P[n]=\begin{cases}1 & \text{for } 0\leqslant n<P\\ 0 & \text{otherwise.}\end{cases}$$

Performing the summation over q:

$$X_2(n_0+\tau,\omega)=\sum_{i=1}^{Q} H(\omega_i-\omega)\exp\left[-j\omega(n_0+\tau)\right]\cdot\tilde{a}_i\sum_{r}\zeta_i^{\tau+rP+D}\,\mathrm{rect}_P\left[\tau+rP+D\right] \tag{49}$$

where

$$\tilde{a}_i=\frac{a_i}{1-\zeta_i^P}.$$

Because the frequency spacing between adjacent formant peaks is assumed to be greater than the bandwidth of $H(\omega)$, $H(\omega_i-\omega)$ and $H(\omega_j-\omega)$ are nonoverlapping on the ω axis for $i\neq j$. Similarly, the rectangular windows $\mathrm{rect}_P[\tau+rP+D]$ and $\mathrm{rect}_P[\tau+qP+D]$ are nonoverlapping on the n axis for $r\neq q$. Thus, for any particular values of ω and n, only a single term in the summations over i and r in (49) contributes to the magnitude of $X_2(n,\omega)$. By substituting $\zeta_i=\rho_i\exp\omega_i$ in (49), the magnitude and phase of the WB-STFT can be written as

$$\begin{aligned}A(n_0+\tau,\omega)&=\left|X_2(n_0+\tau,\omega)\right|\\&=\sum_{i=1}^{Q}\left|H(\omega_i-\omega)\right|\left|\tilde{a}_i\right|\cdot\sum_{r}\rho_i^{\tau+rP+D}\,\mathrm{rect}_P\left[\tau+rP+D\right]\end{aligned} \tag{50}$$

and

$$\begin{aligned}\theta(n_0+\tau,\omega)&=\arg\left[X_2(n_0+\tau,\omega)\right]\\&=\arg\left[H(\omega_i-\omega)\right]+\arg\left[\tilde{a}_i\right]\\&\quad+(\omega_i-\omega)\,\tau+\omega_iD-\omega n_0\\&\quad+\sum_{r}\omega_i rP\,\mathrm{rect}_P\left[\tau+rP+D\right].\end{aligned} \tag{51}$$

For a fixed value of time ($n=n_0,\tau=0$), the magnitude of the WB-STFT as a function of frequency consists of weighted and shifted images of $H(\omega)$, one centered at each of the Q formant frequencies. It is important to note that, because the shape of the resulting "formant" peaks in the WB-STFT depends primarily on the analysis filter, the bandwidths of these peaks do not provide good estimates of the true formant bandwidths.

For a fixed value of frequency, the magnitude of the WB-STFT as a function of time is quasi-periodic with local pitch period $P(n)$. The magnitude of each period is a decaying exponential with a time constant corresponding to the formant contained in the neighborhood of the particular frequency ω at which $X_2(n,\omega)$ is observed.

The phase of the WB-STFT can be expressed as the sum of three components:

$$\theta(n_0+\tau,\omega)=\alpha(n_0+\tau,\omega)+\vartheta(n_0+\tau,\omega)+\varphi(n_0+\tau,\omega) \tag{52a}$$

where

$$\alpha(n_0+\tau,\omega)=\arg\left[H(\omega_i(n_0)-\omega)\right]+\arg\left[\tilde{a}_i(n_0)\right] \tag{52b}$$

$$\vartheta(n_0+\tau,\omega)=(\omega_i(n_0)-\omega)\,\tau+\omega_i(n_0)\,D(n_0)-\omega n_0 \tag{52c}$$

and

$$\varphi(n_0+\tau,\omega)=\omega_i(n_0)\cdot\sum_{r} rP(n_0)\,\mathrm{rect}_P\left[\tau+rP(n_0)+D(n_0)\right]. \tag{52d}$$

As in the NB case, $\alpha(n,\omega)$ is a slowly time-varying phase component, and $\vartheta(n,\omega)$ is a linear phase component which satisfies the recursion relation

$$\vartheta(n_0+\tau,\omega)=\vartheta(n_0,\omega)+(\omega_i(n_0)-\omega)\,\tau, \tag{53}$$

but with slope $(\omega_i(n_0)-\omega)$ rather than $(k\Omega(n_0)-\omega)$. In contrast to the NB case, there is also a third phase component, $\varphi(n,\omega)$, that corresponds to phase jumps of $\omega_i(n_0)$ with each pitch period.

B. Short-Time Fourier Analysis of Unvoiced Speech

Unvoiced speech is modeled as a quasi-stationary random process characterized by its second moments. Recalling the interpretation of the STFT as the output of a demodulator followed by a low-pass filter (Fig. 4), the STFT of unvoiced speech can be interpreted as a set (indexed on ω) of stochastic time series (in n). This section develops the (second-order) statistics of the STFT of unvoiced speech signals.

Suppose, for the moment, that $x(n)$ is a stationary random process. It can be shown that the STFT of $x(n)$ is also a stationary random process for each ω, i.e.,

$$E\{X_2(n+\tau,\omega)\,X_2^*(n,\omega)\}\ \text{is independent of}\ n. \tag{54}$$

Unfortunately, the time series corresponding to the STFT of $x(n)$ evaluated at two different values of ω are not jointly stationary, i.e.,

$$E\{X_2(n+\tau, \omega_1) X_2^*(n, \omega_2)\} \quad \text{depends on } n \quad \text{for } \omega_1 \neq \omega_2. \tag{55}$$

The STFT can be expressed as

$$\begin{aligned} X_2(n, \omega) &= \sum_{m=-\infty}^{\infty} h(n-m) x(m) \exp[-j\omega m] \\ &= \left\{ \sum_{m=-\infty}^{\infty} x(m) h(n-m) \exp[j\omega(n-m)] \right\} \\ &\quad \cdot \exp[-j\omega n] \\ &= \{x(n) *_n (h(n) \exp[j\omega n])\} \exp[-j\omega n] \end{aligned}$$

and, therefore, can be interpreted as the response to $x(n)$ of the linear time-invariant bandpass filter $h(n) \exp[j\omega n]$, multiplied (demodulated) by $\exp[-j\omega n]$. Because the responses of two linear time-invariant systems to the same stationary input are jointly stationary, the two random processes

$$X_2(n, \omega_1) \exp[j\omega_1 n]$$

and

$$X_2(n, \omega_2) \exp[j\omega_2 n]$$

are jointly stationary. Thus,

$$\begin{aligned} &E\{(X_2(n+\tau, \omega_1) \exp[j\omega_1(n+\tau)]) (X_2(n, \omega_2) \\ &\qquad \cdot \exp[j\omega_2 n])^*\} \\ &\quad = E\{X_2(n+\tau, \omega_1) X_2^*(n, \omega_2)\} \\ &\qquad \cdot \exp[j\omega_1 \tau] \exp[j(\omega_1 - \omega_2) n] \end{aligned} \tag{56}$$

is independent of n, and $X_2(n, \omega_1)$ and $X_2(n, \omega_2)$ are jointly stationary "to within a modulation by $\exp[j(\omega_1 - \omega_2) n]$." Letting $\omega_1 = \omega - \epsilon/2$ and $\omega_2 = \omega + \epsilon/2$, the right-hand side of (56) becomes

$$\begin{aligned} &E\{X_2(n+\tau, \omega_1) X_2^*(n, \omega_2)\} \exp[j\omega_1 \tau] \exp[j(\omega_1 - \omega_2) n] \\ &\quad = E\left\{X_2\left(n+\tau, \omega - \frac{\epsilon}{2}\right) X_2^*\left(n, \omega + \frac{\epsilon}{2}\right)\right\} \\ &\qquad \cdot \exp\left[j\left(\omega - \frac{\epsilon}{2}\right)\tau\right] \exp[-j\epsilon n] \\ &\quad = E\left\{X_2\left(n+\tau, \omega - \frac{\epsilon}{2}\right) X_2^*\left(n, \omega + \frac{\epsilon}{2}\right)\right\} \\ &\qquad \cdot \exp\left[-j\left(n + \frac{\tau}{2}\right)\epsilon\right] \exp[j\omega\tau]. \end{aligned} \tag{57}$$

Because (57) is independent of n, a more convenient quantity than the cross-correlation function (55) of the STFT is the modified correlation function, defined by

$$\begin{aligned} K_x(n, \omega, \tau, \epsilon) &= E\left\{X_2\left(n+\tau, \omega - \frac{\epsilon}{2}\right) X_2^*\left(n, \omega + \frac{\epsilon}{2}\right)\right\} \\ &\quad \cdot \exp\left[-j\left(n + \frac{\tau}{2}\right)\epsilon\right] \end{aligned} \tag{58}$$

which is independent of n. The cross-correlation function of the STFT can, therefore, be expressed as

$$\begin{aligned} &E\left\{X_2\left(n+\tau, \omega - \frac{\epsilon}{2}\right) X_2^*\left(n, \omega + \frac{\epsilon}{2}\right)\right\} \\ &\quad = K_x(n, \omega, \tau, \epsilon) \exp\left[j\left(n + \frac{\tau}{2}\right)\epsilon\right]. \end{aligned} \tag{59}$$

If $x(n)$ is no longer stationary, but quasi-stationary, then the modified correlation function (58) becomes a "slowly varying" function of n. By substituting the definition of the STFT (23) into the definition of the modified correlation function (58), the modified correlation function can be expressed in terms of the time-varying power spectrum of $x(n)$ and the analysis filter $H(\omega)$ as

$$\begin{aligned} K_x(n, \omega, \tau, \epsilon) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(n, \omega + \varphi) H\left(\varphi + \frac{\epsilon}{2}\right) \\ &\quad \cdot H^*\left(\varphi - \frac{\epsilon}{2}\right) \exp[j\varphi\tau] \, d\varphi. \end{aligned} \tag{60}$$

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. This expansion is useful for determining the power spectrum of a signal synthesized from a nonlinear modification of the STFT of unvoiced speech [12], [13]. Define the moments up to second order of the analysis filter as

$$J_h = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(\omega)|^2 \, d\omega = \sum_{n=-\infty}^{\infty} |h(n)|^2 \tag{61a}$$

$$M_h = \frac{1}{2\pi J_h} \int_{-\pi}^{\pi} \omega \, |H(\omega)|^2 \, d\omega \tag{61b}$$

$$m_h = \frac{1}{J_h} \sum_{n=-\infty}^{\infty} n \, |h(n)|^2 \tag{61c}$$

$$D_h^2 = \frac{1}{2\pi J_h} \int_{-\pi}^{\pi} \omega^2 \, |h(\omega)|^2 \, d\omega \tag{61d}$$

$$d_h^2 = \frac{1}{J_h} \sum_{n=-\infty}^{\infty} n^2 \, |h(n)|^2 \tag{61e}$$

and

$$\mu_h = \frac{-j}{4\pi J_h} \int_{-\pi}^{\pi} \omega [H'(m) H^*(\omega) - H(\omega) H'^*(\omega)] \, d\omega. \tag{61f}$$

Introducing a suitable time shift to $h(n)$ and frequency shift to $H(\omega)$ so that their first moments vanish, i.e.,

$$M_h = 0 \quad \text{and} \quad m_h = 0,$$

(60) can be expanded in the two-dimensional power series

$$K_x(n, \omega, \tau, \epsilon) = J_x(n, \omega) \{1 - \tfrac{1}{2} [D_h^2 \tau^2 + 2\mu_h \tau\epsilon + d_h^2 \epsilon^2] + \cdots\} \tag{62}$$

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

$$J_x(n, \omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(n, \omega + \varphi) \, |H(\varphi)|^2 \, d\varphi, \tag{63}$$

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.

If 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 NB-STFT of voiced speech. Moreover, it is often convenient to define phase components $\alpha(n, \omega)$ and $\vartheta(n, \omega)$ for unvoiced speech as the values calculated by the same estimator used to calculate these quantities for voiced speech [12], [13]. The major difference between the voiced and unvoiced cases, however, is that unvoiced speech possesses no underlying harmonic structure, across the spectrum, as does voiced speech.

IV. Summary

This paper has developed a model for short-time Fourier analysis of sampled speech signals. A quasi-stationary representation for the speech signal was formulated that made few assumptions about the structure of the speech signal. It was assumed, only, that the speech signal could be modeled as the output of a linear time-varying system driven by a quasi-periodic impulse train or white noise, and that the time variations of the system parameters and the local pitch period were slow compared with the duration of the memory of the system. The techniques of short-time Fourier analysis were applied to this representation to obtain narrow-band and wide-band analyses of voiced speech and a second-order statistical analysis of unvoiced speech. For the case of wide-band analysis of voiced speech, the additional assumption of a formant structure for the speech yielded a simplified interpretation of the WB-STFT. In all cases, convenient quantities to interpret were the magnitude, phase, and instantaneous frequency of the short-time Fourier transform.

The model discussed here is applicable to a number of areas of speech processing. It provides a mathematical model of the speech spectrogram, one of the most widely used tools for speech analysis. It is potentially, a common framework for describing a variety of speech analysis/synthesis systems including phase vocoders, channel vocoders, subband coders, and transform coders. Finally, it has provided the framework for the design and implementation of a high-quality and robust system for time compression and expansion of speech [12], [13].

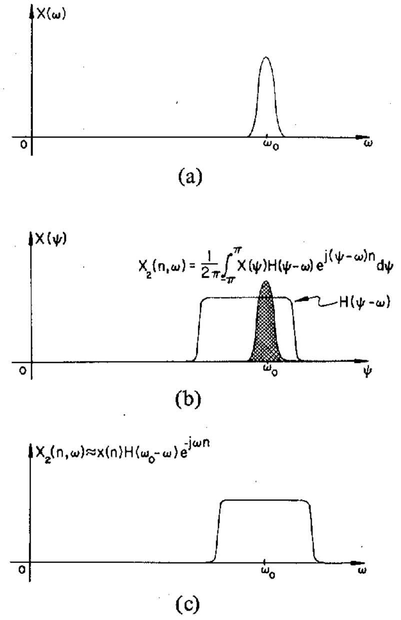

X(ω)
0
ω₀
ω
(a)
X(ψ)
X₂(n,ω) = 1/2π ∫₋π^π X(ψ)H(ψ-ω) e^{j(ψ-ω)n} dψ
H(ψ-ω)
0
ω₀
ψ
(b)
X₂(n,ω)≈x(n)H(ω₀-ω)e^{-jωn}
0
ω₀
ω
(c)

Fig. 7. (a) Fourier transform of a narrow-band signal $x(n)$. (b) Short-time Fourier transform of $x(n)$ as a convolution integral in the frequency domain with ψ as the variable of integration. (c) Short-time Fourier transform of $x(n)$.

Appendix
Short-Time Fourier Analysis of Narrow-Band Signals

In many applications, signals arise that occupy a narrow band of frequencies. One means of tracking such signals, or discriminating among such signals, is short-time Fourier analysis. The term "narrow-band signal" is often applied to such a signal when the bandwidth of the signal is small compared with the frequency resolution of interest. Within the context of short-time Fourier analysis, the term "narrow-band signal" will mean that the bandwidth of the signal is narrow compared with the bandwidth of the analysis filter. Let $x(n)$ represent such a signal occupying a narrow band of frequencies centered about ω_0, so that its Fourier transform, $X(\omega)$, is as illustrated in Fig. 7(a). The short-time Fourier transform of $x(n)$, as defined by (23), can be expressed in terms of $X(\omega)$ and $H(\omega)$ [19] as

$$X_2(n, \omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\psi) H(\psi - \omega) \exp\,[j(\psi - \omega)\, n] \, d\psi. \tag{A1}$$

Interpreting (A1) as a convolution integral and referring to Fig. 7(b), $X_2(n, \omega)$ can be evaluated by approximating $H(\psi - \omega)$ as constant in the region overlapping $X(\psi)$. Thus, the product $X(\psi)\, H(\psi - \omega)$ is replaced by $X(\psi)\, H(\omega_0 - \omega)$, and (A1) becomes

$$\begin{aligned} X_2(n, \omega) &\approx \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\psi) H(\omega_0 - \omega) \exp\,[j(\psi - \omega)\, n] \, d\psi \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\psi) \exp\,[j\psi n] \, d\psi \\ &\quad \cdot H(\omega_0 - \omega) \exp\,[-j\omega n]. \end{aligned}$$

Integrating over ψ gives the result

$$X_2(n, \omega) \approx x(n)\, H(\omega_0 - \omega) \exp\,[-j\omega n]. \tag{A2}$$

Thus, the short-time Fourier transform $X_2(n, \omega)$ of a narrowband signal $x(n)$ is approximately the product of the demodulated (frequency-shifted) signal $x(n) \exp[-j\omega n]$ multiplied by the shifted window $H(\omega_0 - \omega)$. For a fixed value of n, $X_2(n, \omega)$ is illustrated in Fig. 7(c) as the image of $H(-\omega)$ shifted by ω_0 and weighted by $x(n) \exp[-j\omega n]$.

REFERENCES

[1] W. Koenig, H. K. Dunn, and L. Y. Lacy, "The sound spectrograph," *J. Acoust. Soc. Amer.*, vol. 17, pp. 19-49, July 1946.

[2] R. K. Potter, G. A. Kopp, and H. C. Green, *Visible Speech*. New York: Van Nostrand, 1947; republished by Dover, 1966.

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

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

[5] 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.

[6] 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.

[7] B. Gold and C. M. Rader, "The channel vocoder," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, pp. 148-160, Dec. 1967; reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel, Eds. New York: IEEE Press, 1979.

[8] R. W. Schafer and L. R. Rabiner, "Design and simulation of a speech analysis-synthesis system based on short-time Fourier Analysis," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 165-174, June 1973.

[9] 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.

[10] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Amer.*, vol. 60, pp. 911-918, Oct. 1976.

[11] J. A. Moorer, "The use of the phase vocoder in computer music applications," presented at the 55th Conv. Audio Engineering Soc., preprint no. 1146 (E-1), Oct. 1976.

[12] M. R. Portnoff, "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.

[13] —, "Time-scale modification of speech based on short-time Fourier analysis," this issue, pp. 374-390.

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

[15] L. A. Zadeh, "A general theory of signal transmission systems," *J. Franklin Inst.*, vol. 253, pp. 293-312, Apr. 1952.

[16] T. Kailath, "Sampling models for linear time-variant filters," Res. Lab. Electron., M.I.T., Cambridge, Tech. Rep. 352, May 1959.

[17] —, "Channel characterization: Time-variant dispersive channels," in *Lectures on Communication System Theory*, E. J. Baghdaddy, Ed. New York: McGraw-Hill, 1961, pp. 95-123.

[18] A. Gersho and N. DeClaris, "Duality concepts in time-varying linear systems," in *1964 IEEE Int. Conv. Rec.*, part 1, pp. 344-356.

[19] M. R. Portnoff, "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.

[20] —, "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.

[21] 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.

[22] H. K. Dunn, "Methods of measuring vowel formant bandwidths," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1737-1746, Dec. 1961; reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel, Eds. New York: IEEE Press, 1979.

Michael R. Portnoff (S'69-M'77-M'78) was born in Newark, NJ, on July 1, 1949. He was educated at the Massachusetts Institute of Technology, Cambridge, MA, receiving the S.B., S.M., and E.E. degrees in electrical science and engineering in 1973, and the Sc.D. degree in electrical engineering and computer science in 1978.

From 1969 to 1971, he was a cooperative student at Bell Telephone Laboratories, Inc. From 1971 to 1978, he was a Research Assistant in the M.I.T. Research Laboratory of Electronics, Digital Signal Processing Group, and a Teaching Assistant in the M.I.T. Department of Electrical Engineering and Computer Science. From 1978 to 1979, he was a Research Associate in the Digital Signal Processing Group at M.I.T. In 1979 he joined the University of California Lawrence Livermore Laboratory, Livermore, CA, where he is currently a Staff Member in the Engineering Research Division. His research interests are in the theory of digital signal processing and its application to speech, image, and seismic signal processing.

Dr. Portnoff received the 1977 IEEE Browder J. Thompson Memorial Prize Award and the 1980 IEEE ASSP Society Paper Award, and is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.