SPEECH ANALYSIS

The speech wave production mechanism can be divided into three stages: sound source production, articulation by vocal tract, and radiation from the lips and/or nostrils. Sound sources are either voiced or unvoiced. A voiced sound source can be modeled by a generator of pulses or asymmetrical triangular waves which are repeated at every fundamental period. The peak value of the source wave corresponds to the loudness of the voice. An unvoiced sound source, on the other hand, can be modeled by a white noise generator, the mean energy of which corresponds to the loudness of the voice. Articulation can be modeled by the cascade or parallel connection of several single-resonance or antiresonance circuits, which can be realized through a multistage digital filter. Finally, radiation can be modeled as arising from a piston sound source attached to an infinite, plane baffle.

The speech wave can be changed into a processible object by converting it into an electrical signal using a microphone. The electrical signal is usually transformed from an analog into a digital signal prior to almost all speech processing for two reasons. First, digital techniques facilitate highly sophisticated signal processing which cannot otherwise be realized by analog techniques. Second, digital processing is far more reliable and can be accomplished by using a compact circuit. Rapid development of computers and integrated circuits in conjunction with the growth of digital communications networks have encouraged the application of digital processing techniques to speech processing.

SPECTRAL ANALYSIS

Spectral Structure of Speech

The speech wave is usually analyzed using spectral features, such as the frequency spectrum and autocorrelation function, instead of directly using the waveform. There are two important reasons for this. One is that the speech wave is considered to be reproducible by summing sinusoidal waves, the amplitudes and phases of which change slowly. The other is that the critical features for perceiving speech by the human ear are mainly included in the spectral information, with the phase information rarely playing a key role.

The power spectral density in a short interval—that is, the short-time spectrum of speech—can be regarded as the product of two elements: the spectral envelope, which slowly changes as a function of frequency, and the spectral fine structure, which changes rapidly. The spectral fine structure produces periodic patterns for voiced sounds but not for unvoiced sounds, as shown in Fig. 1. The spectral envelope, or the overall spectral feature, reflects not only the resonance and antiresonance characteristics of the articulatory organs,

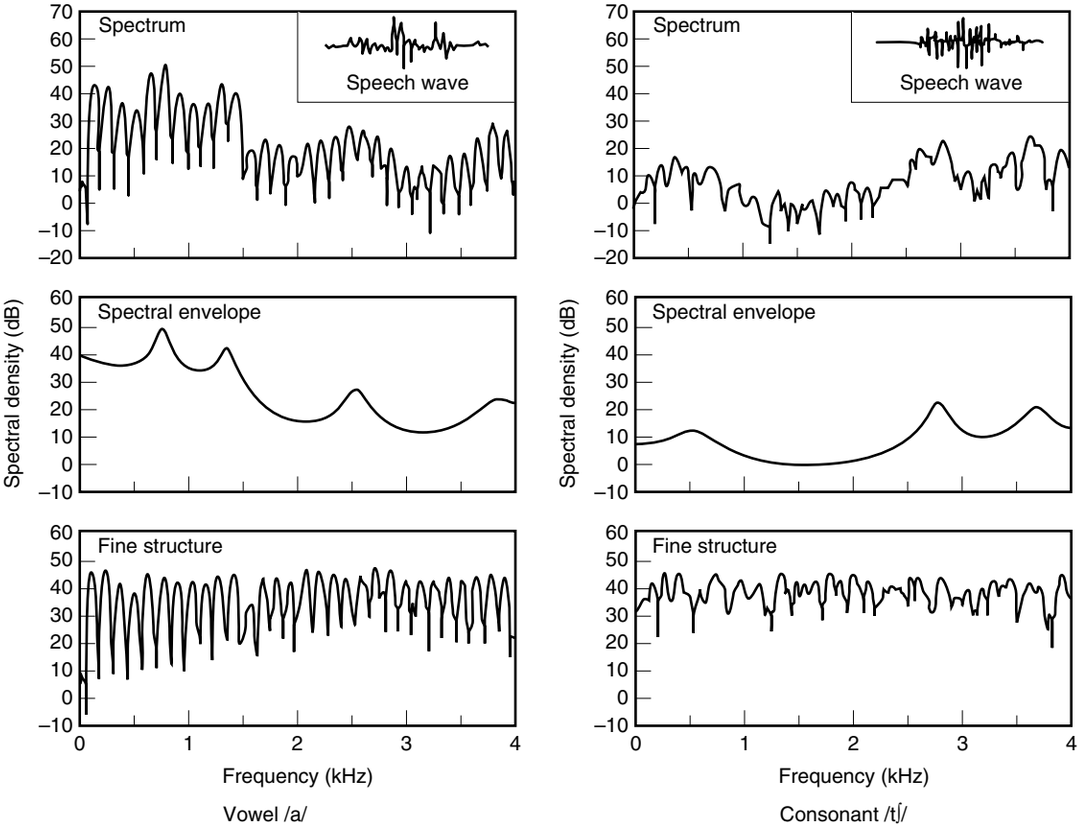

Figure 1. Structure of short-time spectra for male voices when uttering vowel /a/ and consonant /$t\int$/. The short-time spectra of speech can be regarded as the product of the spectral envelope and the spectral fine structure.

but also the overall shape of the glottal source spectrum and radiation characteristics at the lips and nostrils. On the other hand, the spectral fine structure corresponds to the periodicity of the sound source.

Methods for spectral envelope extraction can be divided into parametric analysis (PA) and nonparametric analysis (NPA). In PA, a model which fits the objective signal is selected and applied to the signal by adjusting the feature parameters representing the model. On the other hand, NPA methods can generally be applied to various signals since they do not model the signals. If the model exactly fits the objective signal, PA methods can represent the features of the signal more effectively than can NPA methods.

Autocorrelation and Fourier Transform

When a sampled time sequence is written as $x(n)$ (n is an integer), its autocorrelation function $\phi(m)$ is defined as

$$\phi(m) = \frac{1}{N} \sum_{n=0}^{N-1-|m|} x(n)x(n+|m|), \qquad |m| = 0, 1, \ldots, N-1 \tag{1}$$

where N is the number of samples in the short-time analysis interval. The short-time spectra $S(\lambda)$ and $\phi(m)$ constitute the Fourier transform pair (Wiener–Khintchine theorem):

$$S(\lambda) = \frac{1}{2\pi} \sum_{m=-(N-1)}^{N-1} \phi(m) \cos \lambda m \tag{2}$$

and

$$\phi(m) = \int_{-\pi}^{\pi} S(\lambda) \cos \lambda m \, d\lambda \tag{3}$$

where λ is a normalized radian frequency which can be represented by $\lambda = 2\pi f \Delta T$ (f is a real frequency, and ΔT is a sampling period). $S(\lambda)$ is usually computed directly from the speech wave using the discrete Fourier transform (DFT) facilitated by the fast Fourier transform (FFT) algorithm:

$$S(\lambda) = \frac{1}{2\pi N} \left| \sum_{n=0}^{N-1} x(n) e^{-j\lambda n} \right|^2 \tag{4}$$

The autocorrelation function can also be calculated more efficiently by using the DFT (FFT) compared with the conventional correlation calculation method when higher-order correlation elements are needed. With this method, the autocorrelation function is obtained as the inverse Fourier trans-

form of the short-time spectrum, which is calculated by using Eq. (4).

Window Function

In order to extract the N-sample interval from the speech wave for calculating the spectral features, the speech wave must be multiplied by an appropriate time window. Therefore, $x(n)$, indicated in Eqs. (1) and (14) for calculating $\phi(m)$ and $S(\lambda)$, respectively, is usually not the original waveform but rather the waveform multiplied by the window function.

The Hamming window, $W_{\mathrm{H}}(n)$, defined as

$$W_{\mathrm{H}}(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{N-1}\right) \tag{5}$$

is usually used as the window function for speech analysis. Another window, called the *Hanning window,*

$$W_{\mathrm{N}}(n) = 0.5 - 0.5\cos\left(\frac{2n\pi}{N-1}\right) \tag{6}$$

is also employed.

When the waveform is multiplied by either the Hamming or the Hanning window, the effective analysis interval length becomes approximately 40% shorter since the waveforms near both ends of the window are attenuated. This results in a consequent 40% decrease in the frequency resolution.

Hence, the multiplication of the speech wave by an appropriate window reduces the spectral fluctuation due to the variation of the pitch excitation position within the analysis interval. This is effective in producing stable spectra during the analysis of voiced sounds featuring pitch periodicity. Since multiplication by the window function decreases the effective analysis interval length, the analysis interval should be overlapping and shifted along the speech wave to facilitate tracking the time-varying spectra.

The short-time analysis interval multiplied by a window function and extracted from the speech wave is called a *frame.* The length of the frame is referred to as the *frame length,* and the frame shifting interval is termed the *frame interval* or *frame period.*

A block diagram of a typical speech analysis procedure is shown in Fig. 2. Also indicated at each stage are typical parameter values.

Digital Filter Bank

The digital filter bank—more specifically, a set of bandpass filters—is one of the NPA techniques. The filter bank requires a relatively small amount of calculation and is therefore quite suitable for hardware implementation. Since there is a tradeoff between the time and frequency resolution of each bandpass filter, it is necessary to design various parameters according to the purposes intended. Generally, the bandpass filters are arranged so that the center frequencies are distributed with equal intervals on the logarithmic frequency scale, Mel scale or Bark scale, taking human auditory characteristics into account, and so that the 3 dB attenuation points of the adjacent filters coincide. The output of each bandpass filter is rectified, smoothed by root mean square (rms) value calculation, and sampled every 5 ms to 20 ms to obtain values which represent the spectral envelope.

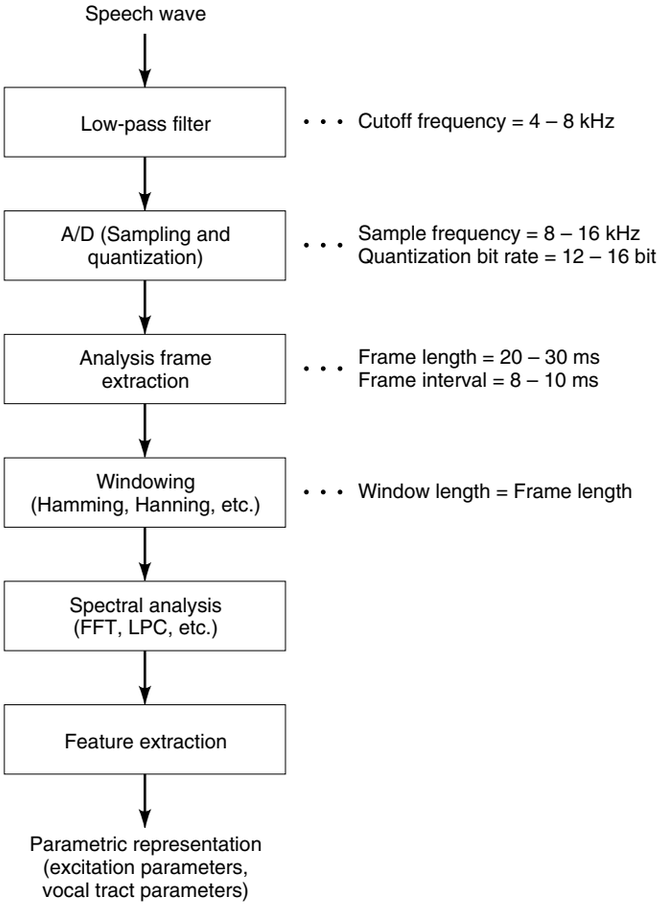

Figure 2. Block diagram of a typical speech analysis procedure. Typical parameter values at each stage are also indicated.

Sound Spectrogram Analysis

Sound spectrogram analysis is a method for plotting the time function of the speech spectrum using density plots. Figure 3 is an example of sound spectrograms for the Japanese word /ikioi/ uttered by a male speaker. The magnitude of the frequency component is illustrated by darkness; in other words, the darker areas reveal higher-intensity frequency components.

Usually the bandwidth of the bandpass filter for the frequency analysis (i.e., the frequency resolution) is either 300 Hz or 45 Hz, depending on the purpose of the analysis. When the frequency resolution is 300 Hz, the effective length of the speech analysis interval is roughly 3 ms; and when the resolution is 45 Hz, the length becomes 22 ms. Because of the tradeoff occurring between the frequency and time resolutions, the pitch structure of speech is indicated by (1) a vertically striped fine repetitive pattern along the time axis in the case of the 300 Hz frequency resolution and (2) a horizontally striped equally fine repetitive pattern along the frequency axis in the case of the 45 Hz resolution.

Many of the sound spectrograms originally produced by analog technology using the sound spectrograph are now produced by digital technology through computers and their peripherals. The digital method is particularly beneficial because it permits easy adjustment of various conditions and

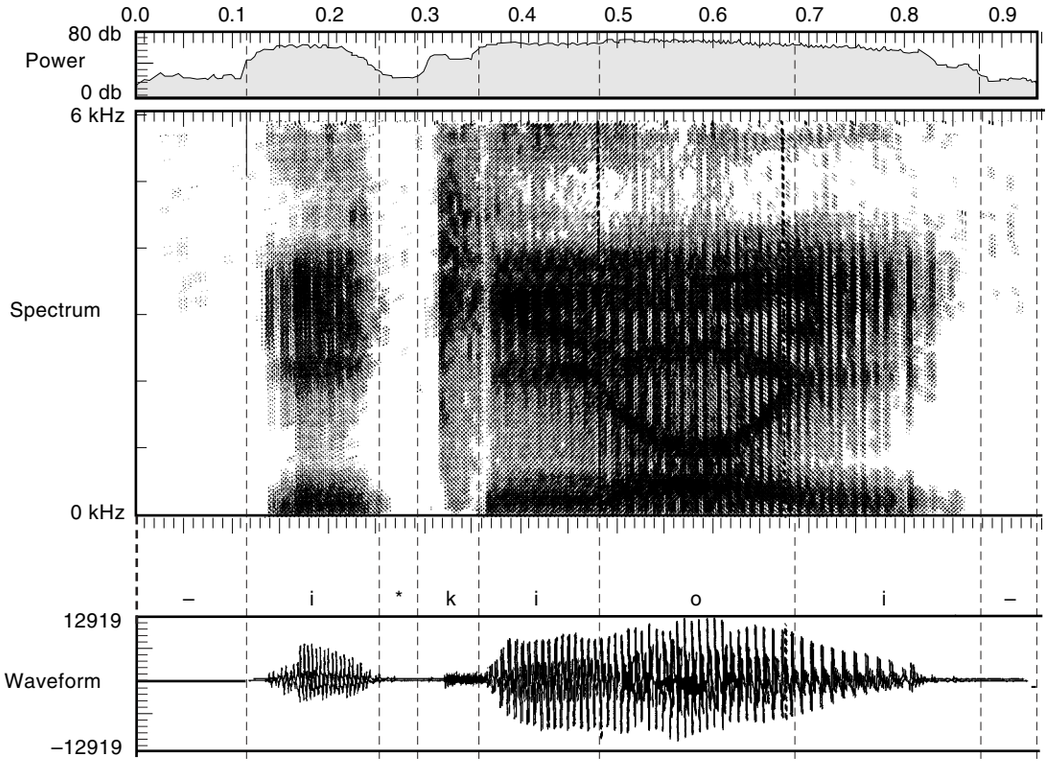

Figure 3. An example of wide-band sound spectrogram for a male voice when uttering the Japanese phase /ikioi/. The magnitude of the frequency component is illustrated by darkness.

also because the spectrograms can be produced sequentially and automatically with good reproducibility.

Zero-Crossing Analysis

The zero-crossing number of the speech wave in a predetermined time inteval, which is counted as the number of times when adjacent sample points have different positive and negative signs, approximately corresponds to the frequency of the major spectral component. Based on this principle, formant frequencies can be estimated by zero-crossing analysis as follows. First, the speech wave is passed through a set of four- or five-octave band-pass filters, and the power and zero-crossing number of the rectified and smoothed output of each filter are measured at short intervals, such as 10 ms. When the power of a filter exceeds the predetermined threshold, this frequency range is regarded as having a formant, with the formant frequency being estimated by the zero-crossing rate. This zero-crossing rate can also be used to detect the periodicity of the sound source as well as to estimate the fundamental period. Although the zero-crossing analysis method is well suited to hardware implementation, its drawback is that it is sensitive to additive noise.

CEPSTRUM

Principles of Cepstrum Analysis

The cepstrum, or cepstral coefficient, $c(\tau)$, is defined as the inverse Fourier transform of the short-time logarithmic amplitude spectrum $|X(\omega)|$ (1–3). The term *cepstrum* is essentially a coined word which includes the meaning of the inverse transform of the logarithmic spectrum. The independent parameter for the cepstrum is called *quefrency,* which is obviously formed from the word *frequency.* Since the cepstrum is the inverse transform of the frequency domain function, the quefrency becomes the time-domain parameter. The special feature of the cepstrum is that it allows for the separate representation of the spectral envelope and fine structure.

Voiced speech $x(t)$ can be regarded as the response of the vocal tract articulation equivalent filter driven by the pseudo-periodic source $g(t)$. Then $x(t)$ can be given by the convolution of $g(t)$ and vocal tract impulse response $h(t)$ as

$$x(t) = \int_0^t g(\tau)h(t-\tau)\,d\tau \tag{7}$$

which is equivalent to

$$X(\lambda) = G(\lambda)H(\lambda) \tag{8}$$

where $X(\lambda)$, $G(\lambda)$, and $H(\lambda)$ are the Fourier transforms of $x(t)$, $g(t)$, and $h(t)$, respectively.

If $g(t)$ is a periodic function, $|X(\lambda)|$ is represented by line spectra, the frequency intervals of which are the reciprocal of the fundamental period of $g(t)$. Therefore, when $|X(\lambda)|$ is calculated by the Fourier transform of a sampled time sequence for a short speech wave period, it exhibits sharp peaks with equal intervals along the frequency axis. Its logarithm $\log|X(\lambda)|$ is

$$\log|X(\lambda)| = \log|G(\lambda)| + \log|H(\lambda)| \tag{9}$$

The cepstrum, which is the inverse Fourier transform of $\log |X(\lambda)|$, is

$$c(\tau) = F^{-1} \log |X(\lambda)| = F^{-1} \log |G(\lambda)| + F^{-1} \log |H(\lambda)| \quad (10)$$

where F is the Fourier transform. The first and second terms on the right side of Eq. (9) correspond to the spectral fine structure and the spectral envelope, respectively. The former is the periodic pattern, and the latter is the global pattern along the frequency axis. Accordingly, large differences occur between the inverse Fourier transform functions of both elements indicated in Eq. (10).

Principally, the first function on the right side of Eq. (10) indicates the formation of a peak in the high-quefrency region, and the second function represents a concentration in the low-quefrency region from 0 to 2 or 4 ms. The fundamental period of the source $g(t)$ can then be extracted from the peak at the high-quefrency region. On the other hand, the Fourier transform of the low-quefrency elements produces the logarithmic spectral envelope from which the linear spectral envelope can be obtained through the exponential transform. The maximum order of low-quefrency elements used for the transform determines the smoothness of the spectral envelope. The process of separating the cepstal elements into these two factors is called *liftering,* which is derived from filtering.

When the cepstrum is calculated by the DFT, it is necessary to set the base value of the transform, N, large enough to eliminate the aliasing similar to that produced during waveform sampling. The cepstrum then becomes

$$C_n = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| e^{j2\pi kn/N}, \qquad 0 \le n \le N-1 \quad (11)$$

The process steps for extracting the fundamental period and spectral envelope using the cepstral method are given in Fig. 4.

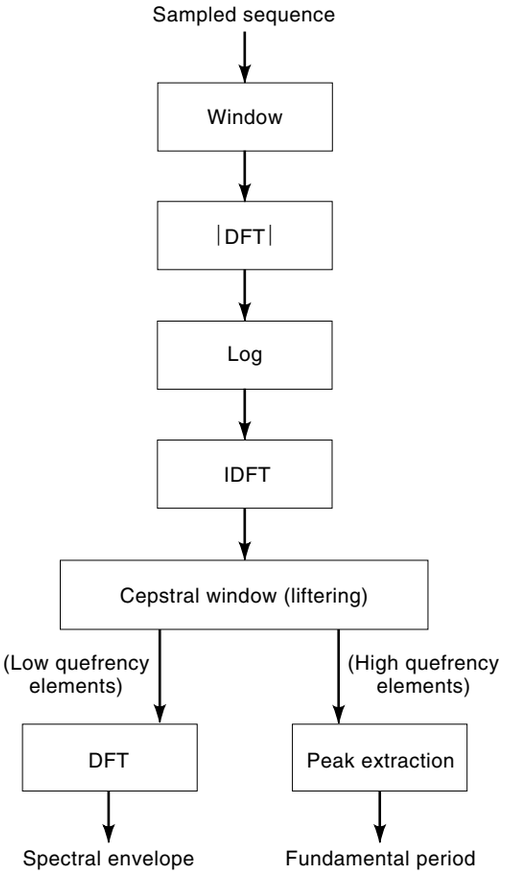

Figure 4. Block diagram of cepstrum analysis for extracting the spectral envelope and fundamental period. The logarithmic spectral envelope can be produced by the Fourier transform of the low-quefrency elements, and the fundamental period of the voice source can be extracted from the peak at the high-quefrency region.

LPC Cepstrum

Let us consider the cepstrum in a special case in which $X(\lambda) = H(z) \,|_{z = \exp(j\lambda T)}$. Here, $H(z)$ is the z-transform of the impulse response of the all-pole speech production system estimated by the linear predictive coding (LPC) analysis method [see section entitled "Linear Predictive Coding (LPC) Analysis"]. Accordingly,

$$H(z) = \frac{1}{1 + \sum_{i=1}^{p} \alpha_i z^{-1}} \quad (12)$$

Equation (12) means that the all-pole spectrum $H(z)$ is used for the spectral density of the speech signal. This is accomplished by expanding the cepstrum into a complex form by replacing the DFT, logarithmic transform, and inverse discrete Fourier transform (IDFT) in Fig. 4 with a dual z-transform, complex logarithmic transform, and inverse dual z-transform, respectively (4). When this complex cepstrum for a time sequence $x(n)$ is represented by $\hat{c}_n$, and their dual z-transforms are indicated by $X(z)$ and $C(z)$, respectively, we obtain

$$\hat{C}(z) = \log[X(z)] \quad (13)$$

If we now differentiate both parts of this equation by z^{-1} and then multiply by $X(z)$, we have

$$X(z)\hat{C}'(z) = X'(z) \quad (14)$$

This equation permits recursive equations to be obtained:

$$\begin{aligned} \hat{c}_1 &= -\alpha_1 \\ \hat{c}_n &= -\alpha_n - \sum_{m=1}^{n-1} \left(1 - \frac{m}{n}\right) \alpha_m \hat{c}_{n-m}, \qquad 1 < n \le p \\ \hat{c}_n &= -\sum_{m=1}^{p} \left(1 - \frac{m}{n}\right) \alpha_m \hat{c}_{n-m}, \qquad p < n \end{aligned} \quad (15)$$

This cepstrum is referred to as the LPC cepstrum, since it is derived through the LPC model. The original cepstrum is sometimes called the FFT cepstrum to distinguish it from the LPC cepstrum.

Figure 5 compares the spectral envelope calculated using the cepstrum directly extracted from the waveform with that calculated using the LPC cepstrum (5). In this figure, the short-time spectrum and the spectral envelope extracted by LPC (maximum likelihood method) are also shown for refer-

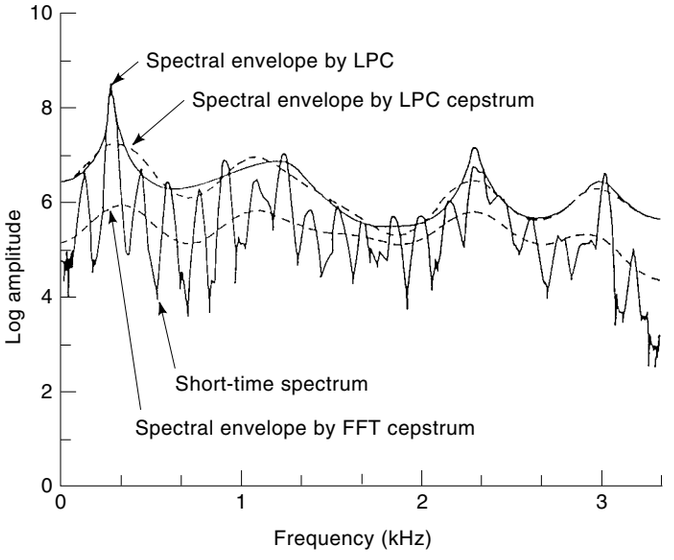

Figure 5. Comparison of spectral envelopes by LPC, LPC cepstrum, and FFT cepstrum methods. The spectral envelope derived from the LPC cepstrum follows the spectral peaks more strictly than does the spectral envelope obtained through the FFT cepstrum.

ence. The spectral envelope derived from the LPC cepstrum clearly tends to follow the spectral peaks more strictly than does the spectral envelope obtained through the FFT cepstrum.

ANALYSIS-BY-SYNTHESIS

Analysis-by-synthesis (A-b-S), presented in Fig. 6, is the process of determining the parameters which characterize the system based on an assumed signal production model (6). The model parameters are adjusted in the course of iterative feedback control so that the error between the observed value and that produced by the model is minimized. Important in A-b-S are selection of the assumed production model, the initial parameter values, the error evaluation measure, and the minimization algorithm. A-b-S is useful not only for speech parameter extraction but also for many applications in which a production model can be used.

PITCH EXTRACTION

Although the accurate extraction of the fundamental frequency (pitch extraction) has been one of the most important study concerns since the beginning of speech analysis research, no definite approach has yet been established. This difficulty with pitch extraction stems from three factors. First, vocal cord vibration does not necessarily have complete periodicity especially at the beginning and end of voiced sounds. Second, it is difficult to extract the vocal cord source signal from the speech wave separately from the vocal tract effects. Third, the dynamic range of the fundamental frequency is very large.

With these factors in mind, recent pitch extraction research has been undertaken from three viewpoints. One is how to reliably extract the periodicity of quasiperiodic signals. Another is how to correct the pitch extraction error owing to the disturbance of periodicity. The other is how to remove the vocal tract (formant) effects. Major errors in pitch extraction are classified into double-pitch and half-pitch errors. The former are those errors occurring when extracting a frequency which is twice as large as the actual value. The latter are errors arising when extracting the half-value of the actual fundamental frequency. The tendency toward which error is most apt to occur depends on the extraction method employed.

The major pitch extraction methods are outlined in Table 1 (7). They can generally be grouped into waveform processing (a), correlation processing (b), and spectral processing (c). Group (a) is composed of methods for detecting the periodic peaks in the waveform. Group (b) methods are those most widely used in digital signal processing of speech, since the correlation processing is unaffected by phase distortion in the waveform. Among the methods in Group (c), the principle of pitch extraction using cepstral analysis has already been described in the section entitled "Principles of Cepstrum

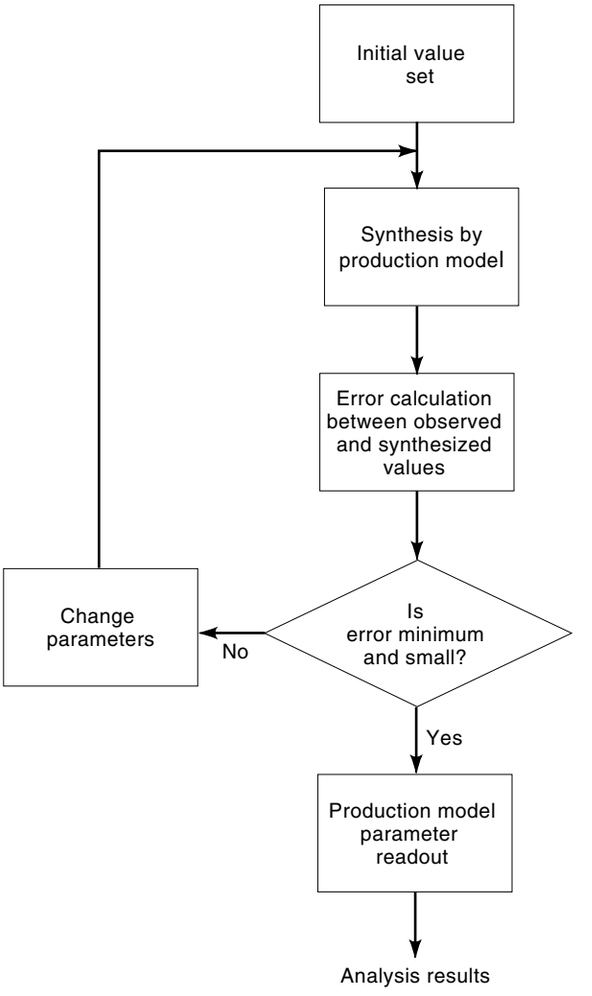

Figure 6. Principle of analysis-by-synthesis method. The model parameters are adjusted in the course of iterative feedback control so that the error between the observed value and that produced by the model is minimized.

Table 1. Classification of Major Pitch Extraction Methods and Their Principal Features

Classification	Pitch Extraction Method	Principal Features
(a) Waveform processing	Parallel processing method	Uses majority rule for pitch periods extracted by many kinds of simple waveform peak detectors.
	Data reduction method	Removes superfluous waveform data based on various logical processing and leaves only pitch pulses.
	Zero-crossing count method	Utilizes iterative patterns in waveform zero-crossing rate.
(b) Correlation processing	Autocorrelation method	Employs autocorrelation function of waveform. Applies center and peak clipping for spectrum flattening and computation simplification.
	Modified correlation method	Utilizes autocorrelation function for residual signal of LPC analysis. Computation is simplified by LPF and polarization.
	SIFT (simplified inverse filter tracking) algorithm	Applies LPC analysis for spectrum flattening after down-sampling of speech wave. Time resolution is recovered by interpolation.
	AMDF method	Uses average magnitude differential function (AMDF) for speech or residual signal for periodicity detection.
(c) Spectrum processing	Cepstrum method	Separates spectral envelope and fine structure by inverse Fourier transform of log-power spectrum.
	Period histogram method	Utilizes histogram for harmonic components in spectral domain. Pitch is decided as the common divisor for harmonic components.

Analysis." The modified correlation method and simplified inverse filter tracking (SIFT) algorithm (8), which are correlation methods, and the cepstral method are generally the most efficient since they explicitly remove the vocal tract effects. The modified correlation method will be described in detail in the section entitled "Source Parameter Estimation from Residual Signals."

The voiced/unvoiced decision is usually made using a method for pitch extraction, since, for the sake of simplicity, the cues for periodic/unperiodic decision are normally regarded as those utilized for voiced/unvoiced decisions. The peak values of the autocorrelation or modified autocorrelation functions are generally implemented in the decision. Because these methods do not work effectively for unperiodic voiced sounds, improvement in decision accuracy has been attempted by employing several other parameters as additional cues (9). These parameters include the speech energy, zero-crossing rate, first-order autocorrelation function, first-order linear predictor coefficient, and energy of the residual signal.

LINEAR PREDICTIVE CODING (LPC) ANALYSIS

Principles of LPC Analysis

Since the term *linear prediction* was first coined by N. Wiener (10), the technique has become popularly employed in a wide range of applications based on a number of formulations. This technique, first used for speech analysis and synthesis by Itakura and Saito (11) and Atal and Schroeder (12), has produced a very large impact on every aspect of speech research (13). The importance of linear prediction stems from the fact that the speech wave and spectrum characteristics can be efficiently and precisely represented using a very small number of parameters. Additionally, these parameters are obtained by relatively simple calculation.

Let us express the discrete speech signal by $\{x_t\}$ (t is an integer), and assume the following first-order linear combination between the present sample value x_t and the previous p samples:

$$x_t + \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} = \epsilon_t \tag{16}$$

where $\{\epsilon_t\}$ is an uncorrelated statistical variable having a mean value of 0 and a variance of σ^2.

This linear difference equation means that the present sample value x_t can be linearly predicted using the previous sample values. That is, if the linearly predicted value $\hat{x}_t$ for x_t is represented by

$$\hat{x}_t = -\sum_{i=1}^{p} \alpha_i x_{t-i} \tag{17}$$

the following equation can be obtained from Eqs. (16) and (17):

$$x_t - \hat{x}_t = \epsilon_t \tag{18}$$

We thus consider Eq. (16) to be the linear prediction model having linear predictor coefficients $\{\alpha_i\}$. ϵ_t is designated as the residual error.

Let us now define the linear predictor filter as

$$F(z) = -\sum_{i=1}^{p} \alpha_i z^{-1} \tag{19}$$

and define $\hat{X}(z) \leftrightarrow \hat{x}_t$ and $X(z) \leftrightarrow x_t$ as the pairs of z-transforms and their sample values. The z-transform of Eq. (17) is then expressed by

$$\hat{X}(z) = F(z)X(z) \tag{20}$$

Based on Eqs. (17) and (18), the linear prediction model in z-transform notation can be given by

$$X(z)(1 - F(z)) = E(z) \tag{21}$$

or

$$X(z)A(z) = E(z) \tag{22}$$

where

$$A(z) = 1 + \sum_{i=1}^{p} \alpha_i z^{-1} = 1 - F(z) \tag{23}$$

and $E(z) \leftrightarrow \epsilon_t$. $A(z)$ is called the inverse filter (14). Based on these definitions, the linear predictive model using the linear predictor filter $F(z)$ and inverse filter $A(z)$ can be block diagrammed as in Fig. 7. This diagram shows that LPC—that is, the process of applying the linear predictive model to the speech wave—minimizes the output σ^2 by adjusting the coefficients $\{\alpha_i\}$ of either the linear predictor filter or the inverse filter.

LPC Analysis Procedure

Let us here consider the method for estimating the linear predictor coefficients $\{\alpha_i\}$ by applying the least mean square error method to Eq. (18). Specifically, let us determine the coefficients $\{\alpha_i\}_{i=1}^{p}$ so that the squared sum of the error ϵ_t between the sample values of x_t and the linearly predicted values $\hat{x}_t$ over a predetermined period of $[t_0, t_1]$ is minimized.

The total squared error β is

$$\begin{aligned} \beta &= \sum_{t=t_0}^{t_1} \epsilon_t^2 \\ &= \sum_{t=t_0}^{t_1} \left(\sum_{i=0}^{p} \alpha_i x_{t-i} \right)^2 \\ &= \sum_{t=t_0}^{t_1} \sum_{i=0}^{p} \sum_{j=0}^{p} \alpha_i \alpha_j x_{t-i} x_{t-j} \end{aligned} \tag{24}$$

where $\alpha_0 = 1$. Defining

$$c_{ij} = \sum_{t=t_0}^{t_1} x_{t-i} x_{t-j} \tag{25}$$

β can then be equivalently written as

$$\beta = \sum_{i=0}^{p} \sum_{j=0}^{p} \alpha_i c_{ij} \alpha_j \tag{26}$$

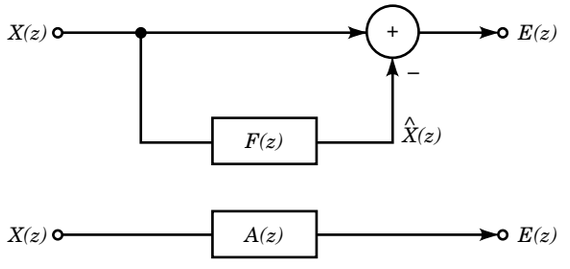

Figure 7. Linear prediction model block diagram. LPC minimizes the output $E(z)$ by adjusting the coefficients $\{\alpha_i\}_{i=1}^{p}$ of either the linear predictor filter $F(z)$ or the inverse filter $A(z)$.

Minimization of β is obtained by setting to zero the partial derivative of β with respect to α_j ($j = 1, 2, \ldots, p$) and solving. Therefore, from Eq. (26),

$$\frac{\partial \beta}{\partial \alpha_j} = 2 \sum_{i=0}^{p} \alpha_i c_{ij} = 0, \qquad j = 1, 2, \ldots, p \tag{27}$$

The predictor coefficients $\{\alpha_i\}$ can be obtained by solving this set of p linear simultaneous equations. The known parameters c_{ij} ($i = 0, 1, 2, \ldots, p; j = 1, 2, \ldots, p$) are defined from the sample data by Eq. (25), which shows that the samples x_t from $t_0 - p$ to t_1 are essential to the solution.

For the actual solution based on a sequence of N speech samples, $\{x_t\} = \{x_0, x_1, \ldots, x_{N-1}\}$, two specific cases have been investigated in detail. These are referred to as the *covariance method* and the *autocorrelation method.*

The covariance method is defined by setting $t_0 = p$ and $t_1 = N - 1$ so that the error is minimized only over the interval $[p, N - 1]$, whereas all the N speech samples are used in calculating the covariance matrix elements c_{ij} (15). Accordingly, Eq. (27) is solved using

$$c_{ij} = \sum_{t=p}^{N-1} x_{t-i} x_{t-j} \tag{28}$$

The covariance method draws its name from the fact that c_{ij} represents the row i, column j element of a covariance matrix.

The autocorrelation method is defined by setting $t_0 = -\infty$ and $t_1 = \infty$ and by letting $x_t = 0$ for $t < 0$ and $t \geq N$ (13). These limits allow c_{ij} to be simplified as

$$\begin{aligned} c_{ij} &= \sum_{t=-\infty}^{\infty} x_{t-i} x_{t-j} \\ &= \sum_{t=-\infty}^{\infty} x_t x_{t+|i-j|} \\ &= \sum_{t=0}^{N-1-|i-j|} x_t x_{t+|i-j|} \\ &= r_{|i-j|} \end{aligned} \tag{29}$$

Thus, α_i is obtained by solving

$$\sum_{i=0}^{p} \alpha_i r_{|i-j|} = 0, \qquad j = 1, 2, \ldots, p \tag{30}$$

where

$$r_\tau = \sum_{t=0}^{N-1-\tau} x_t x_{t+\tau} \qquad \tau \geq 0 \tag{31}$$

Although the error ϵ_t is minimized over an infinite interval, equivalent results are obtained by minimizing it only over $[0, N + p - 1]$. This is because x_t is truncated to zero for $t < 0$ and $t \geq N$ by multiplying by a finite-length window, such as a Hamming window. The autocorrelation method is so named from the fact that for the conditions stated, c_{ij} reduces to the definition of the short-term autocorrelation r_τ at the delay $\tau = |i - j|$.

Equation (31) can be expressed by matrix representation as

$$\begin{bmatrix} r_0 & r_1 \cdots & r_{p-1} \\ r_1 & r_0 & \cdot \\ \cdot & & \cdot \\ \cdot & \ddots & \cdot \\ \cdot & & r_1 \\ r_{p-1} & \cdots r_1 & r_0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_p \end{bmatrix} = - \begin{bmatrix} r_1 \\ r_2 \\ \cdot \\ \cdot \\ \cdot \\ r_p \end{bmatrix} \tag{32}$$

The $p \times p$ correlation matrix of the left term has the form of a Toeplitz matrix, which is symmetrical, and has the same values along the lines parallel to the diagonal. This type of equation is called a *normal equation* or a *Yule–Walker equation.* Since the positive definiteness of the correlation matrix is guaranteed by the definition of the correlation function, an inverse matrix exists for the correlation matrix. Solving the equation then permits $\{\alpha_i\}$ to be obtained. On the other hand, the positive definiteness of the coefficient matrix is not necessarily guaranteed in the covariance method.

The equations for the covariance and correlation methods can be efficiently solved by the Cholesky decomposition method and by Levinson–Durbin's recursive solution methods, respectively. Levinson–Durbin's method is equivalent to the PARCOR (partial autocorrelation) coefficient extraction process which will be presented in the section entitled "PARCOR Analysis." Although the covariance and autocorrelation methods give almost the same results when $\{x_t\}$ is long ($N \gg 1$) and stationary, their results differ when $\{x_t\}$ is short and features temporal variations.

In linear system identification in modern control theory, the process exemplified by Eq. (16) is called the *autoregressive* (AR) process, in which ϵ_t and x_t are the system input and output, respectively. This system is also referred to as the *all-pole model* since it has an all-pole system function.

Maximum Likelihood Spectral Estimation

Formulation of Maximum Likelihood Spectral Estimation. Maximum likelihood estimation is the method used to estimate parameters which maximize the likelihood based on the observed values. Here, the likelihood is the probability of occurrence of the actual observations (the speech samples) under the presumed parameter condition. The maximum likelihood method is better than any other estimation method in the sense that the variance of the estimated value is minimized when the sample size is sufficiently large.

In order to accomplish maximum likelihood spectral estimation, let us make two assumptions for the speech wave (11):

1. The sample value x_t can be regarded as the sample derived from a stationary Gaussian process characterized by the power spectral density $S(\lambda)$.
2. The spectral density $S(\lambda)$ is represented by an all-pole polynomial spectral density function of the form

$$\begin{aligned} S(\lambda) &= \frac{\sigma^2}{2\pi} \frac{1}{\left| \prod_{i=1}^{p} (1 - z/z_i) \right|^2} \\ &= \frac{\sigma^2}{2\pi} \frac{1}{\left| 1 + \sum_{i=1}^{p} \alpha_i z^{-i} \right|^2} \\ &= \frac{\sigma^2}{2\pi} \frac{1}{A_0 + 2 \sum_{i=1}^{p} A_i \cos i\lambda} \end{aligned} \tag{33}$$

where z_i is the root of

$$1 + \sum_{i=1}^{p} \alpha_i z^{-i} = 0 \tag{34}$$

and A_i is defined as

$$A_i = \sum_{j=0}^{p-|i|} \alpha_j \alpha_{j+|i|'}, \quad \alpha_0 = 1 \qquad i = 0, \pm 1, \ldots, \pm p \tag{35}$$

Furthermore, σ^2 is the scaling factor for the magnitude of spectral density, and p is the number of poles necessary for approximating the actual spectral density. Here, a pair of conjugate poles is counted as two separate poles.

Assumption 2 corresponds to the AR process described in the previous section. That is, the signal $\{x_t\}$, exhibiting the spectral density of Eq. (33), satisfies the relationship of Eq. (16) in the time domain. This correspondence can be understood if one traces back from Eq. (23) to Eq. (16).

Zeros are not included in the hypothesized spectral density for two reasons. First, the human auditory organs are sensitive to poles and insensitive to steep spectral valleys. Second, removing zeros simplifies as well as facilitates the mathematical process and the parameter extraction procedure.

$\{\alpha_i\}_{i=1}^{p}$ values obtained by maximum likelihood spectral estimation are actually equal to the values derived by the autocorrelation method. This means that linear predictive analysis employing the autocorrelation method and maximum likelihood spectral estimation, respectively, solve the same passive linear system (acoustic characteristics of the vocal tract, including the source and radiation characteristics) in the time domain and frequency domain, respectively. The maximum likelihood spectral estimation method is equivalent to the process of adjusting the coefficients to minimize the output power σ^2 when the input signal is passed through an adjustable pth-order inverse filter. Hence, this method is often referred to as the *inverse filtering method* (14).

Physical Meaning of Maximum Likelihood Spectral Estimation. In spectral matching using the maximum likelihood method, the matching error for neglecting a local valley in $\hat{S}(\lambda)$ is evaluated as being smaller than that for neglecting a local peak having the same shape. The nonuniform weighting in the maximum likelihood method is preferred over uniform weighting since the peaks play a dominant role in the perception of voiced speech.

The poles of the spectral envelope, z_i ($i = 1, 2, \ldots, p$), can be obtained as roots of the equation

$$1 + \sum_{i=1}^{p} \alpha_i z^{-i} = 0 \tag{36}$$

in which complex poles correspond to quadrature resonances. Their resonance frequencies and bandwidths are given by the equations

$$F_i = \frac{\arg z_i}{2\pi \Delta T} [\text{Hz}] \tag{37}$$

and

$$B_i = \frac{\log |z_i|}{\pi \Delta T}[\text{Hz}] \tag{38}$$

where ΔT is the sampling period. The formants can be extracted by selecting the poles whose bandwidth-to-frequency ratios are relatively small.

Source Parameter Estimation from Residual Signals

Let us consider the spectral fine structure of the residual signal

$$\epsilon_t = x_t - \hat{x}_t = \sum_{i=0}^{p} \alpha_i x_{t-i} \tag{39}$$

Since the fine structure is obtained by inverse-filtering the short-term spectrum of input speech, $\hat{S}(\lambda)$, using the spectral envelope $S(\lambda)$, it is almost flat along the frequency axis and exhibits a harmonic structure for periodic speech. Therefore, the autocorrelation for the residual signal, called the *modified autocorrelation function,* produces large correlation values at the delays having the integer ratio of the fundamental period for voiced speech, whereas no specific correlation is demonstrated for unvoiced speech (11).

In this way, the vocal source parameters can be obtained using the modified autocorrelation function regardless of the spectral envelope shape. The modified autocorrelation function can be easily calculated by the Fourier expansion of $\hat{S}(\lambda)/S(\lambda)$ as follows:

$$\begin{aligned} \gamma_\tau &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\hat{S}(\lambda)}{S(\lambda)} \cos \tau\lambda \, d\lambda \\ &= \frac{1}{\sigma^2} \int_{-\pi}^{\pi} \hat{S}(\lambda) \sum_{s=-p}^{p} A_s \cos(\tau - s)\lambda \, d\lambda \\ &= \frac{1}{N\sigma^2} \sum_{s=-p}^{p} A_s r_{\tau - s} \end{aligned} \tag{40}$$

where A_s is a correlation function of linear predictor coefficients as previously defined by Eq. (35). Equation (40) means that γ_τ can be calculated by the convolution of the short-term autocorrelation function and $\{A_s\}_{s=1}^{p}$ for input speech, followed by normalization using $N\sigma^2$. γ_τ can also be obtained by directly calculating the correlation function for ϵ_t using Eq. (39).

In the course of pitch extraction, low-pass filtering is widely applied to speech waves or residual signals for improving the resolution of the extracted pitch period. Low-pass filtering is effective for removing the influence of high-order formants and for compensating for the insufficiency of the time resolution arising in the autocorrelation function. The latter effect is especially important for pitch extraction using this modified autocorrelation function. The double-period pitch error due to the time resolution insufficiency can be considerably minimized by employing low-pass filtering.

PARCOR Analysis

Formulation of PARCOR Analysis. The same two assumptions made for the maximum likelihood estimation (see section entitled "Formulation of Maximum Likelihood Spectral Estimation") are also made for the speech wave. When the prediction errors for the linear prediction of x_t and x_{t-m} using the sampled values $\{x_{t-i}\}_{i=1}^{m-1}$ are written as

$$\epsilon_{ft}^{(m-1)} = \sum_{i=0}^{m-1} \alpha_i^{(m-1)} x_{t-i} \tag{41}$$

and

$$\epsilon_{bt}^{(m-1)} = \sum_{i=1}^{m} \beta_i^{(m-1)} x_{t-i} \tag{42}$$

the PARCOR (partial autocorrelation) coefficient k_m between x_t and x_{t-m} is defined by

$$k_m = \frac{E\{\epsilon_{ft}^{(m-1)} \epsilon_{bt}^{(m-1)}\}}{[E(\epsilon_{ft}^{(m-1)})^2\} E\{(\epsilon_{bt}^{(m-1)})^2\}]^{1/2}} \tag{43}$$

This equation means that the PARCOR coefficient is the correlation between the forward prediction error $\epsilon_{ft}^{(m-1)}$ and the backward prediction error $\epsilon_{bt}^{(m-1)}$ (11). The definitional concept behind the PARCOR coefficient is presented in block diagram form in Fig. 8. Since the prediction errors, $\epsilon_{ft}^{(m-1)}$ and $\epsilon_{bt}^{(m-1)}$, are obtained after removing the linear effect of m sample values between x_t and x_{t-m} from these sample values, k_m represents the pure or partial correlation between x_t and x_{t-m}.

When Eqs. (41) and (42) are put into Eq. (43), the PARCOR coefficient sequence k_m ($m = 1, 2, \ldots, p$) can be written as

$$\begin{aligned} k_m &= \frac{\sum_{i=0}^{m-1} \alpha_i^{(m-1)} r_{m-i}}{\sum_{i=0}^{m-1} \alpha_i^{(m-1)} r_i} \\ &= \frac{w_{m-1}}{u_{m-1}} \end{aligned} \tag{44}$$

where r_i is the short-term autocorrelation function for the speech wave. k_1 is equal to r_1—that is, to the first-order autocorrelation coefficient. This is also clear from the definition of k_m.

Using Eq. (44) and the fact that the prediction coefficients $\{\alpha_i^{(m-1)}\}_{i=0}^{m-1}$ and $\{\beta_i^{(m-1)}\}_{i=1}^{m}$ constitute the solutions of the simultaneous equations

$$\sum_{i=0}^{m-1} \alpha_i^{(m-1)} r_{i-j} = 0, \quad \alpha_0^{(m-1)} = 1 \quad (j = 1, 2, \ldots, m-1) \tag{45}$$

and

$$\sum_{i=1}^{m} \beta_i^{(m-1)} r_{i-j} = 0, \quad \beta_m^{(m-1)} = 1 \quad (j = 1, 2, \ldots, m-1) \tag{46}$$

the following recursive equations can be obtained ($m = 1, 2, \ldots, p$):

$$\begin{aligned} &\alpha_i^{(m)} = \alpha_i^{(m-1)} - k_m \beta_i^{(m-1)}, \quad \alpha_m^{(m-1)} = 0, \\ &\qquad\qquad\qquad\qquad\qquad\qquad i = 1, 2, \ldots, m \\ &\beta_i^{(m)} = \beta_{i-1}^{(m-1)} - k_m \alpha_{i-1}^{(m-1)}, \quad \beta_0^{(m-1)} = 0 \\ &u_m = u_{m-1}(1 - k_m^2) \qquad\qquad (i = 1, 2, \ldots, m+1) \end{aligned} \tag{47}$$

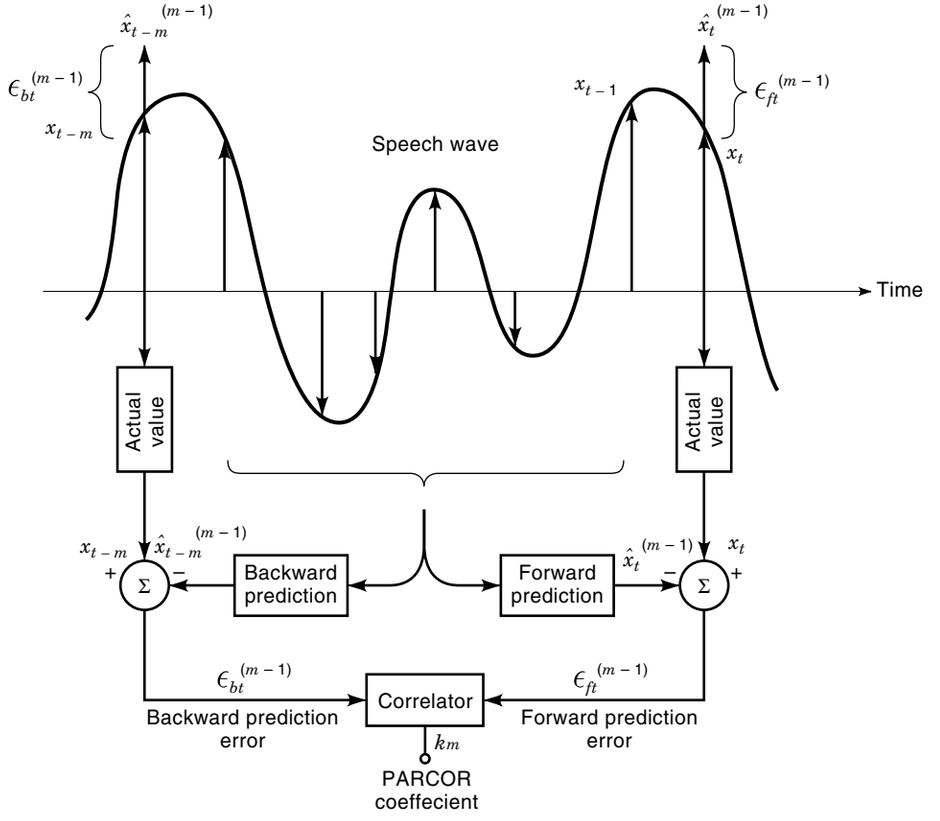

Figure 8. Definition of PARCOR coefficients. The PARCOR coefficient is the correlation between the forward prediction error $\epsilon_{ft}^{(m-1)}$ and the backward prediction error $\epsilon_{bt}^{(m-1)}$.

Additionally, the following equation is obtained from Eqs. (45) and (46):

$$\beta_i^{(m-1)} = \alpha_{m-i}^{(m-1)}, \qquad i = 1, 2, \ldots, m-1 \tag{48}$$

Based on these results, the PARCOR coefficients $\{k_m\}_{m=1}^p$ and linear predictor coefficients $\{\alpha_m\}_{m=1}^p$ are obtained from $\{r_i\}_{i=1}^p$ through the flowchart given in Fig. 9 by using Eqs. (44) and (47). This iterative method is equivalent to Levison–Durbin's recursive solution for simultaneous linear equations. The numbers of multiplications, summations, and divisions necessary for this computation are roughly $p(p+1)$, $p(p+1)$, and p, respectively. When these computations are done using a short word length, the truncation error in the computation accumulates as the analysis progresses. In the iteration process, each k_m ($m = 1, 2, \ldots, p$) is obtained one by one, whereas the α_m values change at every iteration. Finally, α_m values are obtained as

$$\alpha_m = \alpha_m^{(p)}, \qquad 1 \le m \le p \tag{49}$$

Since the normalized mean square error σ^2 is equal to u_p from its definition, σ^2 can be calculated using PARCOR coefficients, instead of linear predictor coefficients, from

$$\sigma^2 = \prod_{m=1}^{p} (1 - k_m^2) \tag{50}$$

This equation is obtained from Eq. (47).

In order to derive $\{k_m\}_{m=1}^p$ directly from the signal $\{x_t\}$ let us define the forward and backward prediction error operators $A_m(D)$ and $B_m(D)$ as

$$A_m(D) = \sum_{i=0}^{m} \alpha_i^{(m)} D^i \tag{51}$$

and

$$B_m(D) = \sum_{i=1}^{m+1} \beta_i^{(m)} D^i \tag{52}$$

where D is the delay operator such that $D^i x_t = x_{t-i}$. Equations (41) and (42) can then be written as

$$\epsilon_{ft}^{(m-1)} = A_{m-1}(D) x_t \tag{53}$$

and

$$\epsilon_{bt}^{(m-1)} = B_{m-1}(D) x_t \tag{54}$$

From Eq. (47), we can arrive at the recursive equations

$$A_m(D) = A_{m-1}(D) - k_m B_{m-1}(D) \tag{55}$$

and

$$B_m(D) = D(B_{m-1}(D) - k_m A_{m-1}(D)) \tag{56}$$

Based on Eqs. (44), (53), (54), (55), and (56), the PARCOR coefficients $\{k_m\}$ can subsequently be produced directly from the speech wave x_t using a cascade connection of variable pa-

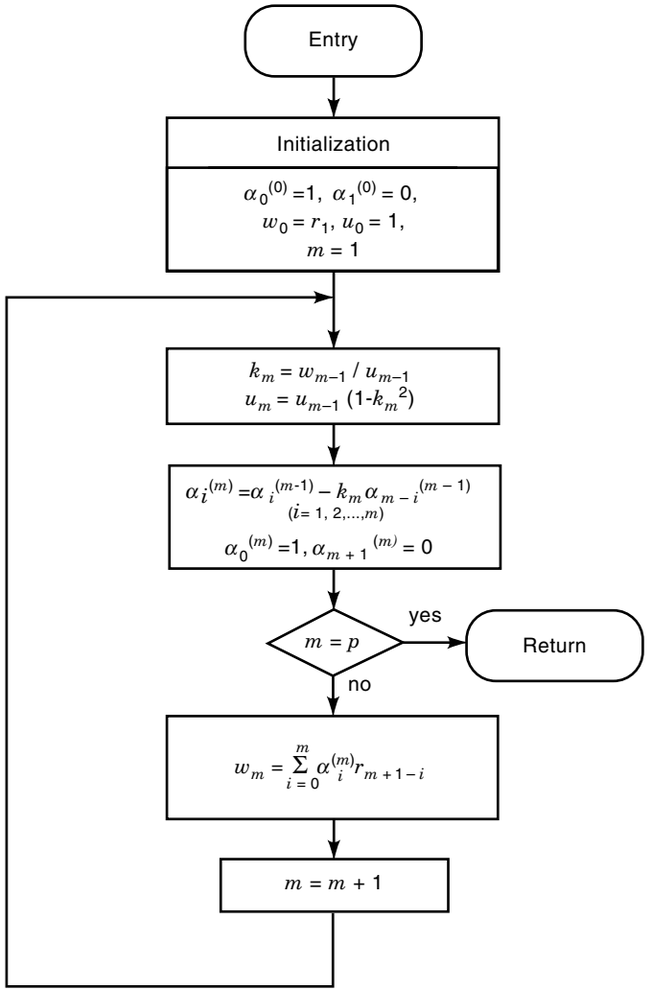

Figure 9. Flowchart for calculating $\{k_m\}_{m=1}^{p}$ and $\{\alpha_m\}_{m=1}^{p}$ from $\{r_i\}_{i=1}^{p}$. This iterative method is equivalent to Levinson-Durbin's recursive solution of simultaneous linear equations.

rameter digital filters (partial correlators), each of which includes a correlator as indicated in Fig. 10(a). Since $E\{(\epsilon_{ft}^{(m-1)})^2\} = E\{(\epsilon_{bt}^{(m-1)})^2\}$, the correlator can be realized by the structure indicated in Fig. 10(b), which consists of square, addition, subtraction, and division circuits and low-pass filters.

The process of extracting PARCOR coefficients using the partial correlators involves successively extracting and removing the correlations between adjacent samples. This is an inverse filtering process which flattens the spectral envelope successively. Therefore, when the number of partial correlators p is large enough, the correlation between adjacent samples, which corresponds to the overall spectral envelope information, is almost completely removed by passing the speech wave through the partial correlators. Consequently, the output of the final stage—namely, the residual signal—includes only the correlation between the distant samples which relates to the source (pitch) information. Hence, the source parameters can be extracted from the autocorrelation function for the residual signal—in other words, from the modified autocorrelation function.

Relationship Between PARCOR and LPC Coefficients. If either one of the set of $\{k_m\}_{m=1}^{p}$ or $\{\alpha_m\}_{m=1}^{p}$ is given, the other can be obtained by recursive computations. For example, when $\{k_m\}_{m-1}^{p}$ are given, $\{\alpha_m\}_{m=1}^{p}$ are derived by recursive computations ($m = 1, 2, \ldots, p$) using a part of Levinson–Durbin's solution:

$$\begin{aligned} \alpha_m^{(m)} &= -k_m \\ \alpha_i^{(m)} &= \alpha_i^{(m-1)} - k_m\alpha_{m-i}^{(m-1)}, \qquad 1 \le i \le m-1 \end{aligned} \tag{57}$$

On the other hand, $\{k_m\}_{m=1}^{p}$ can be drawn from $\{\alpha_m\}_{m=1}^{p}$ using the recursive computations in the opposite direction ($m = p, p-1, \ldots, 2, 1$) as indicated below, where the initial condition is $\alpha_m^{(p)} = \alpha_m$ ($1 \le m \le p$):

$$\begin{aligned} k_m &= -\alpha_m^{(m)} \\ \alpha_i^{(m-1)} &= \frac{\alpha_i^{(m)} - \alpha_m^{(m)}\alpha_{m-i}^{(m)}}{1-k_m^2}, \qquad 1 \le i \le m-1 \end{aligned} \tag{58}$$

Line Spectrum Pair (LSP) Analysis

As with PARCOR analysis, LSP analysis is based on the all-pole model. The PARCOR coefficients are essentially parameters operating in the time domain as are the autocorrelation coefficients, whereas the LSPs are parameters functioning in the frequency domain. Therefore, the LSP parameters are advantageous in that the distortion they produce is smaller than that of the PARCOR coefficients even when they are roughly quantized and linearly interpolated.

The polynomial expression for z, which is the denominator of the all-pole model, satisfies the following recursive equations, as previously demonstrated in Eqs. (55) and (56):

$$A_m(z) = A_{m-1}(z) - k_m B_{m-1}(z) \tag{59}$$

and

$$B_m(z) = z^{-1}(B_{m-1}(z) - k_m A_{m-1}(z)) \tag{60}$$

where $A_0(z) = 1$ and $B_0(z) = z^{-1}$ (initial conditions).

Let us assume that $A_p(z)$ is given; and represent two $A_{p+1}(z)$ types, $P(z)$ and $Q(z)$, under the conditions $k_{p+1} = 1$ and $k_{p+1} = -1$, respectively. Then a pair of delta function-like resonance characteristics (a pair of line spectra) which correspond to each boundary condition at the glottis are obtained. The number of resonances are $2p$.

From Eqs. (59) and (60), $P(z)$ and $Q(z)$ can be represented as

$$P(z) = A_p(z) - B_p(z) \tag{61}$$

and

$$Q(z) = A_p(z) + B_p(z) \tag{62}$$

Although $P(z)$ and $Q(z)$ are both $(p + 1)$st-order polynomial expressions, $P(z)$ has inversely symmetrical coefficients whereas $Q(z)$ has symmetrical coefficients. Using Eqs. (59) through (62), we get

$$A_p(z) = \frac{P(z) + Q(z)}{2} \tag{63}$$

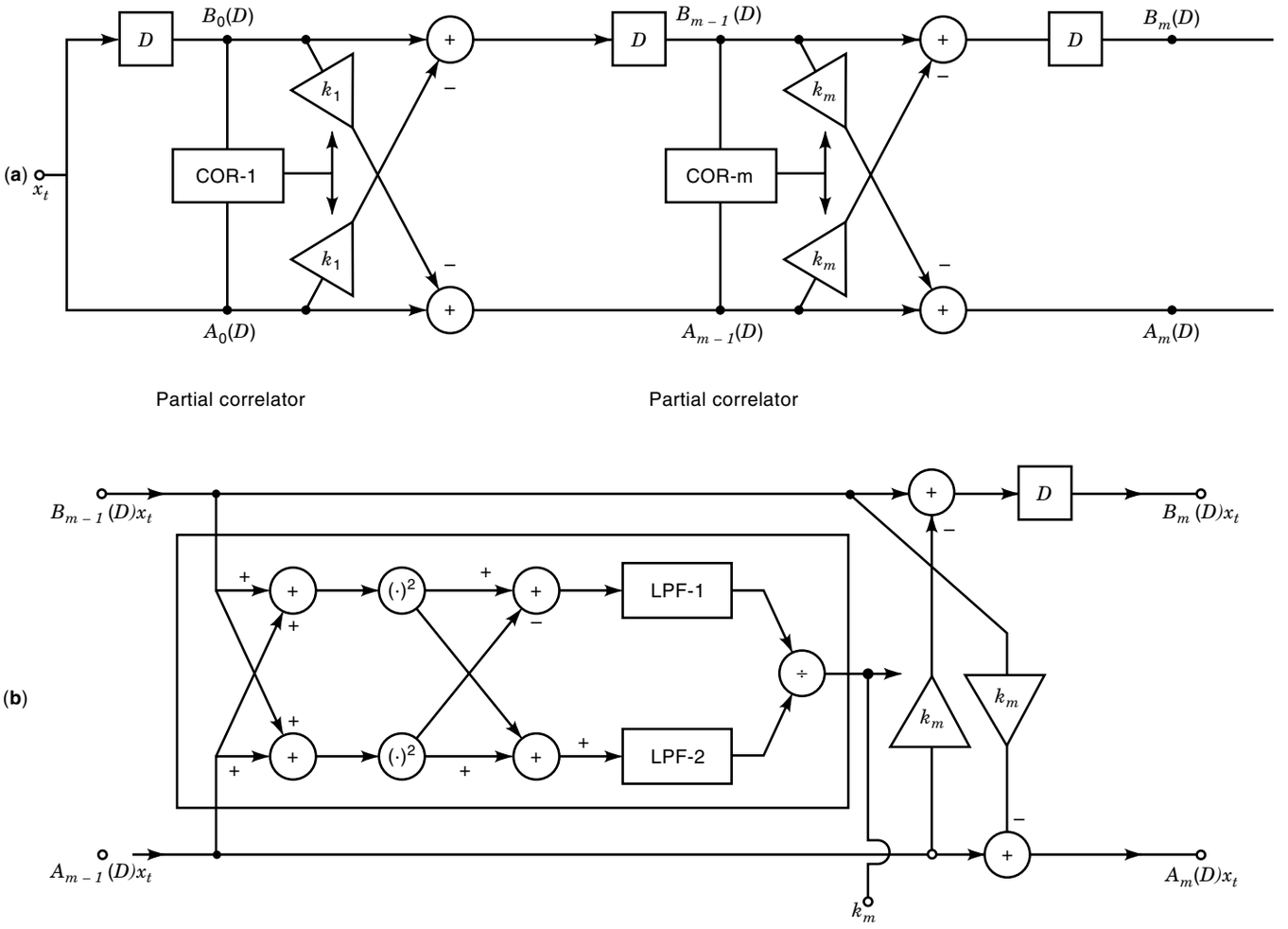

Figure 10. (a) PARCOR coefficient extraction circuit constructed by cascade connection of partial autocorrelators and (b) structure of each partial autocorrelator. The PARCOR coefficients $\{k_m\}_{m=1}^{p}$ can be produced directly from the speech wave $\{x_t\}$ using the cascade connection of variable parameter digital filters (partial autocorrelators).

and

$$\begin{aligned} B_p(z) &= z^{-(p+1)}A_p(z^{-1}) \\ &= z^{-(p+1)} + z^{-p}\alpha_1 + \cdots + z^{-1}\alpha_p \end{aligned} \tag{64}$$

If p is assumed to be even, $P(z)$ and $Q(z)$ are factorized as

$$P(z) = (1 - z^{-1}) \prod_{i=2,4,\ldots,p} (1 - 2z^{-1}\cos\omega_i + z^{-2}) \tag{65}$$

and

$$Q(z) = (1 + z^{-1}) \prod_{i=1,3,\ldots,p-1} (1 - 2z^{-1}\cos\omega_i + z^{-2}) \tag{66}$$

The factors $1 - z^{-1}$ and $1 + z^{-1}$ are found by calculating $P(1)$ and $Q(-1)$ after putting Eq. (64) into Eqs. (65) and (66). The coefficients $\{\omega_i\}$ which appear in the factorization of Eqs. (65) and (66) are referred to as LSP parameters. $\{\omega_i\}$ are ordered as

$$0 < \omega_1 < \omega_2 < \cdots < \omega_{p-1} < \omega_p < \pi \tag{67}$$

Even-suffixed $\{\omega_i\}$ are proved to separate each element of odd-suffixed $\{\omega_i\}$, and vice versa. In other words, even-suffixed $\{\omega_i\}$ and odd-suffixed $\{\omega_i\}$ are interlaced. Under the condition that p is odd, the LSP is obtained in the same way.

Using Eq. (63), the power transmission function for $H(z)$ can be represented as

$$\begin{aligned} |H(e^{-j\omega})|^2 &= \frac{1}{|A_p(e^{-j\omega})|^2} \\ &= 4|P(e^{-j\omega}) + Q(e^{-j\omega})|^{-2} \\ &= 2^{1-p}\left\{\sin^2\frac{\omega}{2}\prod_{i=2,4,\ldots,p}(\cos\omega - \cos\omega_i)^2\right. \\ &\quad \left. + \cos^2\frac{\omega}{2}\prod_{i=1,3,\ldots,p-1}(\cos\omega - \cos\omega_i)^2\right\}^{-2} \end{aligned} \tag{68}$$

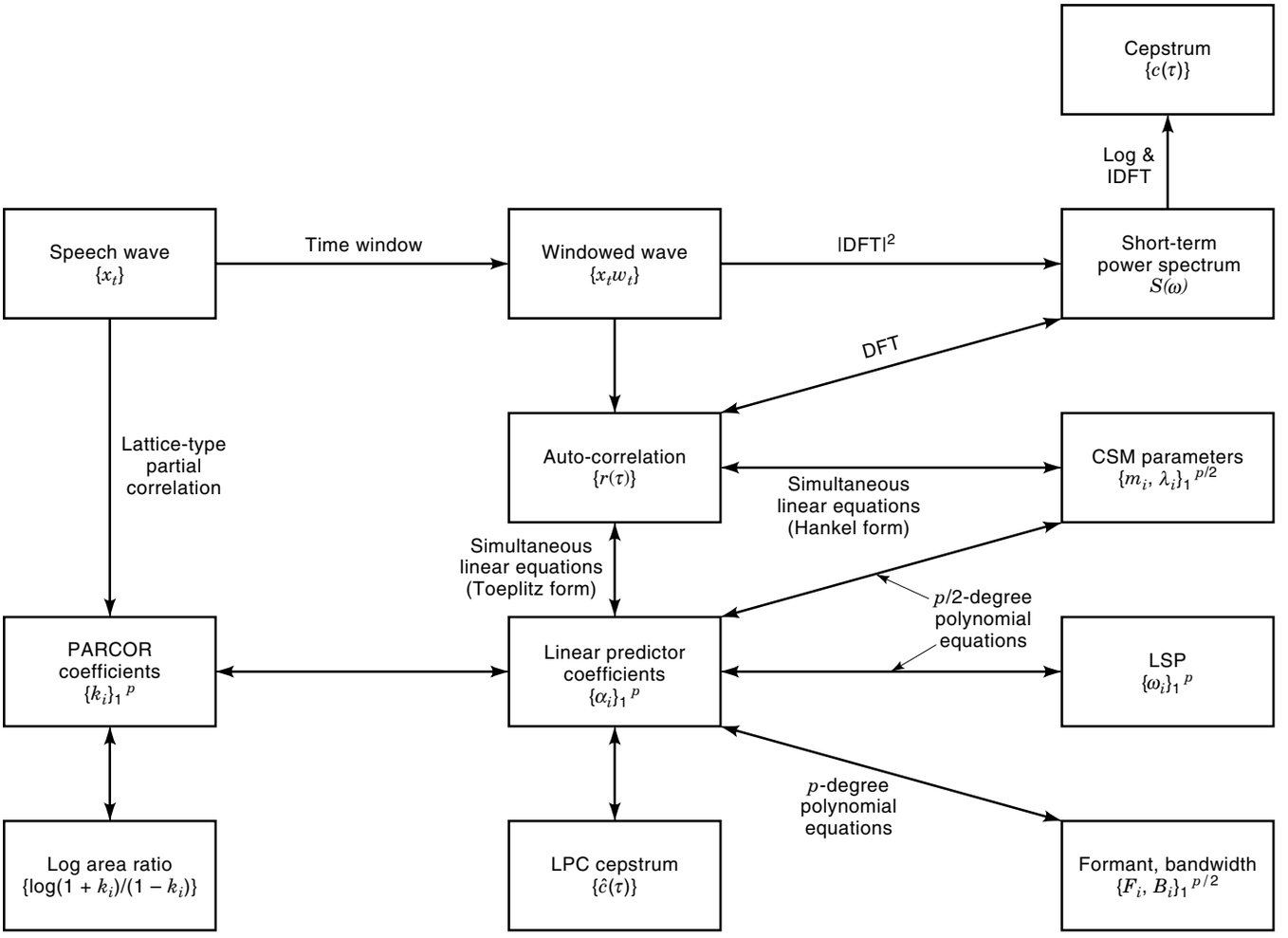

Figure 11. Mutual relationships between parameters based on all-pole spectrum modeling (LPC modeling).

The first term in braces approaches 0 when ω approaches 0 or one of the $\{\omega_i\}$ ($i = 2, 4, \ldots, p$), and the second term approaches 0 when ω approaches π or one of the $\{\omega_i\}$ ($i = 1, 3, \ldots, p - 1$). Therefore, when two LSP parameters, ω_i and ω_j, are close together and when ω approaches both of them, the gain of $H(z)$ becomes large and resonance occurs. Strong resonance occurs at frequency ω when two or more ω_i's are concentrated near ω. That is, the LSP method represents the speech spectral envelope through a distribution density of p discrete frequencies $\{\omega_i\}$.

Either of the following methods can be used to obtain the zeroes for $P(z)$ and $Q(z)$ with respect to z^{-1} after deriving the coefficients for $A_p(z)$—that is, the linear predictor coefficients $\{\alpha_i\}$.

1. Root finding in algebraic quations. Equations (65) and (66) can be transformed into

$$\prod_{j=1}^{m}(1-2z^{-1}\cos\omega^j+z^{-2}) = (2z^{-1})^m \prod_{j=1}^{m}\left(\frac{z+z^{-1}}{2}-\cos\omega_j\right) \tag{69}$$

Then, by replacing $(z + z^{-1})/2|_{z=\exp(-j\omega)} = \cos\omega$ with x, the equations $P(z)/(1 - z^{-1}) = 0$ can be solved as a pair of $(p/2)$th-order algebraic equations with respect to x using the Newton iteration method.

2. DFT for the coefficients of the equations. The values of $P(z)$ and $Q(z)$ at $z_n = e^{-jn\pi/N}$ ($n = 0, \ldots, N$) are first obtained through the DFT using the coefficients of $P(z)$ and $Q(z)$. Zeros can then be estimated by the interpolation of two points which produce a zero between them. The procedure for searching for the zeros is largely reduced using the relationship $0 < \omega_1 < \omega_2 < \cdots < \omega_p < \pi$. A value between 64 and 128 is considered large enough for N.

Mutual Relationships between LPC Parameters

The mutual relationships between each parameter obtained based on all-pole spectral modeling (LPC modeling) are indicated in Fig. 11 (16). For the parameters such as log-area ratios and CSM parameters, please refer to Furui (17).

The relationship existing between the autocorrelation function for the impulse response of the all-pole system $\tilde{r}_\tau$ and

$\{\alpha_i\}_{i=0}^{p}$ can be expressed as

$$\sum_{i=0}^{p} \alpha_i \tilde{r}_{|i-j|} = 0, \qquad j \geq 1 \tag{70}$$

$\tilde{r}_\tau$, which is often called the LPC correlation function, agrees with the autocorrelation function for the signal r_τ in the range of $\tau = 1$ to p.

BIBLIOGRAPHY

1. B. P. Bogert, M. J. R. Healy, and J. W. Turkey, The frequency analysis of time-series for echoes, *Proc. Symp. Time Series Analysis,* 1963, Chapter 15, pp. 209–243.
2. A. M. Noll, Short-time spectrum and 'cepstrum' techniques for vocal-pitch detection, *J. Acoust. Soc. Am.,* **36** (2): 296–302, 1964.
3. A. M. Noll, Cepstrum pitch determination, *J. Acoust. Soc. Am.,* **41** (2): 293–309, 1967.
4. B. S. Atal, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *J. Acoust. Soc. Am.,* **55** (6): 1304–1312, 1974.
5. S. Furui, Cepstral analysis technique for automatic speaker verification, *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-29** (2): 254–272, 1981.
6. C. G. Bell et al., House, Reduction of speech spectra by analysis-by-synthesis techniques, *J. Acoust. Soc. Am.,* **33** (12): 1725–1736, 1961.
7. F. Itakura and Y. Tohkura, Feature extraction of speech signal and its application to data compression, *Joho-shori,* **19** (7): 644–656, 1978.
8. J. D. Markel, The SIFT algorithm for fundamental frequency estimation, *IEEE Trans. Audio. Electroacoust.,* **AU-20** (5): 367–377.
9. B. S. Atal and L. R. Rabiner, A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition, *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-24** (3): 201–212, 1976.
10. N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series,* Cambridge, MA: MIT Press, 1966.
11. F. Itakura and S. Saito, Analysis synthesis telephony based on the maximum likelihood method, *Proc. 6th Int. Congress on Acoustics,* 1968, pp. C-5-5.
12. B. S. Atal and M. R. Schroeder, Predictive coding of speech signals, *Proc. 6th Int. Congress on Acoustics,* 1968, pp. C-5-4.
13. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech,* New York: Springer-Verlag, 1976.
14. J. D. Markel, Digital inverse filtering—A new tool for formant trajectory estimation, *IEEE Trans. Audio Electroacoust.,* **AU-20** (2): 129–137, 1972.
15. B. S. Atal and S. L. Hanauer, Speech analysis and synthesis by linear prediction of the speech wave, *J. Acoust. Soc. Am.,* **50** (2) (Part 2): 637–655, 1971.
16. F. Itakura, Speech analysis-synthesis based on spectrum encoding, *J. Acoust. Soc. Jpn.,* **37** (5): 197–203, 1981.
17. S. Furui, *Digital Speech Processing, Synthesis, and Recognition,* New York: Marcel Dekker, 1989.
18. F. Itakura and S. Saito, Digital filter techniques for speech analysis and synthesis, *Proc. 7th Int. Congress on Acoustics,* Budapest, 1971, pp. 25-C-1.

SADAOKI FURUI
Tokyo Institute of Technology