Formant Estimation Method Using Inverse-Filter Control

Akira Watanabe

***Abstract*—This paper proposes a new method for estimating formant frequencies of speech signals, based on inverse-filter control and zero-crossing frequency distributions. In this method, which is called the inverse-filter control (IFC) method, we use 32 basic inverse filters that are mutually controlled by weighted means of zero-crossing frequency distributions. After quick convergence of the inverse filters, we can gain four to six formant frequencies as final mean-values of the zero-crossing frequencies. The proposed method (IFC) has a specific feature that it directly estimates resonant frequencies of a vocal tract, unlike analysis-by-synthesis (A-b-S) or linear predictive coding (LPC) as a spectral matching method. Therefore, spectral shapes influence indirectly alone the formant estimation in IFC. Although the superiority of IFC to LPC was not necessarily prominent in the systematic evaluation using synthetic speech, the estimates showed satisfactorily small errors for the practical analysis. On the other hand, when observing some analysis examples of real speech, we found many fewer gross errors in IFC than in LPC. Last, we describe in brief a method for estimating a spectral envelope (or formant bandwidths) based on the obtained formant frequencies and the spectrum to be analyzed. According to the results, it is understandable that existence of the wide-band formants also contributes to stable formant trajectories.**

***Index Terms*—Formant estimation, inverse-filter control, zero-crossing frequency.**

I. Introduction

FORMANT frequency is one of the most useful speech parameters to be specified by a vocal tract shape or its movements in various pronunciations. The formants are physically defined as poles in a system function expressing the characteristics of a vocal tract. Therefore, we can point out clearly their existence and properties using a typical model of acoustic tube approximating a vocal tract, such as the three parameter vocal tract model or line electric analog (LEA), etc. [1]. However, capturing and tracking formants accurately from natural speech is not so easy because of the variety of speech sounds. This paper proposes a new formant estimation method, called the inverse-filter control (IFC) method, in which 32 basic inverse filters are mutually controlled by zero-crossing frequencies.

Most speech researchers will agree that the two historically representative methods for estimating formant frequencies are the analysis-by-synthesis (A-b-S) method [2] and the linear predictive coding (LPC) method [3]. Their ideas are still brilliant and many modified methods have stemmed from them [4]–[6]. However, these methods are ultimately based on the best matching between a spectrum to be analyzed and a synthesized one so that formant frequencies are estimated through spectral shapes. Hence, the estimates may be sensitive to spectral distortions or modifications, because the analysis model seeks a set of parameters to represent correctly the spectrum by minimizing errors. On the other hand, the IFC method does not need any criterion to minimize errors for spectral estimation, unlike A-b-S or LPC, but estimates formant frequencies directly as average zero-crossing frequencies of the approximate single resonant waves, which have been separated from speech signals using inverse filters. Thus, spectral shapes influence indirectly alone the estimates of formant frequency in IFC. Based on the principle of IFC method, we devised a real-time hardware system for estimating the lowest three formants [7] and then applied it to the color display system of speech for the hearing impaired [8]. Although the obtained estimates were visually effective for the color representation of vowels, it was necessary to refine the algorithm for attaining higher accuracy in the formant estimation. However, since there had been no logical or no experimental paradigm for the IFC method except our research, we needed trial and error many times to complete the fine procedure applicable to comprehensive speech data. This paper describes the refined software system and evaluates accurateness and stability of the formant frequencies estimated by IFC method. In order to display the high performance in analyzing many varieties of speech, not only the principle but also all of the processing to be described in this paper should be implemented consistently.

Manuscript received December 21, 1999; revised September 12, 2000. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Philip C. Loizou.

The author is with the Department of Computer Science, Faculty of Engineering, Kumamoto University, Kumamoto 860-8555, Japan.

Publisher Item Identifier S 1063-6676(01)02735-3.

II. Principle and Number of Formants to be Estimated

A. Principle

To briefly describe the IFC method, we use an algorithm to separate speech signals into approximate single resonant waves by controlling inverse filters. Next, formant frequencies are estimated as a mean frequency from each of the separated waves. The principle is described as follows.

Each inverse filter to be used in the proposed system (IFC) has a pair of complex conjugate zeros necessary to approximately eliminate a pair of complex conjugate poles of a vocal tract transfer function. Thus, the system function of the inverse filter is shown by

$$\begin{aligned} H(z) &= \gamma(1 - z_1/z)(1 - z_1^*/z) \\ &= \gamma(1 - (z_1 + z_1^*)z^{-1} + (z_1 z_1^*)z^{-2}) \end{aligned} \tag{2.1}$$

where $z_1 = e^{-\pi B + j\omega}, z_1^* = e^{-\pi B - j\omega}$.

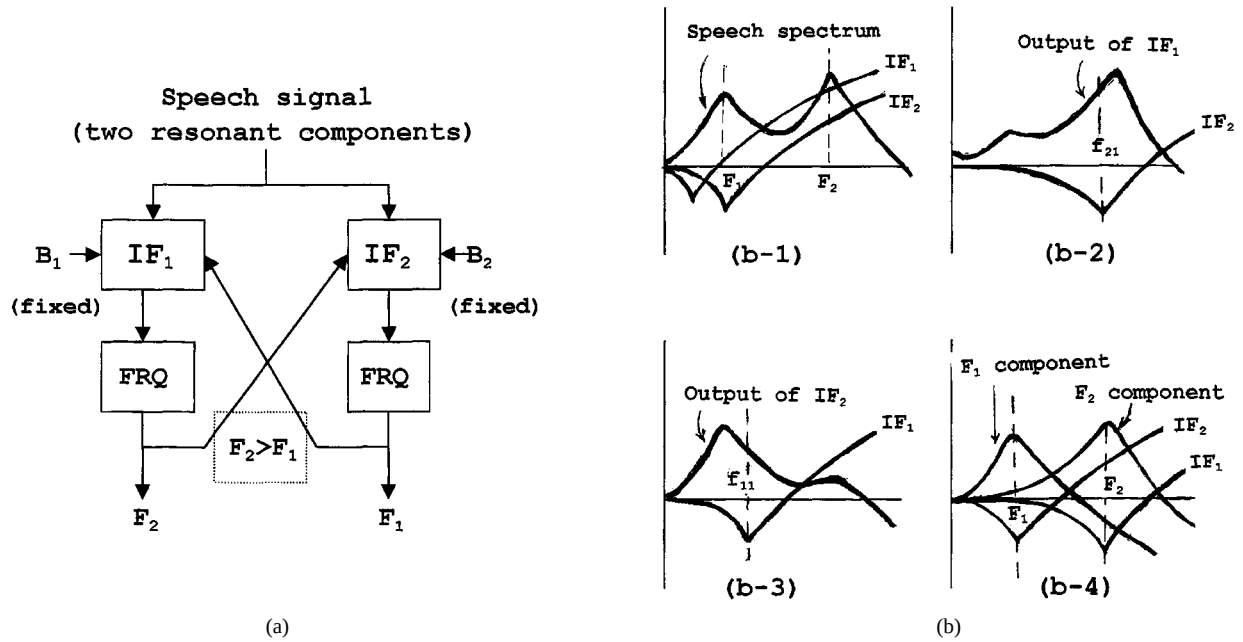


Fig. 1. Principle of formant estimation by the IFC method. (a) Simplest model and (b) control process.

Substituting z_1 and z_1^* into (2.1) and arranging it

$$H(z) = \gamma(1 + \beta z^{-1} + \alpha z^{-2}) \tag{2.2}$$

where

α = $e^{-2\pi B}$;
β = $-2e^{-\pi B}\cos\omega$;
γ = $1/(1+\alpha+\beta)$;

to be specified by a zero (notch) frequency, $\omega\ (= 2\pi F)$ and a bandwidth, B. We call the inverse filter of $H(z)$ "basic filter." Let us show the simplest model of IFC system to introduce the principle next. Fig. 1(a) and (b) show the simplest model and its control process, respectively, which are used for extracting the first and second formant frequencies (F_1, F_2) from speech signals containing the lowest two resonant components. The system in Fig. 1(a) is constructed with two basic filters (IF_1 and IF_2) and a routine (FRQ) to compute a mean frequency from the output signal of each basic filter. A variable notch with a fixed bandwidth characterizes the basic filter, which eliminates each of two resonant components. When the speech signals are put in the system shown in Fig. 1(a), the mean frequency f_{21} is computed by FRQ from output signals of IF_1, whose notch frequency is put at a low frequency (see b-1 in Fig. 1), and then, the notch frequency of IF_2 shifts to that frequency f_{21} (see b-2). Next, the mean frequency f_{11}, which is computed from output signals of IF_2, likewise controls the notch frequency of IF_1 (see b-3) and then the mean frequency from IF_1-output is computed again, and so on. These two mean frequencies (f_{21} and f_{11}) are always compared each other so that the larger frequency controls IF_2 and the smaller one IF_1. By repeating the above mutual control between two basic filters a few times, the notch frequencies converge and two resonant components are individually extracted from the respective outputs of the basic filters, as shown in (b-4) of Fig. 1. Last, two mean frequencies of the separated components indicate two formant frequencies, F_1 and F_2, respectively. Since the approximate single resonant waves, which are separated from speech signals, typically form the damped sinusoids synchronizing with an excitation signal, both the frequency of spectral peak and the zero-crossing frequency of each wave approximately indicate the resonant (formant) frequency. So, we use two steps for estimating the mean frequencies in the routine FRQ. One is a rough estimation from the power spectrum and the other is a more accurate estimation based on the zero-crossing frequency distribution. In speech signals including some resonance, the power spectrum forms some relatively clear peaks due to formants, but the zero-crossing frequency distribution shapes a dull hump alone. However, as the spectrum approaches a single resonant one, the zero-crossing frequency distribution is more rapidly getting local near the resonant frequency than the spectral peak, which is significantly influenced by harmonic components. Therefore, the center of gravity of spectrum is suitable for the first estimation of the mean frequency, and the zero-crossing frequency distribution should be used for the more accurate second estimation. In the second estimation, the formant frequency is approximated by the weighted mean of the zero crossing frequency distribution using a triangular weighting function whose base is reduced step-by-step in the control of the system as will be described in Section IV-B.

B. Number of Formants to be Estimated

When applying the above principle to the practical system, one of the problems is in determining the number of formants necessary to estimate in a given frequency band of speech signals. In general, it is well known that resonant (formant) frequencies of a uniform tube, which is a model of the vocal tract uttering a neutral vowel, are given by the following equation [9]:

$$F_n = (2n - 1)c/4L \tag{2.3}$$

where

F_n nth formant frequency [Hz];
c sound velocity [m/s];
L vocal tract length [m].

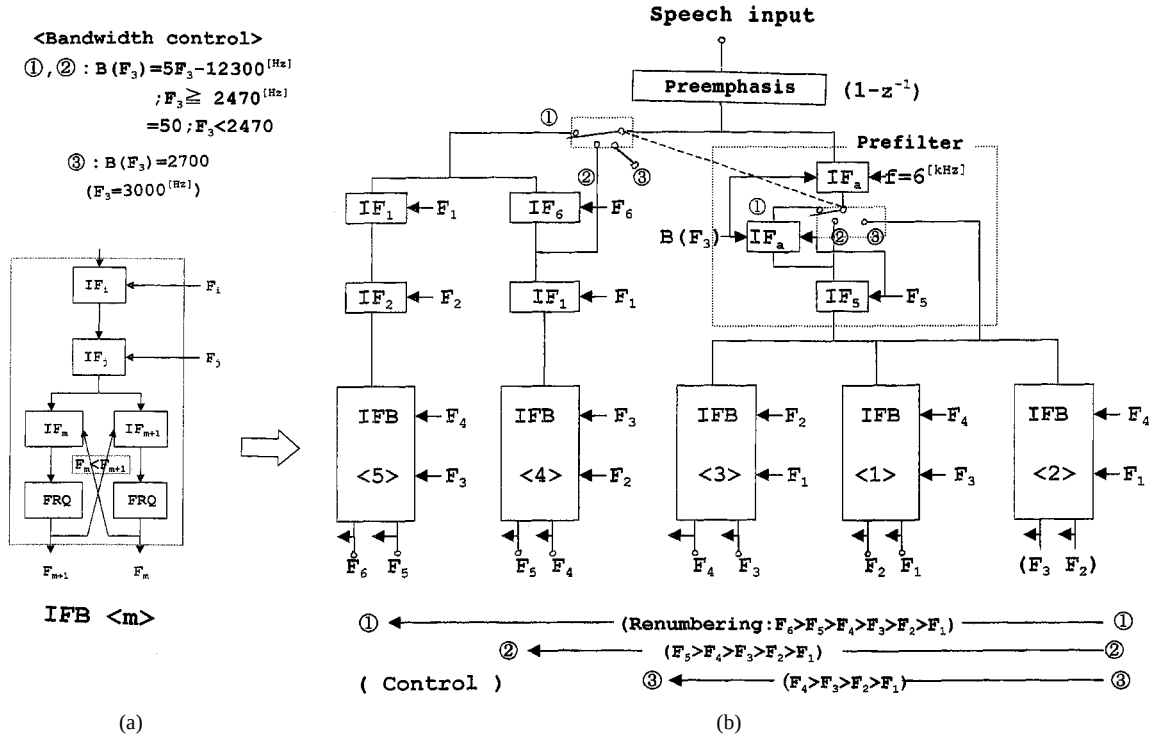


Fig. 2. Formant estimation system by the IFC method. (a) Fundamental block and (b) IFC system.

Therefore, if the average length of vocal tracts is 17.5 [cm] for adult males, the first formant appears at about 500 [Hz] and the upper formants are calculated as odd multiples of the first one. So, if we limit a frequency band of speech signals to 0 to 6 [kHz], six formants should be expected for voice of adult male speakers. In addition, according to the fact that the vocal tract lengths of adult females are approximately 15–26% shorter than the average length of adult males [10], the lowest five formants should be estimated in the same frequency band. Likewise, since 7–10 year old boys and girls are approximately 20–34% and 27–39% shorter [10], respectively, the lowest four or five formants need to be found for boys and the lowest four for girls. Thus, we have developed the system to seek any of three combinations in the formant frequencies, i.e., F_1-F_6, F_1-F_5, or F_1-F_4 in the signal band 0–6 [kHz].

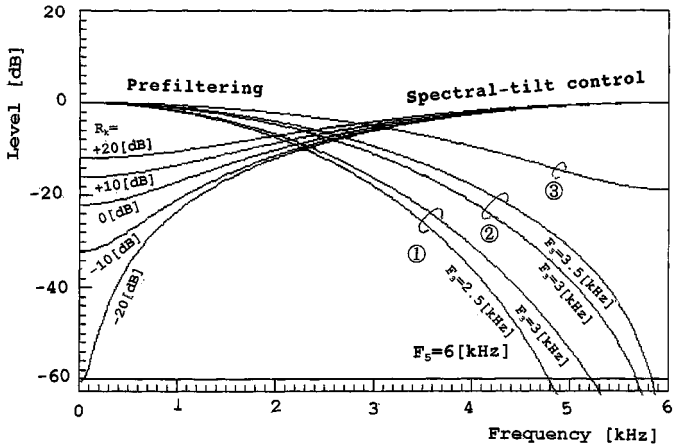


Fig. 3. Characteristics of prefiltering and spectral-tilt control.

III. Inverse Filter Control System

A. Construction and Control

The proposed system is shown in Fig. 2(b). After limiting speech signals to the frequency band f_u-5800 Hz, for antialiasing and for reducing influence of the strong fundamental components on the estimation of F_1, the signals are put in the system. The frequency f_u is changed according to the class of vocal tract lengths to be described later, that is, $f_u = 200$, 300 Hz for the switch ①, ② (or ③), respectively. This system operates at a 12 kHz sampling frequency. The frame length and the frame shift are 20 [ms] and 10 [ms], respectively. A block of inverse filters, called "fundamental block," which is constructed using four basic filters as shown in Fig. 2(a), is one of the main elements for the system. When the boxes [called IFB in Fig. 2(b)] denote the fundamental block, the proposed system consists of a preemphasis, seven basic filters and five fundamental blocks (IFB⟨1⟩-IFB⟨5⟩). The control signals allocate a specific roll individually to each of them. Most of the basic filters, including those in the fundamental blocks, have the fixed bandwidths, which are suitable to the respective resonance. (The arrows for controlling bandwidths are omitted from the figures.) In the whole system, three basic filters, that is, two IF_a's and IF_5, construct a prefilter for removing resonant components higher than fourth. The characteristic of the prefilter is switched by any of three categories (①–③) to be selected based on the class of vocal tract lengths. Those frequency responses of the prefilter are shown in Fig. 3. The bandwidth of IF_a to be determined by the third formant frequency controls the fine change of a frequency region for estimating F_1-F_4 in the category ① or ②, since the third formant is relatively independent of most phonemes in those categories and influenced by a vocal tract length. This prefilter

plays a roll as the buffer to mildly cut off the influence of higher formants on the signal band of lower ones. The left two vertical-blocks of Fig. 2(b) are used for extracting F_4-F_6 in ① and F_4-F_5 in ② Thus, we can gain F_1-F_6, F_1-F_5 and F_1-F_4 by switching ①, ② and ③, respectively. Three categories—①, ②, and ③—roughly correspond to utterances of adult males, adult females, and girls, respectively. Utterances of boys will be appropriate for the characteristic, ② or ③. (The girls and boys are assumed to be under 12 years of age.)

For each frame of speech signals, all basic filters of the whole system shown in Fig. 2(b) are controlled repeatedly from the right box to the left one, that is, from IFB $\langle 2\rangle$ to IFB $\langle 5\rangle$ in the switch ①; from IFB $\langle 2\rangle$ to IFB $\langle 4\rangle$ in ②; and from IFB $\langle 2\rangle$ to IFB $\langle 3\rangle$ in ③. We call these control loops "large loop." In addition, two basic filters that are put at the bottom of each fundamental block [see Fig. 2(a)] are controlled mutually three times as "small loop." In the large loop, the notch frequencies, which are controlled by arrows from the right side of the fundamental blocks or the basic filters, are always updated by the newest output frequencies in that period. After the whole system has repeatedly been controlled until the mean frequencies of FRQ outputs converge, formant frequencies are extracted as outputs of the fundamental blocks. Although we fixed the control times as will be described in Section V-A, the results always showed sufficient convergence. When two outputs of the same formant-number appear in the large loop, the output on the left side of them becomes a better estimate. During the control, the formant frequencies before convergence are always checked and have to be arranged in order for renumbering.

B. Modification of the System for Control of Spectral Tilt

A modification for the control of spectral tilt is indispensable to improve the formant estimation accuracy in speech signals that have a weak resonance adjacent to a strong one. The modification is simply realized using a basic filter IF_0 cascaded from IF_1. IF_0 operates in a constant notch frequency with a variable bandwidth. To control the bandwidth of IF_0, a five-channel filter bank is used, in which each frequency-band approximately covers each formant region. As the notch frequency is fixed at 60 [Hz] in IF_0, the spectral tilt at frequencies higher than 60 [Hz] varies with the bandwidth that is controlled by

$$\begin{aligned} B_k^{[\mathrm{Hz}]} &= 100R_k + 2200; \quad R_k \geqq -20 \text{ [dB]} \\ &= 200; \quad R_k < -20 \end{aligned} \tag{3.1}$$

where R_k [dB] $= 20\log(y_{k+1}/y_k)$; $k = 1-4$, and $R_5 = R_4$ ($B_5 = B_4$).

In the above equations, k is a channel-number of the filter bank and y_k indicates rms value of kth channel-output signal. Therefore, R_k means a gain [dB] represented by the ratio of rms values of $(k+1)$th to kth-channel output. The number k corresponds to the IFB $\langle k\rangle$ of Fig. 2(b). Namely, IF_1, which is put inside or in the upper part of IFB $\langle k\rangle$, is connected to IF_0 to be controlled by R_k. The characteristics of the tilt control by the parameter R_k are indicated in Fig. 3 together with those of the prefilter. Since all IF_1's are replaced with IF_1 and IF_0, IFC system requires 32 basic filters in all.

IV. Computational Methods of Mean Frequency From the Basic Filter Output

A. Mean Frequency of a Power Spectrum as First Estimation

First of all, let us assume that N sample sequence of a frame in a time signal is defined as follows:

$$\text{Time sequence, } \{x_n\}; \qquad n = 0, 1, 2, \cdots, N-1. \tag{4.1}$$

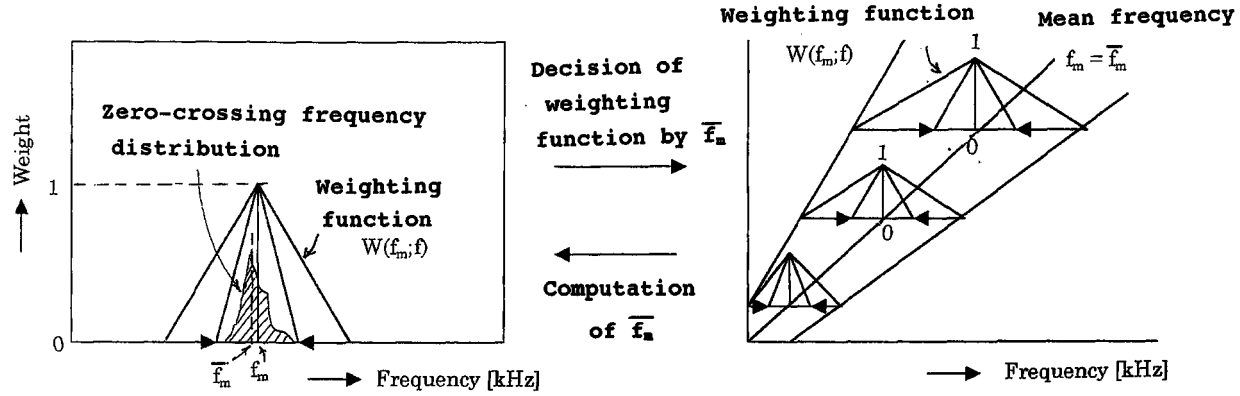
The well-known relation [11] between a power spectrum and an auto-correlation function ϕ_m leads the following equation for computing a mean frequency $\bar{\omega}$, which is a center of gravity of the power spectrum

$$\bar{\omega} = \cos^{-1}\frac{\phi_1}{\phi_0} = \cos^{-1}\frac{\sum_{n=0}^{N-1} x_n x_{n-1}}{\sum_{n=0}^{N-1} x_n^2}. \tag{4.2}$$

Thus, $\bar{\omega}$ can be calculated at a high speed in the time domain. Although the above estimation is convenient to search approximately a mean frequency from the power spectrum, the estimation error will be getting somewhat large as the resonant frequency approaches 0 or π even in the case of perfect single-resonant waves. By this reason, the next algorithm, which is independent of a power spectrum, is needed for more accurate estimation.

B. Weighted Mean of Zero-Crossing Frequency Distribution as the Second Estimation

The following algorithm for the subroutine FRQ has been determined experimentally through the processing of a great number of speech data, which cover a comprehensive area. Let a frame of input signal to FRQ after the control by the first estimation to be given by the time sequence (4.1) again. When two successive samples of the sequence have opposite signs (polarities) of each other, a zero-crossing point exists in a period between them. So, linear interpolation between the two samples makes it possible to estimate a zero-crossing point in high resolving power. By using linear interpolation to seek all zero-crossing points in N samples of a frame, we made the algorithms to get two sets of zero-crossing intervals to be defined as half and full cycles of waveforms. In the first formant, which is estimated from waveforms of a low zero-crossing rate, the half cycles' set is effective to form the zero-crossing frequency distribution. We defined a pseudo-full cycle's interval as two times a half cycle's one. The true full cycles' set has been used for the estimation of the second formant frequency or the higher ones, which have many zero-crossing intervals in a frame. When a zero-crossing frequency is defined as a reciprocal of the full or the pseudo-full cycle's interval, an example of the relation between the zero-crossing frequency distribution and a weighting function will be illustrated as a caricature in Fig. 4(a). The weighting function $W(f_m; f)$ has a maximum value (1) at the mean frequency f_m, to be decided from the power spectrum at first, and decreases linearly from 1 to 0, as the zero crossing frequency becomes distant from the mean. In addition, the weighting function forms a triangle in which a length of

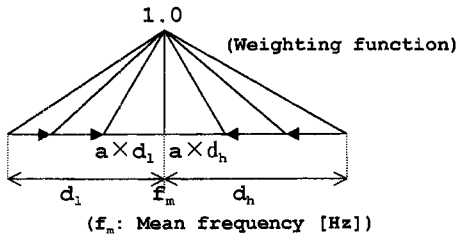


(a) (b)

Fig. 4. Computational method of weighted mean from zero-crossing frequency distribution. (a) Computation of mean frequency $\overline{f_m}$ and (b) decision of weighting function by $\overline{f_m}$.

TABLE I
TRIANGLE FOR THE WEIGHTING FUNCTION AND ITS MODIFICATION

		Base of triangle [Hz] (Initial length)		Multipliers (Reduction factor for base)		
	sw	d_l	d_h	loop 1	loop 2	loop 3-9
F_1	①	$0.3f_m + 140$	$0.3f_m + 140$	a = 1.0	a = 1.0	a = 1.0
	②	$0.3f_m + 240$	$0.3f_m + 340$			
	③	$0.3f_m + 340$	$0.3f_m + 440$	0.75	0.70	0.65
F_2-F_3		$0.63f_m + 174$	$0.63f_m + 174$			
F_4-F_6		$0.84f_m + 230$	$0.84f_m + 230$	0.50	0.40	0.30



the base is a function of the mean frequency, as in Fig. 4(b). This triangular weighting function is effective to remove reasonably zero-crossing intervals disturbed by influence of sound source, which are far from the center of the distribution. When a set of zero-crossing frequencies is given as $\{f_i\}$ in $i = 1, 2, \cdots M$, a weighted mean $\overline{f_m}$ is computed using the weighting function $W(f_m; f)$ as

$$\overline{f_m} = \sum_{i=1}^{M} W(f_m; f_i) f_i \Big/ \sum_{i=1}^{M} W(f_m; f_i) \qquad (4.3)$$

Thus, a new mean-frequency $\overline{f_m}$ is obtained as a weighted mean of the set of zero-crossing frequencies, and next, a vertex of the triangle showing the weighting function shifts to a position of the updated mean-frequency $f_m = \overline{f_m}$. By this operation, a new weighting function is generated in Fig. 4(b). Successively, the mean frequency is updated by computing the weighted mean as shown in Fig. 4(a). During repeating, the calculation of weighted mean and the succeeding decision of weighting function, the base of the triangle is reduced step by step. The weighting function shapes isosceles triangles for formants higher than the first one and the first formant of a long vocal tract to be processed with the switch ① in Fig. 2. On the other hand, a scalene triangle, in which the slope of the low-frequency side of the vertex is steeper than that of the high-frequency one, has been used for the first formant of a short vocal tract, which is implemented with the switch ② or ③ in Fig. 2, to avoid the influence of fundamental component with high pitch on the estimation. In Table I, the lengths of the triangle's base to represent the weighting function are summarized together with its reduction rate (multiplier) in each of the large loops. The meaning of symbols in Table I is indicated in the lower figure. The multiplier ("a" in Table I) always starts from 1 in the input of all FRQs and is reduced two times for the same set of zero-crossing intervals, that is, only when a new set of intervals is put in FRQ as the result

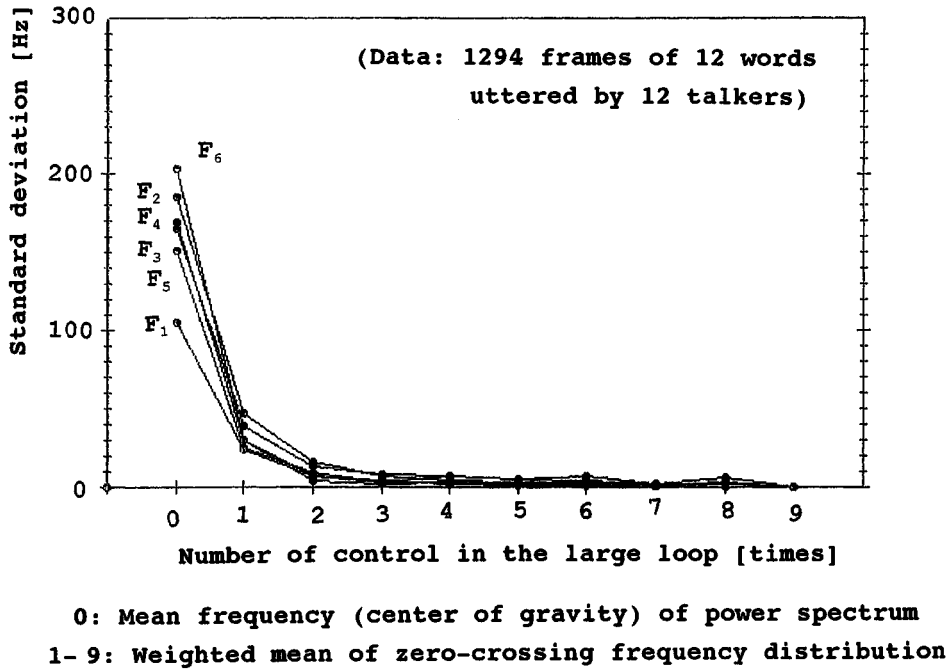


Fig. 5. Convergence of the estimated mean frequency.

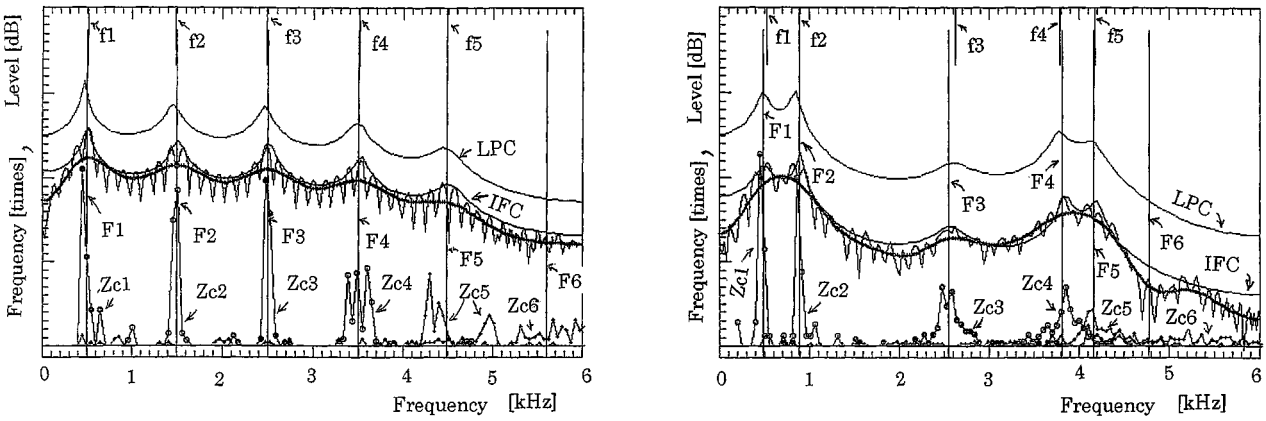


Fig. 6. Examples of formants estimated by LPC and IFC and zero-crossing frequency distributions of components separated by IFC. (a) Synthetic speech (neutral vowel) (first formant: 500 [Hz], pitch: 130 [Hz]). (b) Real speech (male voice) (first /o/ in /aoisora/). f1–f6: Formant frequencies estimated by LPC. F1–F6: Formant frequencies estimated by IFC. Zc1–Zc6: Zero-crossing frequency distributions in IFC. LPC: LPC spectrum. IFC: IFC spectrum (described in Section VI).

of control in the small loop, the multiplier is put back 1. At last, the formant frequencies are estimated by the weighting function of the triangle whose base is equal to 30% of the initial one. Thus, in the subroutine FRQ, the parameters in Table I are just changed by the formant number, the switch number and the loop number to be given from the main routine.

V. Fundamental Characteristics of IFC Method

A. Process of Convergence

We have investigated the convergent process using 1294 frames of 12 compound words uttered by three adult males, three adult females, three boys, and three girls. Fig. 5 shows the convergent process as changes of standard deviations from the respective final values for every formant. In this processing, the same condition as the loop 3 shown in Table I has extendedly been repeated from loop 4 to 9. The estimated formant frequencies satisfactorily converge (S.D. $=$ 2.4–8.2 [Hz]) after three times of the control in the large loop. Based on this result, we have decided on the necessary control times to be three. The formant estimation accuracy after three times' control of the large loop will be evaluated by Monte Carlo simulation using synthetic speech as described in the Section V-C1.

B. Zero-Crossing Frequency Distributions After Convergence

Fig. 6(a) indicates FFT and LPC (order of the model, $p = 14$) spectra of speech synthesized using Rosenberg's sound source [12] and the system function of a vocal tract with stationary parameters. In the figure, the zero-crossing frequency distributions and the formant frequencies estimated after convergence in IFC method are shown together with the spectra. Each zero-crossing frequency distribution is extremely localized in Zc1–Zc3 and

TABLE II
STANDARD DEVIATION OF ERRORS (ESTIMATION BY MONTE CARLO METHOD)

(F_0)	(80-180)[Hz]*		(150-350)[Hz]*		(200-400)[Hz]*	
	IFC	LPC	IFC	LPC	IFC	LPC
	①**	(p=14)***	②**	(p=12)***	③**	(p=10)***
F_1	±8.7[Hz]	±17.2[Hz]	±24.1[Hz]	±40.7[Hz]	±41.9[Hz]	±44.8[Hz]
F_2	±5.9	±12.1	±15.5	±29.4	±18.9	±41.0
F_3	±4.2	±9.3	±18.0	±27.4	±24.1	±35.3
F_4	±11.3	±10.1	±42.5	±33.6	±62.9	±82.5

the extracted formant frequencies are very close to those given for synthetic speech (a neutral vowel). And that is correct even in the case when the formant frequency is given between two harmonic components. An example of real speech shown in Fig. 6(b) has the same tendency as it of the synthetic speech. Moreover, the mean values of zero-crossing frequency distributions approximately correspond with the frequencies of LPC spectral peaks and the formant frequencies estimated by the root-solving method of LPC. Thus, it is likely that the mean frequencies of the zero-crossing frequency distributions indicate formant frequencies.

C. Comparative Evaluations of Formant Estimation Accuracy between IFC and LPC

1) Evaluation using Synthetic Speech: For comparing the accuracy of formant frequencies estimated by IFC and LPC, we have adopted the Monte Carlo method using synthetic speech. For the test, a simulated terminal analog synthesizer has synthesized speech signals using pitch (F_0) and formant frequencies (F_1-F_4), which we randomly selected from the respective frequency ranges. The frequency ranges are the same as those used for the spectral tilt control in IFC except the third formant. Only the third formant frequency is randomly selected from the range of ±500 [Hz] on the regression functions estimated by the first and second formants of real vowels [13]. The formant frequencies higher than fourth are fixed. The bandwidths of the lowest five formants (four in the highest pitch range to be described later in this paper) were given as a function of formant frequency, assumed on the basis of the experimental data by Fant [14]. Moreover, we also gave wide bandwidths to the fixed formants higher than the fifth formant (fourth formant in the highest pitch range), to keep an approximately suitable tilt in the spectrum. The number of formants to use for synthetic speech has been changed in accordance with the pitch frequency range, assuming correlation between pitch and a vocal tract length. That means seven formants in a pitch range of 80–180 [Hz], six in 150–350 [Hz], and five in 200–400 [Hz], respectively. Radiation characteristic was approximated by the first derivative of a sound-source wave. A cascade form of second-order subsystems simulated the vocal tract.

In the IFC method, the analysis condition for each of the three pitch ranges, 80–180 [Hz], 150–350 [Hz], and 200–400 [Hz], has been switched by ①, ②, and ③, respectively, in the system of Fig. 2(b). Likewise, the order of the LPC model to be used for the analysis has been 14 for 80–180 [Hz] in pitch, 12 for 150–350 [Hz], and 10 for 200–400 [Hz]. In LPC, after the speech signals were preprocessed by preemphasis and windowing, linear predictor coefficients have been estimated by the autocorrelation method. Then, we have used the polynomial root-solving procedure to get formants. The preemphasis is the first backward difference in time sequence like IFC.

Each test consists of 500 trials whose combinations for pitch and the lowest four formant-frequencies are all different. We have evaluated the errors by standard deviations due to the difference between a given formant frequency and the estimated one. The errors, as summarized in Table II, are smaller in IFC than in LPC except two cases of F_4. However, these error magnitudes for both methods may somewhat change due to gross spectral tilt, which is given by higher formants of the synthetic speech. So, it might be difficult to insist constant superiority of IFC to LPC by these data alone. Still, based on the magnitudes of errors, it is no doubt that IFC method provides good estimates with high accuracy like LPC.

2) Analysis Examples of Real Speech: Synthetic speech is well regulated for evaluation meaning that bandwidths given for the lowest four formants are relatively small. To make the difference between IFC and LPC clearer, let us observe some analysis examples of real speech next. Fig. 7 shows the raw (no-smoothing) formant trajectories of four compound words uttered by an adult male, a boy, an adult female, and a girl. (To manifest the gross errors, two adjacent points are connected with a straight line.) Though two kinds of trajectories, which have been estimated from the same utterance using each of both methods, are very similar to each other, we can find more gross errors in LPC than in IFC. In LPC data in the right column of Fig. 7, we have extracted the formants only that have a bandwidth below 1000 [Hz]. Augmenting the allowable bandwidth may reduce loss of formants, but cannot reduce excess of it. On the contrary, decreasing the bandwidth may reduce excess but cannot restore loss. Therefore, when loss and excess of formants are mixed under a condition as shown in LPC data of Fig. 7, it is impossible to remove perfectly the gross errors by varying the allowable bandwidth. In contrast, the formant trajectories

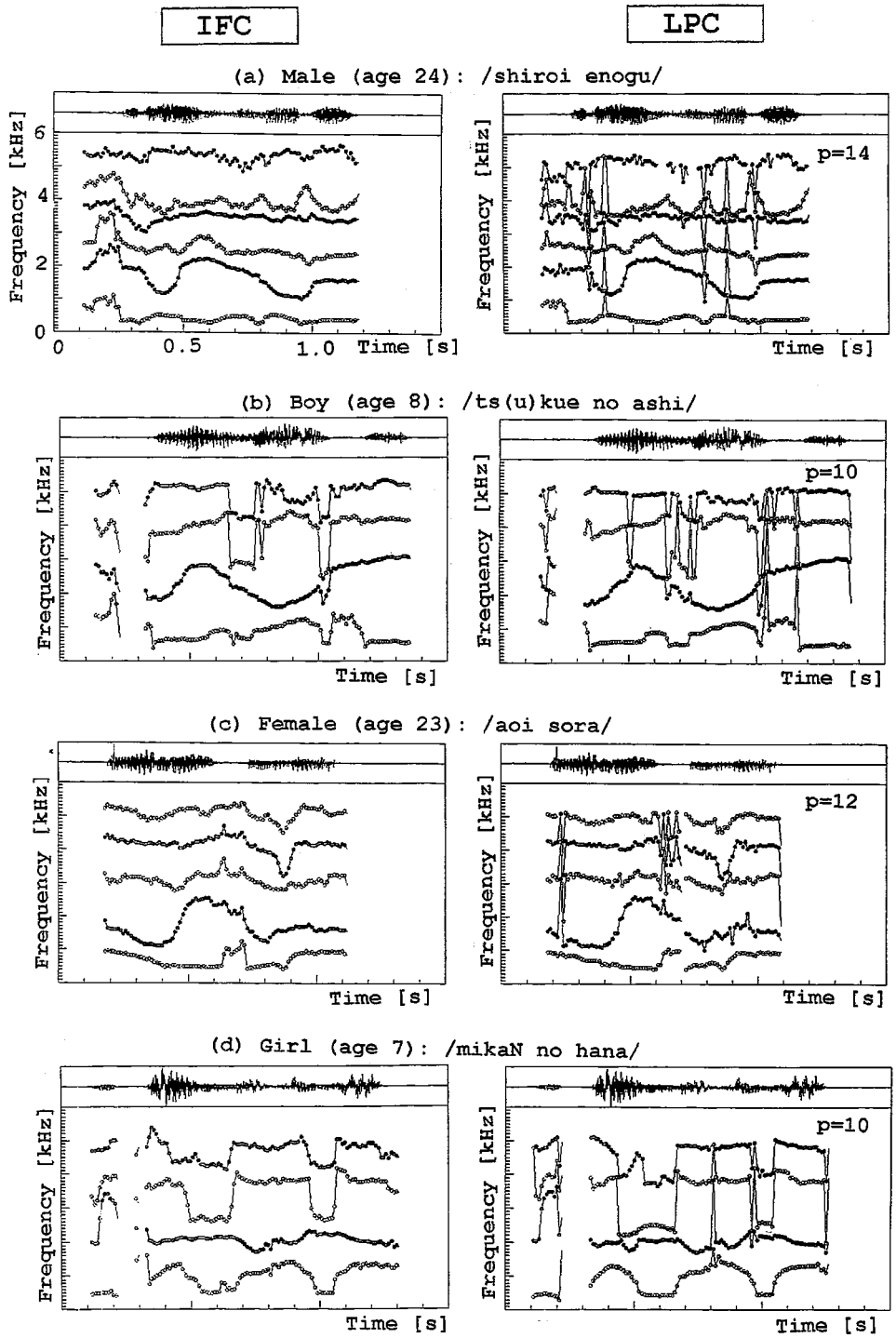


Fig. 7. Examples of formant trajectories estimated by IFC and LPC (no smoothing).

by IFC method are continuous and legible. In addition to that, the trajectories by IFC convince us that they are caused by the correct phoneme sequence of a given word. It seems that IFC is superior to LPC in the stability of the estimation.

VI. Specific Features of IFC Method

The IFC method does not need any algorithm to estimate spectral shapes, but searches directly for the resonant frequencies. Therefore, when the number of formants assumed in the model equals to that of relatively clear resonance in speech signals, we can always estimate the fixed number of formant frequencies. In other words, partial loss or excess of formants never occurs in IFC method. This is important for acquiring stable formant trajectories, because if there is no loss and no excess of formants, we can apply a mild smoothing method to the raw data. Finally, we tried to estimate a spectral envelope based on the least-mean-square error criterion, leaving the estimated formant frequencies unchanged. After the estimation of formant frequencies, a spectral envelope has been synthesized by an

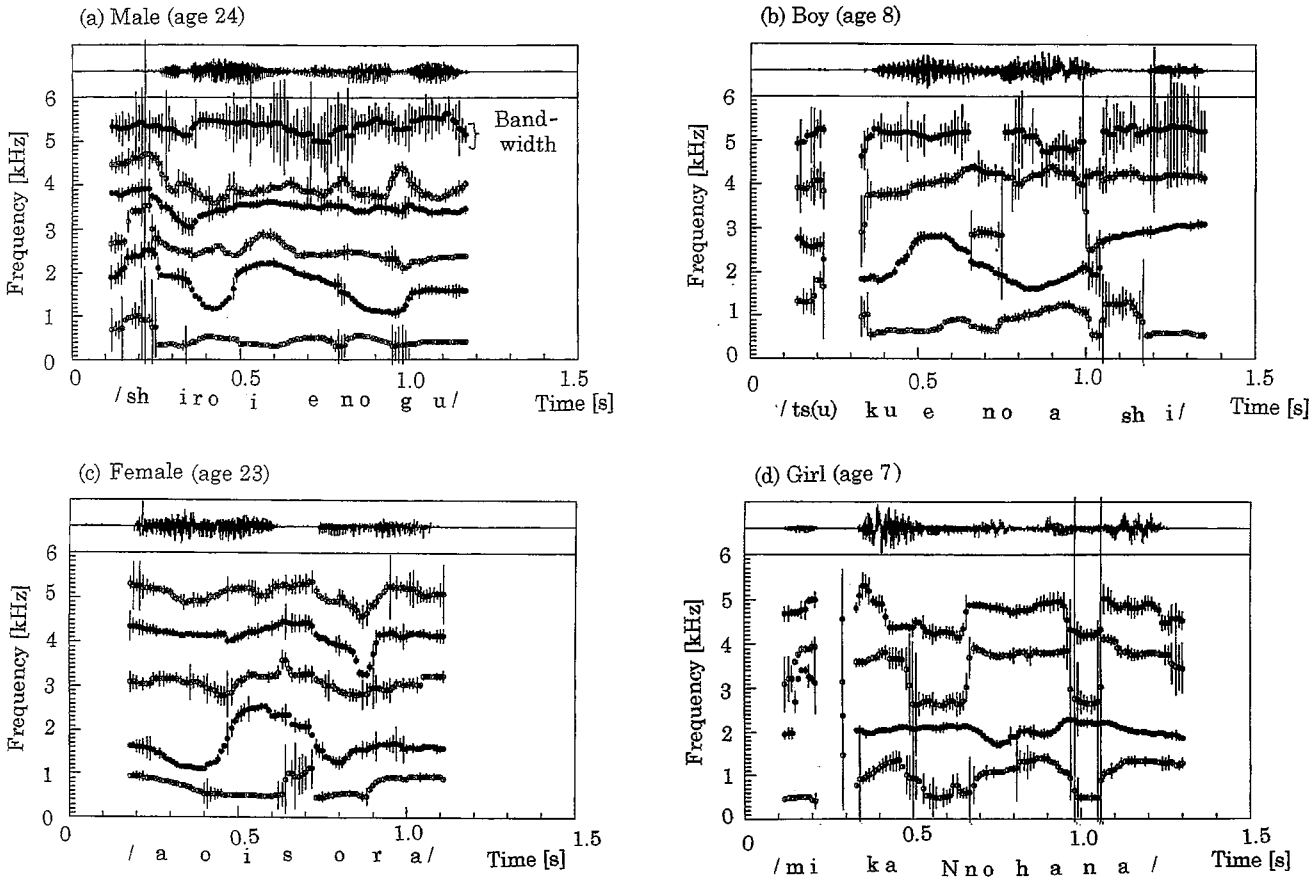


Fig. 8. Estimated formant bandwidths and the effect of smoothing in formant trajectories (after three-point median smoothing).

all-pole model with those formant frequencies and an additive resonance whose peak is fixed at a frequency of 0 [Hz] to control the spectral tilt by the change of bandwidth. The desired spectral-envelope can be estimated by controlling all of the formant bandwidths so as to minimize the mean-square-error between the synthetic envelope and the analyzed one at frequencies of harmonic peaks, which are defined as peaks larger than the magnitudes of smoothed spectrum by FFT cepstrum of 0th–26th order (see the thick line in Fig. 6). We used the steepest descent method to seek a set of the bandwidths for the least-mean-square error. Examples of the estimated spectral-envelopes are shown by the mark of "IFC" in Fig. 6. Fig. 8 shows formant trajectories for the same compound words as those of Fig. 7. In Fig. 8, the vertical bars represent the estimated formant bandwidths. Only the formant frequencies are processed using three-point median smoothing method. As can be seen in these figures, the reliable estimates of continuous or quickly shifting formant-trajectories are obtained without rejecting formants of wide bandwidths.

VII. Conclusion

We have proposed a new method for estimating formant frequencies, called the IFC method. In this method, 32 basic (notch) filters, whose system function is characterized by a pair of complex-conjugate zeros, are mutually controlled so as to separate speech waves into approximate single-resonant waves. After convergence of zeros (notches) in all the basic filters, formant frequencies are estimated from weighted means of zero-crossing frequencies of the separated waves. This method is substantially different from LPC or A-b-S, which are spectral matching methods, because of the direct estimation of resonant frequencies from the zero-crossing frequency distributions. We compared the performance of the IFC and LPC methods using synthetic speech and by analysis examples of real speech. Although superiority of IFC to LPC was not necessarily prominent in the test by synthetic speech, the errors were sufficiently small. On the other hand, the gross errors in some formant trajectories of real speech were much fewer in IFC than in LPC. IFC presents more stable formant trajectories than LPC. Finally, we have searched the formant bandwidths by means of the spectral envelope matching, based on the formant frequencies estimated by IFC. As a result, it has been shown that the IFC method acquires reliable and stable formant trajectories by allowing formants of wide bandwidths.

The proposed IFC method has already been applied to three specific studies: speech visualization [15], speech recognition [16], and estimation of ratio of vocal tract lengths between two speakers' [10]. All studies showed that the formant frequencies estimated by the IFC method were practically useful due to stability and high accuracy.

Acknowledgment

The author deeply appreciates the encouragement of Dr. G. Fant when starting this research in the Royal Institute of Technology (KTH). He also wishes to thank Dr. Y. Ueda and many graduates who collaborated in this research.

REFERENCES

[1] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1970.

[2] C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House, "Reduction of speech spectra by analysis-by-synthesis techniques," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1725–1736, 1961.

[3] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Berlin, Germany: Springer-Verlag, 1976.

[4] J. C. Wood and D. J. B. Pearce, "Excitation synchronous formant analysis," *Proc. IEEE*, vol. 136, pp. 110–118, 1989.

[5] B. Yegnaratama and R. N. J. Veldhuis, "Extraction of vocal tract system characteristics from speech signals," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 313–327, July 1998.

[6] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 36–48, Jan. 1998.

[7] A. Watanabe, "A real-time formant tracker using inverse filters," *STL-QPSR*, vol. 3–4, pp. 1–30, 1980.

[8] A. Watanabe, Y. Ueda, and A. Shigenaga, "Color display system for connected speech to be used for the hearing impaired," *IEEE Trans. Audio, Speech, Signal Processing*, vol. ASSP-33, pp. 164–173, 1985.

[9] J. M. Pickett, *The Sounds of Speech Communication*. Baltimore, MD: Univ. Park Press, 1980, pp. 46–49.

[10] A. Watanabe, T. Sakata, and R. Masuda, "Estimation of ratio of vocal tract lengths based on formant trajectories" (in in Japanese), *IEICE*, vol. J83-A, pp. 230–233, 2000.

[11] Y. W. Lee, *Statistical Theory of Communication*. New York: Wiley, 1967, pp. 56–59.

[12] A. E. Rosenberg, "Effect of glottal pulse on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, pp. 583–588, 1971.

[13] A. Watanabe, R. Hamasaki, and Y. Ueda, "Parameters' extraction and estimation methods for visualization of telephonic speech" (in in Japanese), *J. TV. Eng. Jpn.*, vol. 45, pp. 233–243, 1991.

[14] G. Fant, *Speech Sounds and Features*. Cambridge, MA: MIT Press, 1973, pp. 6–15.

[15] A. Watanabe, S. Tomishige, and M. Nakatake, "Speech visualization by integrating features for the hearing impaired," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 454–466, July 2000.

[16] T. Dutono, N. Ikeda, and A. Watanabe, "Effects of compound parameters on speaker-independent word recognition," *J. Acoust. Soc. Jpn.*, vol. 19, pp. 1–11, 1998.

Akira Watanabe received the B.E. degree in electrical engineering in 1962, and the M.E. and D.E. degrees in electrical and communication engineering from Tohoku University, Sendai, Japan, in 1964 and 1968, respectively.

He has been with the Faculty of Engineering, Kumamoto University, Kumamoto, Japan, since 1967. He investigated a real-time formant estimation system from 1978 to 1979 as a Guest Researcher at the Royal Institute of Technology (KTH), Stockholm, Sweden. He is now a Professor with the Department of Computer Science, Kumamoto University. His research interests include speech processing and coding for the hearing impaired.