SPEECH SIGNAL CODING USING LINE SPECTRUM PARAMETERS

(線スペクトルパラメータによる音声符号化に関する研究)

by

Noboru SUGAMURA

July, 1984

ACKNOWLEDGEMENTS

The author would like to express his sincere appreciation to Dr. Fumitada Itakura who recommended the thesis presentation, for his kind encouragement, guidance and discussions throughout this study.

The author would like to express his sincere gratitude to Professor Katsuhiko Fujii, Faculty of Engineering, Osaka University, for encouragement, guidance, suggestions and kind arrangement concerning the thesis presentation.

The author also would like to express his sincre gratitude to Professors Osam Kakusho, Institute of Scientific and Industrial Research of Osaka University, Yutaka Suzuki and Yoshikazu Tezuka, Faculty of Engineering of Osaka University, for suggestions and kind arrangement concerning the thesis presentation.

The author wishes to thank Professors, Hitoshi Kinoshita, Chiyoe Yamanaka, Yoshio Inuishi, Faculty of Engineering of Osaka University, for kind arrangement concerning the thesis presentation.

The author wishes to express his sincere thank to Drs. Noriyoshi Kuroyanagi, Shinichiro Hashimoto, Sadakuni Shimada and Mr. Tsunehiko Koike for their kind encouragement and guidance, and for giving him a chance to achieve this study. Sincere thanks are also due to Professors Shuzo Saito of Tokyo University and Kenichi Noda of Tokyo University of Agriculture and Technology for encouragement and guidance.

The author is deeply indebted to Drs. Yoh'ichi Tohkura and Nobuhiko Kitawaki, for kind suggestions and discussions about PARCOR speech analysis synthesis method in this study. Sincere thanks are also due to Drs. Sadaoki Furui, Kiyohiro Shikano and Mr. Kazuo Hakoda, for invaluable discussions about word recognition in this study.

The author wishes to express his thanks to the many colleagues who

cooperated in several listening experiments.

The author is greatly indebted to Mr. Shigeki Sagayama for making editor programs in writing up this thesis.

ABSTRACT

This dissertation describes the research on narrow band transmission in speech coding.

Two equivalent representations in the frequency domain of linear predictive coefficients or PARCOR coefficients are introduced. One is "line spectrum representation (LSR)" and the other is "line spectrum pair (LSP)". The characteristics of these parameters in quantization and interpolation are clarified experimentally comparing PARCOR coefficients. The optimal coding methods in these new vocoders are proposed. Synthesized speech quality is assessed by objective and subjective measurements. These results lead to the superiority of new parameters proposed in this dissertation.

Moreover, a transmission technique according to the rate of parameter transition is presented. Speech events occur generally at non-uniformly spaced time intervals according to the articulator movements. Then, sampling at a variable rate in accordance with changing speech signal characteristics is more efficient.

Another technique to eliminate the redundancy in parameter distribution space was studied. Vector quantization technique using LSP parameters is executed to achieve an extreme low bit rate transmission.

This quantization technique is applied to isolated word recognition from the standpoint of quantizing speech signals at a larger unit than a frame.

TABLE OF CONTENTS

LIST OF SYMBOLS

SYMBOL	CONTENTS	REMARKS
$x(i)$	sampled value of speech waveform	
N	number of speech samples	Chapter 3
Δ T	sampling period	
f_N	sampling frequency	Chapter 2
α_i	linear predictive coefficients	
$e(n)$	excitation source signal	
σ	root mean square of e(n)	
$X(z)$	z-transformation of x(n)	
$E(z)$	z-transformation of e(n)	
$H(z)$	transform function (=X(z)/E(z))	
$A_p(z)$	all zero filter	
$S(\omega)$	power transform function of H(z)	
V_i	autocorrelation coefficients	
P	pitch period	
p	number of prediction (analysis order)	
$T(\omega)$	short time spectrum	
$f^{(P-1)}(n)$	forward prediction residual	
$g^{(P-1)}(n)$	backward prediction residual	
k_i	PARCOR coefficients	
$A_p(D)$	polynomial of lattice filter	
$B_p(D)$	polynomial of lattice filter	
D	shift operator for unit time	
β_i	backward predictive coefficients	

SYMBOL	CONTENTS	REMARKS
$R(m)$	autocorrelation coefficient of prediction residuals	
A_i	maximum likelihood parameters	
A_n	cross section of acoustic tube	
S_{ij}	S-matrix components	
J_n	equivalent circuit for n-th acoustic tube junction	
$w_n(z,s)$	impulse response at right side of J_n	
s	impulse response at left side of J_n	
$R_n(z)$	immitance function	
m_i	line spectrum amplitude	
f_i (λ_i)	line spectrum frequency	
(m_i, f_i)	line spectrum parameters (LSR parameters)	
$S_0(\omega)$	estimated spectrum without quantization	
$S(\omega)$	distorted spectrum by quantization	
$\Delta S(\omega)$	spectral distortion in a frame	
$\overline{\Delta S(\omega)}$	averaged spectral distortion over a long period	
C_{m_i}	spectral sensitivity of m_i	
C_{f_i}	spectral sensitivity of f_i	
W	input speech bandwidth	
$\Delta\tilde{S}$	estimated spectral distortion	
Δ_i	quantization error of i-th spectral parameter	
$P(z)$	polynomial at $k_{p+1} = +1$	
$Q(z)$	polynomial at $k_{p+1} = -1$	

SYMBOL	CONTENTS	REMARKS
(ω_i, θ_i) or (f_i, g_i)	line spectrum pair (LSP parameters)	
a_i	LSP digital filter's coefficient ($=-2\cos\omega_i$)	
b_i	LSP digital filter's coefficient ($=-2\cos\theta_i$)	
T	duration of speech signal in a segment	Chapter 5
N	number of division	Chapter 5
$G(t_{n-1}, t_n)$	cost function of spectral distortion	
$S(t_{n-1}, t_n)$	spectral distortion in a segment ($t_{n-1} < t < t_n$)	
n(t)	noise signal	
s(t)	speech signal	
k	gain factor for noise added speech	Chapter 5
g_i	vectors of spectral pattern	Chapter 6
θ	threshold for clustering	Chapter 6,7
$d(g_i, g_j)$	spectral distance between two vectors g_i and g_j	
N(i)	number of vectors in same cluster	
$R(\theta, 1)$	covering rate	
C_i	cepstrum coefficients	Chapter 7
M_i	number of frames in i-th word	
N_S	number of phoneme-like patterns	
N_a	spectral parameters' accuracy	
N	number of spectral parameters	Chapter 7
n_b	bits of phoneme-like patterns ($= \log_2 N_S$)	

SYMBOL	CONTENTS	REMARKS
L	vocabulary size	
N_W	window length in dynamic time warping	

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 PROLOGUE

1.1 Introduction

1.1.1 Mechanism of Acoustical Speech Waveform Generation [1],[2],[3]

Speech signals are the most convenient and easiest means for human beings to communicate with each other. The acoustical speech waveform is generated by physiological movements of the articulator shown in Fig. 1.1. Air in the lungs is expelled into the trachea and then forced between the vocal cords. In the case of voiced sounds, such as /o/ in OSAKA, the air pushed toward the lips from the lungs causes the vocal cords to open and close at a rate dependent upon the air pressure in the trachea and the physiological adjustment of the vocal cords. This adjustment includes changes in the length, thickness, and tension of the vocal cords. The greater the tension, the higher the perceive pitch or fundamental frequency of the voice. The opening between the vocal cords is defined as the glottis. The vocal tract is a non-uniform acoustic tube which extends from the glottis to the lips. The vocal tract varies in shape as a function of time, according to the voice sounds being emitted. The major components causing this time-varying change are the lips, jaw, tongue and velum. In the case of unvoiced sounds, such as /f/ in FUJI, these sounds are generated by voluntarily holding the vocal cords open, forcing air past them, and then using the articulation to create construction. A plosive sound, such as /p/ in SAPPORO, is generated by building up air pressure in the mouth and then suddenly releasing air.

1.1.2 Acoustical Waveform and Its Spectrum

To illustrate the acoustical implication of speech production in both the time domain and frequency domain, sounds such as /a/ and /tʃ/ were spoken and analyzed. Figures 1.2A and 1.3A show the acoustical waveform. This waveform was obtained by low-pass filtering the tape recorder output to a 4 kHz bandwidth, performing 14 bit analog to digital conversion into a computer system at an 8 kHz sampling rate. The waveform in Fig. 1.2A is nearly periodic in comparison with that in Fig. 1.3A. The distance between major peaks shows the pitch period of the glottal vibration. The waveform in Fig. 1.3A exhibits no discernible pitch period, since the sound /tʃ/ is produced by turbulence noise generated by directing the air stream past a construction formed by the tongue and teeth. Figures 1.2B, C, D and Figures 1.3B, C, D show spectral shapes on a log amplitude scale (in units of dB) versus frequency. Figures C and D correspond to the vocal tract resonance and the vocal cords vibration, respectively. The locations of the major peaks on the envelope in Fig. 1.2C, labeled F_1, F_2 and F_3, which define estimates of the formant frequency locations, correspond with the vocal tract resonance. The spectrum in Fig. 1.2D shows cyclic behavior every 1/P units of frequency. Its fundamental frequency corresponds to the vibration of vocal cords. The spectrum in Fig. 1.2D is nearly flat versus frequency. In addition, a spectrum in Fig. 1.2B is superimposed upon each log magnitude spectrum of Fig. 1.2C and D. On the contrary, in the case of unvoiced sound, there are no strong resonance points, as shown in Fig. 1.3C. In addition, the spectrum in Fig. 1.3D has a nearly white noise property.

Other frequency domain representations of this utterance from sonograms (also referred to as voiceprints or spectrograms), are shown in Fig. 1.4 and Fig. 1.5. using wideband and narrowband filters, respectively. The sonogram shows speech energy as a parameter on a continuous frequency scale versus time. During voiced regions, the dark bars indicate the locations of the resonances as functions of time. The voiced regions show vertical striations, corresponding to the epoch or beginning of each pitch period. During unvoiced intervals, the dark areas indicate major concentrations of energy. With the wideband sonograph filter, pitch period resolution is obtained in the time domain, but a larger amount of averaging or smearing in the frequency domain occurs. With the narrowband sonograph filter, frequency resolution is obtained at the expense of time resolution. With the narrowband filter, the harmonic structure of voiced speech is obtained; that is, the narrow horizontal lines indicate the harmonics of the fundamental frequency. During unvoiced sounds, no discernible harmonic structure is evident.

1.2 Digital Coding of Speech

In order to transmit speech signal waveforms, many coding techniques have been developed. Recently, many digital coding transmission techniques have been proposed. These techniques can be divided into two categories. One is speech waveform coding, the other is speech analysis synthesis system, the so called "vocoder". Several waveform coding methods are listed in Table 1.1 [4] [10]. These methods aim at encoding the original speech waveform as precisely as possible with a lower bit rate. Typical speech waveform coding, PCM (Pulse Code

Modulation), needs 64 kbps to encode speech signals. On the other hand, in the speech analysis synthesis field, many methods have been developed, too, as shown in Table 1.2 [11] [21]. These methods are based on speech spectrum reservation principle, not on the speech waveform. They don't aim at constructing the original speech waveform itself. Conserving the speech spectrum is sufficient to ensure adequate speech quality and intelligibility for many purposes. The natural speech signal is redundant and there are many efficient speech coding methods available both for transmission and storage of voice using these properties. In fact, the movement of articulators is very slow, about 50 Hz, never 8 kHz. In a speech analysis synthesis system, some feature parameters corresponding to speech spectra are extracted at the analyzer and transmitted. At the receiver, the speech waveform is synthesized from these feature parameters. The basic configuration of a speech analysis synthesis system is shown in Fig. 1.6. The difference among several methods in Table 1.2 is the difference in feature parameters which represent the speech spectra. The coding efficiency and synthesized speech quality mainly depend on the properties of the feature parameters and their extraction method. A typical, and the first, speech analysis synthesis method is "Channel Vocoder", invented in 1939 by Homen Dudley of Bell Telephone Laboratories. Vocoder is a term derived from the words VOice CODER [22].

A spectrum channel vocoder blockdiagram is shown in Fig. 1.7. The speech signal is separated into 14 contiguous spectral bands with bandwidths ranging from 200 Hz to 3200 Hz. The output from each filter is connected to a rectifier and low pass filter, whose output represents the time-varying average signal amplitude for each frequency band. Together, these 14 channel signals represent the envelope of the

short time spectrum of the speech signal. Also shown in Fig. 1.7 are a voiced-unvoiced detector and a pitch detector which determine the fine structure of the speech signal and produce a corresponding narrowband signal. These 15 narrowband signals are combined into a single signal with a total bandwidth of 15 times 20 Hz. Thus, the transmission bandwidth is only one-tenth of that required for the original speech signal. At the synthesizer, the original channel signals are recovered and utilized to control the frequency response of a time-varying filter to correspond to the spectral envelope measured at the analyzer. The input for this time-varying filter is supplied with a flat spectrum excitation signal of the proper spectral fine structure. Namely, quasi-periodic pulses are used for voiced speech sounds, or white noise is used for unvoiced sounds.

After his work, much effort has been expended to realize efficient speech signal coding. Most of the research interest was concentrated on finding feature parameters which express speech characteristics efficiently. Among these, the most effective and successful method is the speech analysis synthesis, based on linear prediction of a speech signal, to be introduced in Chapter 2. Moreover, recently, research interest in speech coding is focused in extreme low bit rate transmission, under 1000 bps, as listed in Table 1.3 [23] [32]. A very interesting problem involves how to eliminate the speech signal redundancy. In most of these methods, redundancies in feature parameters, in both the space domain and the time domain, are utilized as much as possible. Theoretically, the speech information rate is assumed to be 50 bps and it is possible to compress speech to this rate. Realizing the ultimate speech signal compression requires phoneme, syllable, word or continuous speech recognition at the

analyzer. At the receiver, it is necessary to reconstruct synthesized speech signal from these units by synthetic rules. This is an ultimate speech analysis synthesis form. Furthermore, it may be possible to transmit only the meaning of utterances or conversation, when research on speech understanding is much advanced. To sum up, the hierarchy of speech coding is shown in Fig. 1.8. Typical coding techniques are shown chronologically in relation with information rates.

1.3 Study Objective and Outline for This Thesis

This thesis is mainly concerned with a speech signal data compression method at an extremely low bit rate.

In Chapter 2, a typical vocoder system, based on linear prediction of speech signal, named PARCOR, is reviewed briefly for reference in succeeding chapters.

In Chapter 3, another representation of linear predictive coefficients in the frequency domain is introduced. The quantization and interpolation characteristics are investigated, comparing with PARCOR coefficients. A new speech analysis synthesis system, named "Pseudo Formant Vocoder", is proposed.

In Chapter 4, another representation equivalent to PARCOR coefficients, called line spectrum pair (LSP) is explained. LSP parameter properties are examined theoretically and experimentally. An optimum bit allocation and interpolation methods are proposed. Synthesized speech quality is assessed through objective and subjective measurements.

In Chapter 5, the technique to eliminate the parameter redundancy in the time domain is proposed. A transmission technique according to

the rate of parameter transition, named variable frame rate transmission (VFR) is studied. The effectiveness of this method is clarified by a quality assessment subjectively.

In Chapater 6, a technique to eliminate the parameter distribution redundancy in the feature space is proposed. Using the good LSP parameters interpolation property, an efficient pattern matching vocoder is studied. Typical spectral patterns are obtained using vector quantization. A spectrum in each frame is represented as a vector of feature parameters. The differences between scalar and vector quantization are examined as a function of spectral distortion.

In Chapter 7, the vector quantization technique, proposed in Chapter 6, is applied to an isolated word recognition based on dynamic time warping. A new efficient isolated word recognition method is realized with representing word templates as sequences of discrete symbols which correspond to a spectrum.

In Chapter 8, the conclusions of this thesis are summed up.

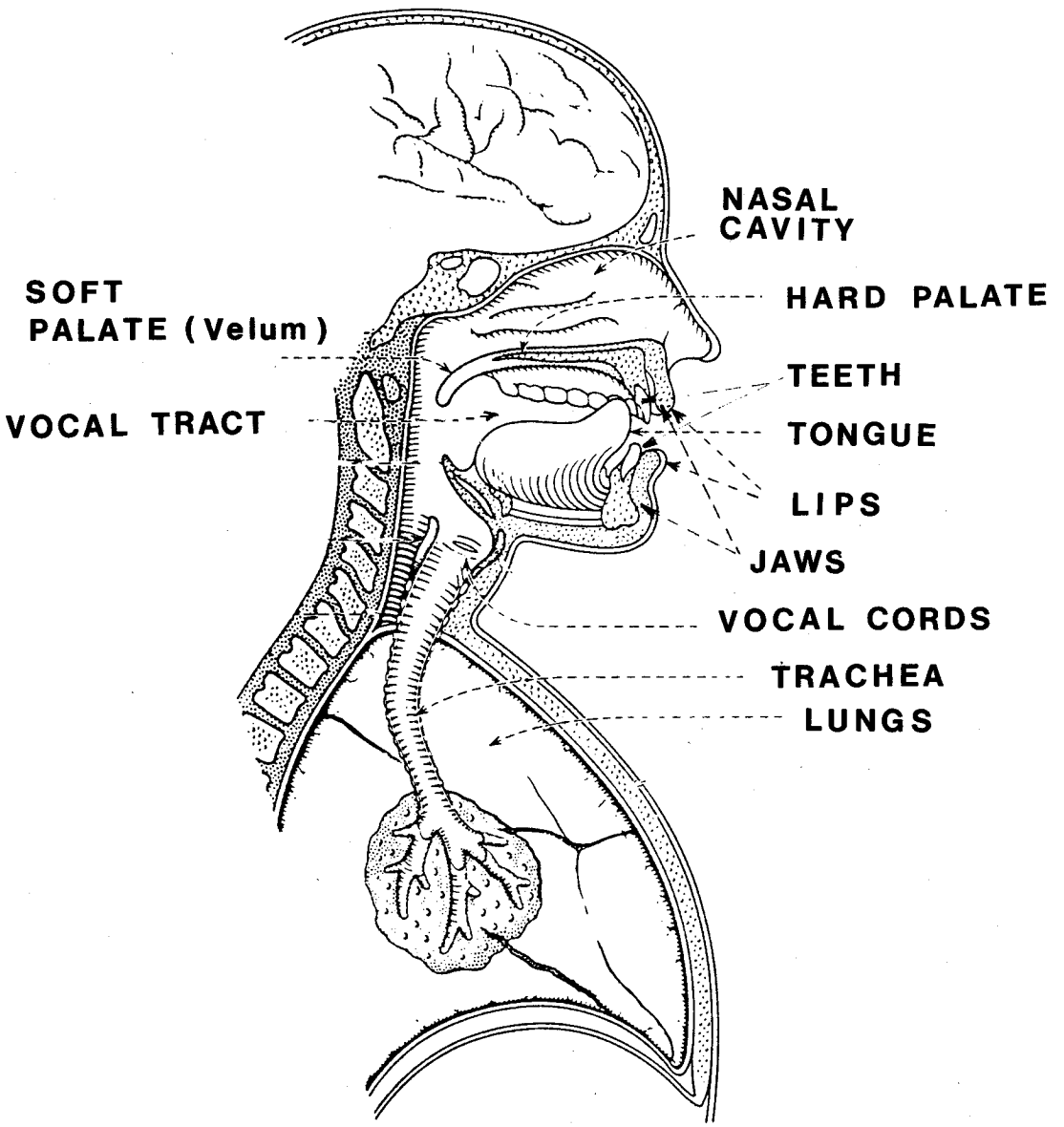


Fig. 1.1 Cross-sectional view of the vocal mechanism involved in speech production (from reference [1], p.10)

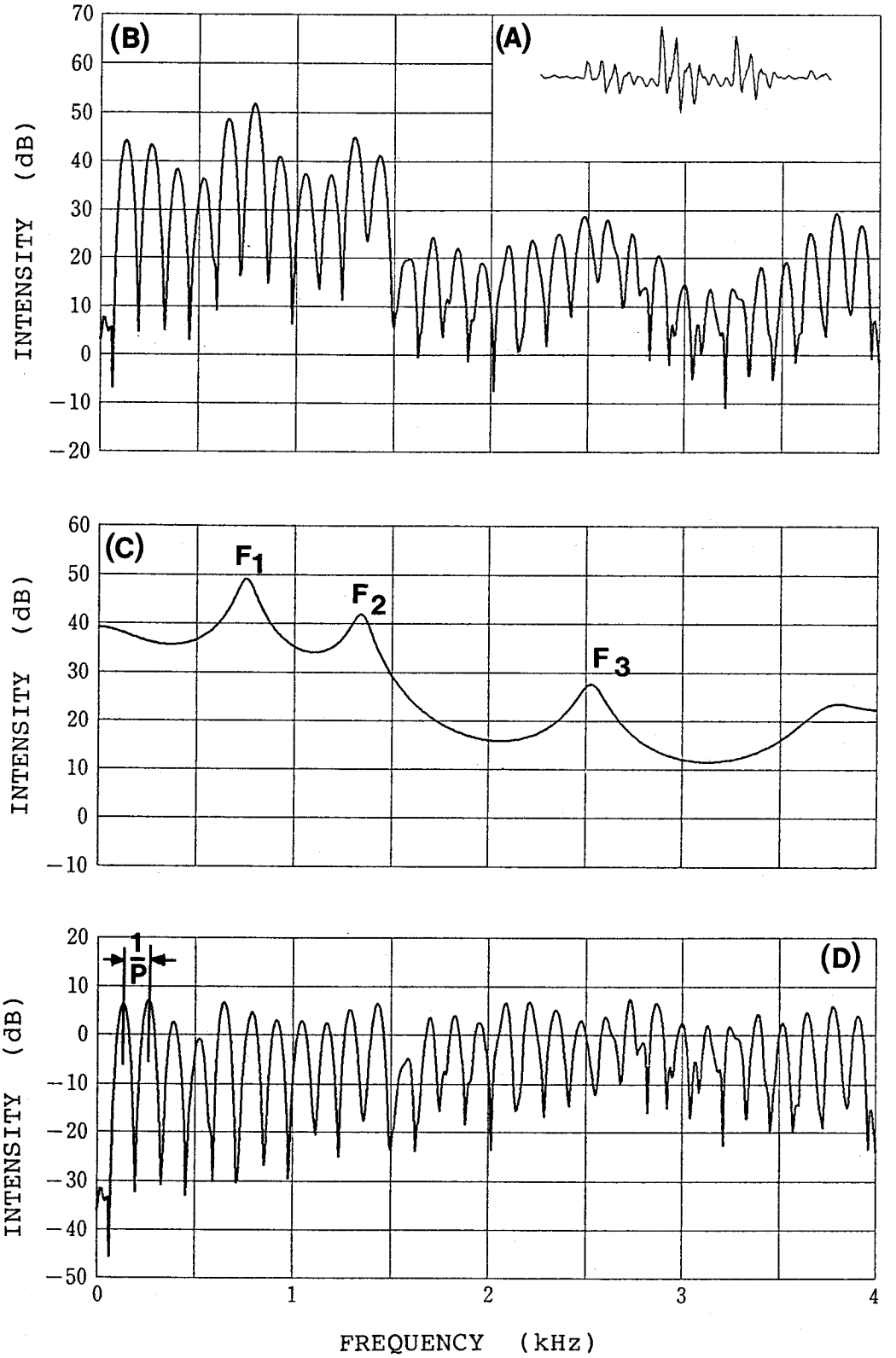


Fig. 1.2 Time and frequency representations of the utterance /a/.

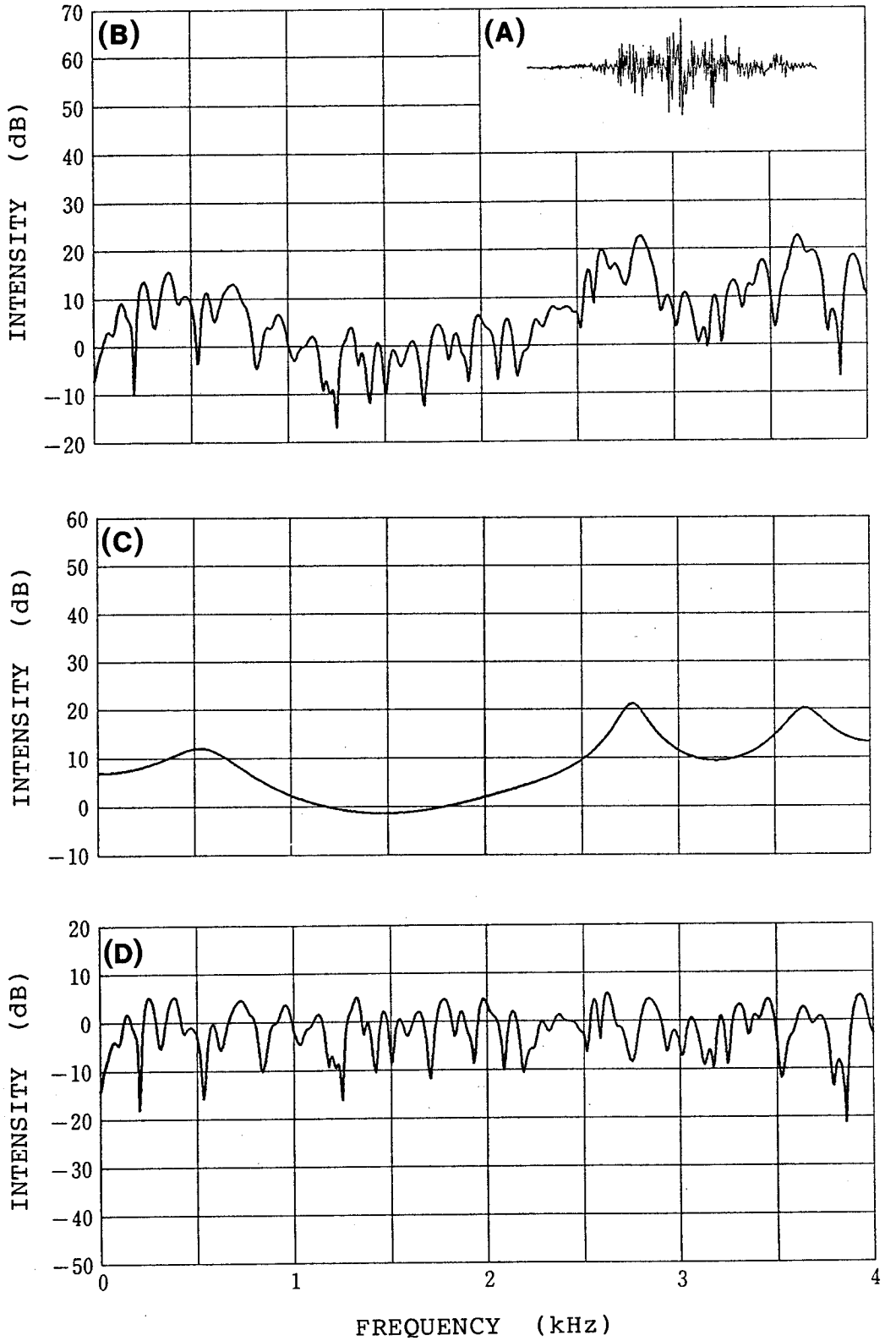


Fig. 1.3 Time and frequency representations of the utterance /tʃ/.

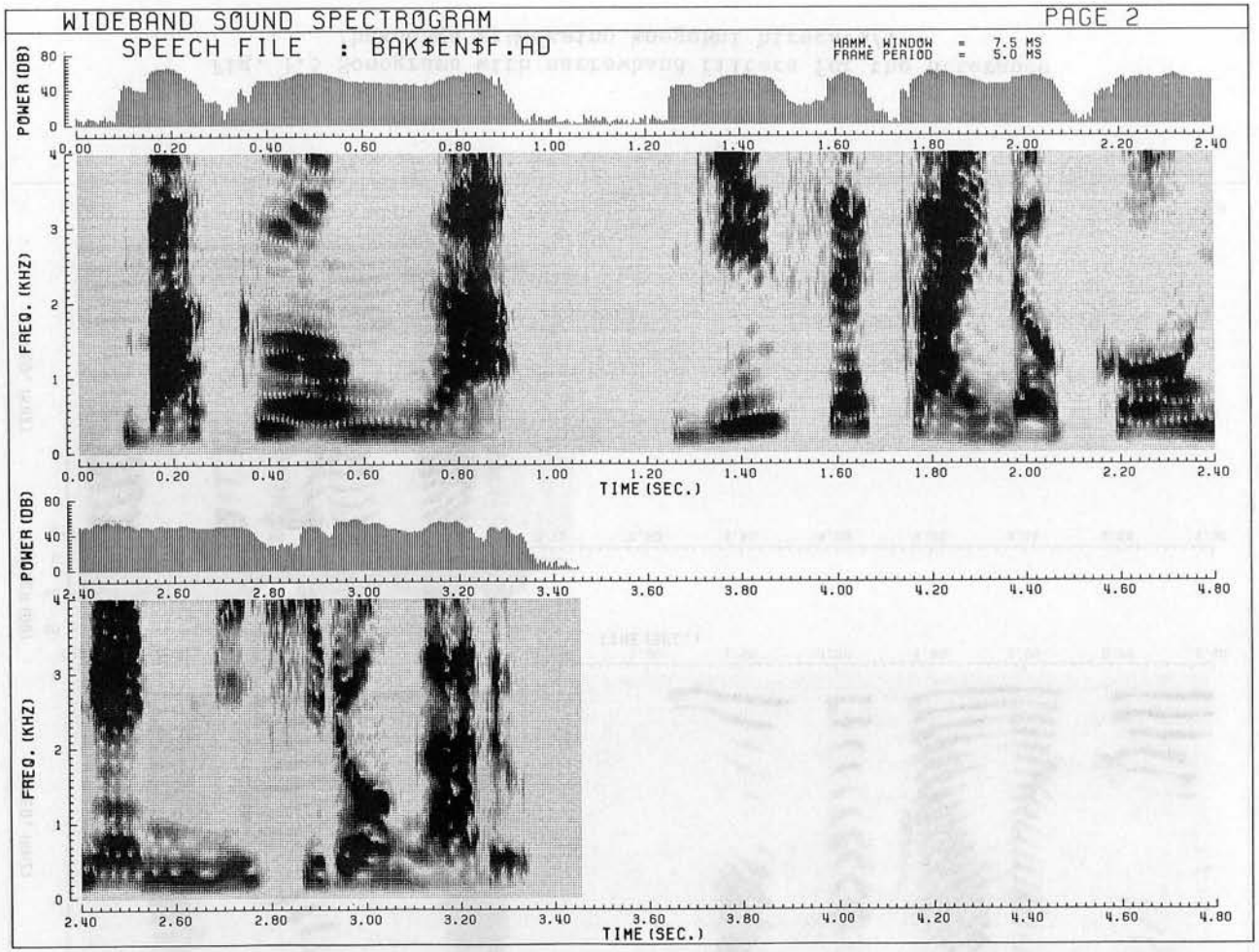


Fig. 1.4 Sonograms with wideband filters for the utterance /bakuoNga giNsekaino koogeNni hirogaru/.

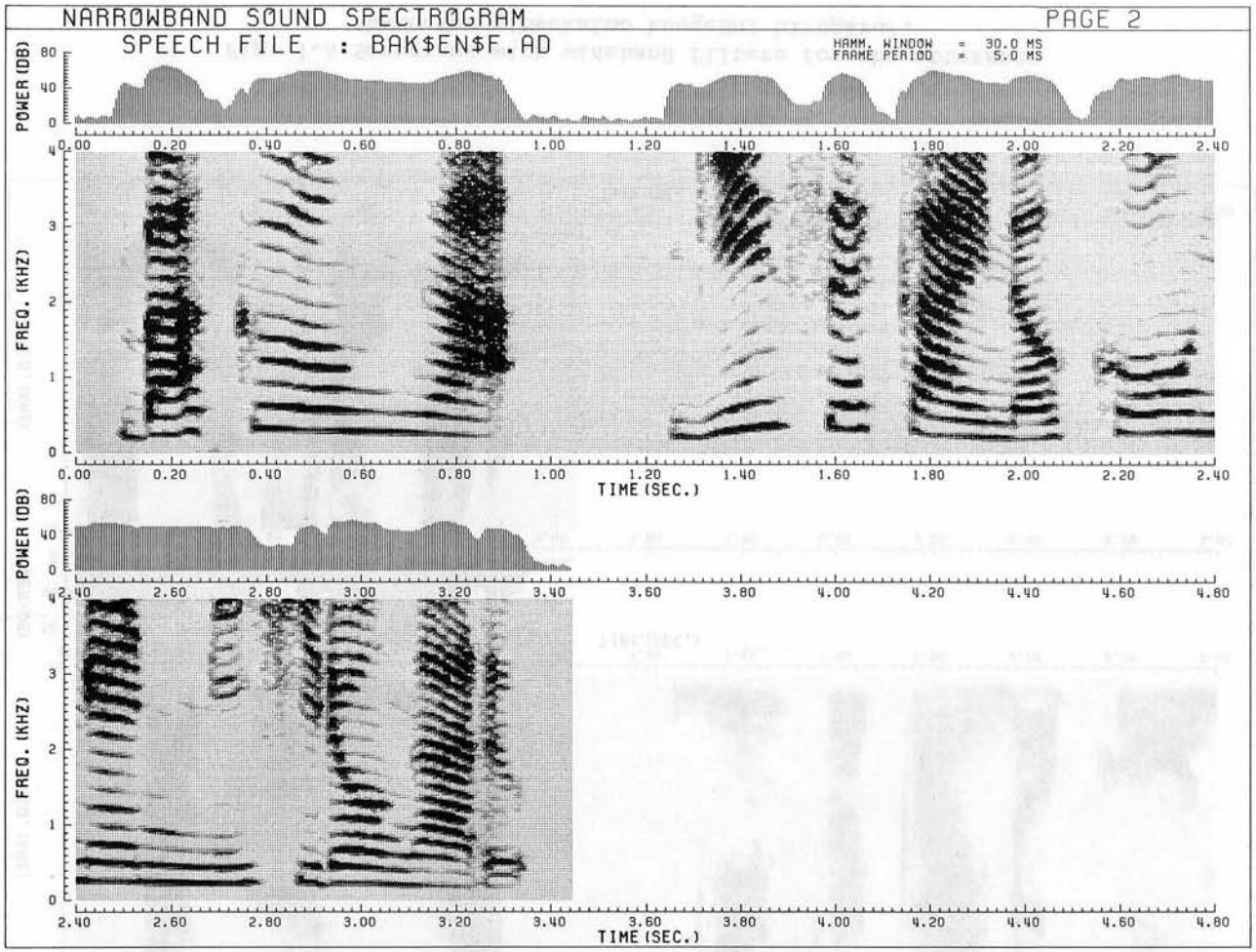


Fig. 1.5 Sonograms with narrowband filters for the utterance /bakuoNga giNsekaino koogeNni hirogaru/.

Table 1.1 Research on speech waveform coding.

CODING TECHNIQUES	PROPOSER(S)	FEATURES	kbps
Pulse Code Modulation (PCM)	A.H.Reeves (1939)	Sampling of speech waveform	56-64
Delta Modulation (Δ M)	F.de.Jager (1952)	Coding of difference between adjacent samples with 1 bit	
Differential Pulse Code Modulation (DPCM)	H.Van de Weg (1953)	Coding of difference between adjacent samples with N bits	
Adaptive Predictive Coding (APC)	B.S.Atal (1968)	Adaptive coding of linear predictive coefficients	9.6-24
Adaptive Delta Modulation (AΔM)	N.S.Jayant (1970)	Δ M with adaptive quantization step size	16-32
Adaptive Differential Pulse code Modulation (ADPCM)	N.S.Jayant (1973)	DPCM with adaptive quantization step size	24-32
Sub-band Coding (SBC)	R.E.Chrochiere (1977)	ADPCM of output in each sub-band channel	16-32
Adaptive Transform Coding (ATC)	R.Zelinski (1977)	Coding of transform coefficients of speech waveform	4.8-16
Adaptive Predictive Coding with adaptive bit allocation (APC-AB)	M.Honda F.Itakura (1979)	Predictive coding with adaptive bit allocation in sub-band coding	9.6-24

Table 1.2 Research on speech analysis synthesis coding (vocoder).

CODING TECHNIQUES	PROPOSER(S)	FEATURE PARAMETERS	kbps
Channel Vocoder	H.Dudley (1939)	Outputs of band-pass filters	300 Hz (analog) 2.4 (digital)
Formant Vocoder	W.A.Munson (1950)	Outputs of band-pass filters and the number of zero-crossing	300 Hz (analog) 2.4 (digital)
Pattern Matching Vocoder	C.P.Smith (1957)	Phoneme spectral patterns	0.9
Correlation Vocoder	M.R.Schroeder (1959)	Auto-correlation coefficients	400 Hz
Voice Exicited Channel Vocoder	M.R.Schroeder (1960)	Outputs of band-pass filters and base band signal	1000 Hz
Phase Vocoder	J.L.Flanagan (1966)	Outputs of band-pass filters and their phase signal	1500 Hz 7.2-9.6
Maximum Likelihood Vocoder	F.Itakura S.Saito (1968)	Linear predictive coefficients by maximum likelihood estimation	5.4
Homomorphic Vocoder	A.V.Oppenheim (1969)	Cepstrum coefficients	7.8
PARCOR Vocoder	F.Itakura S.Saito (1969)	Partial auto-correlation coefficients	2.4-9.6
Linear Prediction Vocoder	B.S.Atal (1971)	Linear predictive coefficients by covariance method	3.6
Residual Excited Linear Predictive Vocoder	D.T.Magill (1974)	PARCOR coefficients by linear prediction	8.0
LSP Vocoder	F.Itakura N.Sugamura (1979)	Line spectrum pair (LSP)	1.6-9.6

Table 1.3 Research on speech analysis synthesis coding at extreme low bit rate.

VOCODER TYPE	PROPOSER(S)	FEATURES	TRANSMISSION RATE (bps)
Phonemic Vocoder	H.Dudley (1958)	4 consonant 6 vowel patterns	100
Formant Vocoder	G.S.Kang (1976)	6 formant parameters	600
LPC Vocoder	A.W.F. Huggins (1977)	Variable frame rate coding	1100-2500
LPC Vocoder	R. Reddy (1977)	Segment coder label coder	165-1800
Formant Vocoder	C.K.Un (1978)	Estimation of bandwidth using amplitude	1200
LPC Vocoder	R.Viswanathan (1977)	Variable frame rate coding	1700
Pattern Matching Vocoder based on LPC	A.Buzo (1979)	Vector quantization	800-900
Pattern Matching Vocoder based on LPC	D.Y.Wong (1981)	Vector quantization with full search or binary search	800
Segment Vocoder	S.Rocus (1983)	Segment clustering segment network	150
Pattern Matching Vocoder based on LPC	D.Y.Wong (1983)	Vector and matrix quantization	200-400

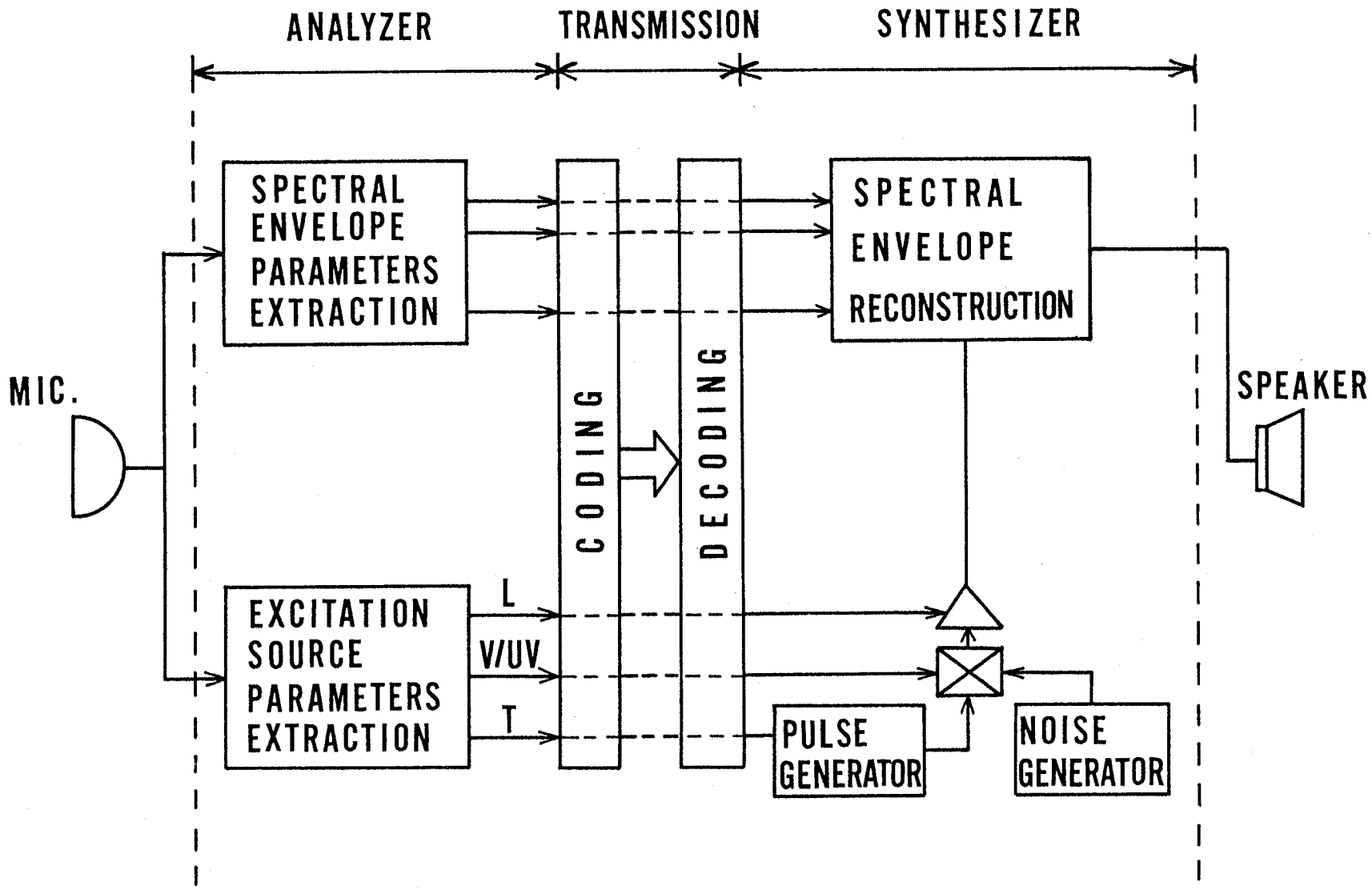


Fig. 1.6 Speech analysis synthesis system organization.

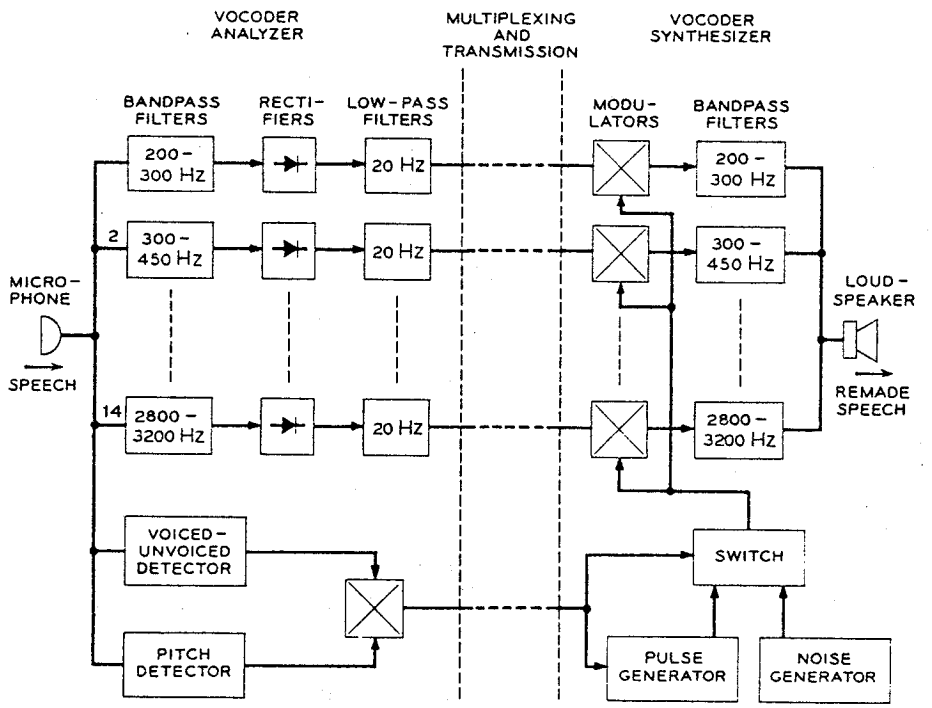


Fig. 1.7 Spectrum channel vocoder blockdiagram. (from reference [22], p.724)

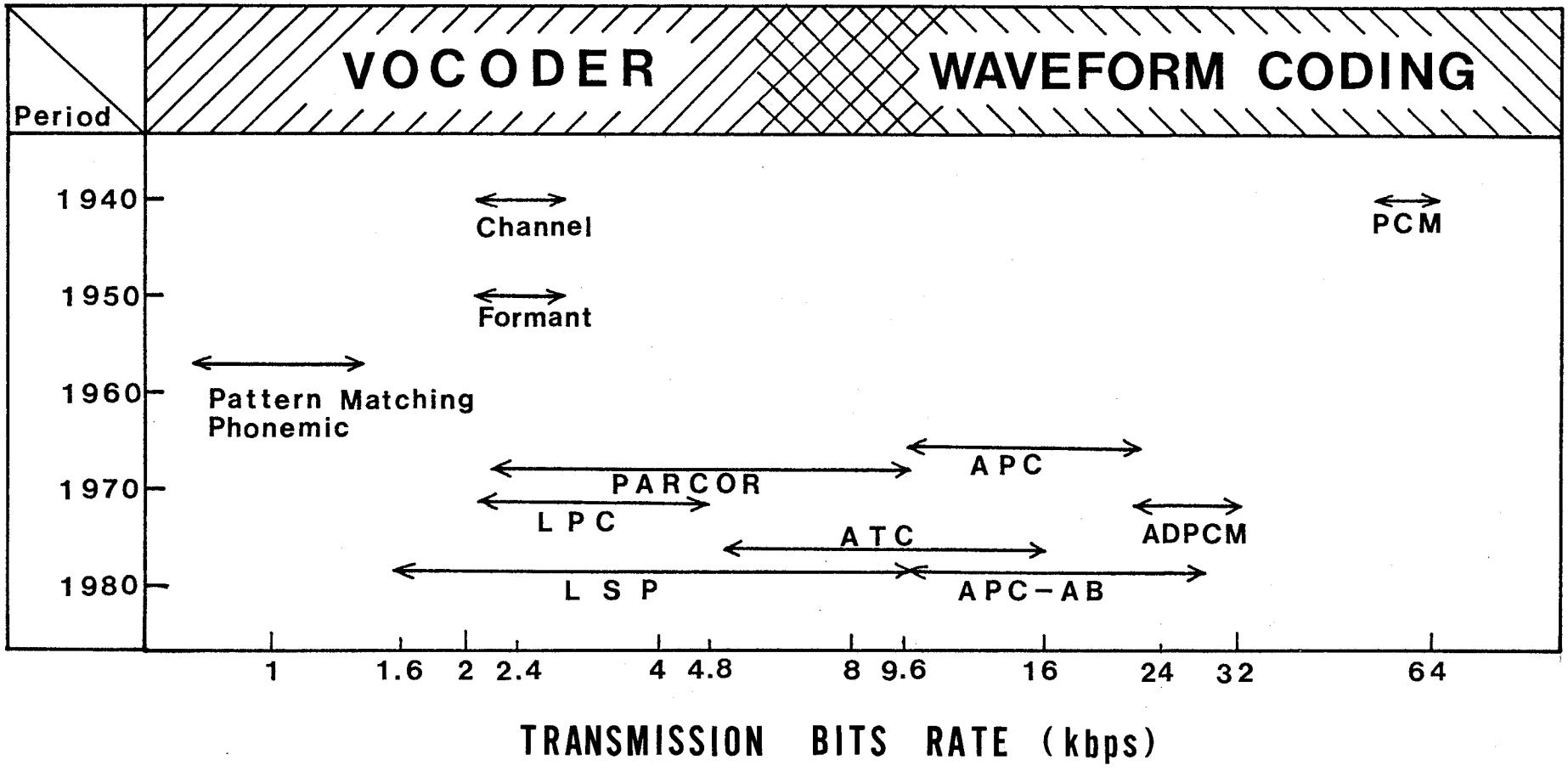


Fig. 1.8 Hierarchy organization of speechy coding.

CHAPTER 2 LINEAR PREDICTION MODEL OF SPEECH WAVEFORM [33],[34]

2.1 Introduction

Various vocoder systems have been studied in transmitting speech spectra efficiently. These vocoder systems are based on the spectral preservation principle. Linear predictive coding (LPC) is one of most successful attempts among those, which was first presented in 1966. The LPC approach itself is not so efficient, because of its quantization characteristics. Some extended ideas resulted from the principle concept that speech signal can be approximately represented by the output signal from all pole digital filter. The PARCOR scheme, proposed in 1969 by F.Itakura and S.Saito [19], is a successful example which inherits the basic idea of all pole modeling. Additionally, it accomplishes significant information reduction through detailed investigation. In this chapter, the PARCOR speech analysis synthesis method is reviewed in order to assure understanding the following chapters easily.

2.2 Speech Production Model Based On Linear Prediction of Speech Waveform

The acoustic speech waveform has a very complex structure, as mentioned in Chapter 1. Ideally, it is desirable to have models that are both linear and time-invariant. Unfortunately, the human speech mechanism does not satisfy either of these properties. Speech is a continually time-varying process. However, by making reasonable assumptions, it is possible to develope linear time-invariant models

over short intervals of time for describing important speech events. The speech production model separates the smoothed envelope structure from an actual spectrum, and attaches a physiological significance to each of the components of the model.

A typical speech production model, based on linear prediction of speech signal, the PARCOR method, is described. PARCOR is a term derived from the words PARtial autoCORrelation coefficients, which are a set of spectral parameters in this model. In the PARCOR speech analysis synthesis system, there are mainly two types of feature parameters. These feature parameters are composed of excitation source and spectral envelope parameters. The source excitation parameters represent vocal cords vibration. They are composed of fundamental frequency, power and voicing (voiced or unvoiced) information. The period according to fundamental frequency is sometimes called "pitch" period. Spectral parameters represent the vocal tract frequency transmission characteristics according to the articulator movements. In the PARCOR system, PARCOR coefficients are spectral envelope feature parameters.

Many papers concerning the PARCOR system, especially optimum coding of PARCOR coefficients, have been published to present ways to improve the synthesized speech quality [35],[36],[37]. The basic idea of the PARCOR system will be introduced in the following explanations, starting with an LPC model of speech signal.

2.2.1 Autocorrelation Method

A time series, obtained by sampling speech signal, shows a significant autocorrelation between adjacent samples. The short time

autocorrelation function is related with the running spectrum, which plays the most important role in speech perception.

Let (x(n), n=----,-2,-1,0,1,2,------) be a discrete time series obtained by sampling the speech waveform every Δ T seconds. Time spacing Δ T should be chosen according to the formula $\Delta T = 1/2f_N$, where f_N is the maximum frequency to be preserved. This frequency range should be chosen to maintain high intelligibility and good quality for human speech signal. In the following discussion, Δ T is fixed at 1/8000 second.

Assume x(n-1), x(n-2),---,x(n-p) are p dimensional random variables taken from a stochastic process, which is stationary within a short interval, for from 10 to 30 mseconds, for example. Letting $\tilde{x}(n)$ be predicted linearly from (x(n-1),---,x(n-p)), then $\tilde{x}(n)$ is represented as

$$\tilde{x}(n) = - \sum_{i=1}^{p} \alpha_i x(n-i) \qquad (2.1)$$

where α_i is defined as a linear predictive coefficient. When x(n) is a real sampled value, the next relation is assumed.

$$x(n) - \tilde{x}(n) = \sigma\, e(n) \qquad (2.2)$$

where e(n) is excitation source signal and σ is its root mean square (RMS) value and e(n) has a signal which has flatness power spectrum and a unit RMS.

Equation (2.2) is converted into the following form easily.

$$x(n) + \sum_{i=1}^{p} \alpha_i x(n-i) = \sigma\, e(n)$$

$$\sum_{i=1}^{p} \alpha_i x(n-i) = \sigma e(n) \qquad (2.3)$$

$$\alpha_0 = 1$$

This equation is interpreted as linear prediction model of speech signal and $\sigma e(n)$ is regarded as prediction residual. An explicit representation of Eq. (2.3) is a direct filter form, shown in Fig. 2.1, where D is unit time delay Δ T. Figure 2.1 shows the speech synthesis filter by α parameters, too. Linear prediction of the speech signal results in an equivalent linear model for speech production.

The transfer function H(z) equivalent to Eq. (2.3) is

$$H(z) = \frac{X(z)}{E(z)} = \frac{\sigma}{1 + \sum_{i=1}^{p} \alpha_i z^i} \qquad (2.4)$$

$A_p(z)$ is defined as

$$A_p(z) = 1 + \sum_{i=1}^{p} \alpha_i z^i \qquad (2.5)$$

A power transfer function of H(z) is

$$S(\omega) = |H(z)|^2/2$$

$$= \sigma^2 / 2\pi \left| 1 + \sum_{i=1}^{p} \alpha_i z^i \right|^2 \qquad (2.6)$$

where $z=e^{-j\omega}$, $(-\pi < \omega < \pi)$

Equation (2.3) is interpreted as the auto-regressive (AR) process and its system function of H(z) is an all pole model in control theory. Namely, $S(\omega)$ is an all pole filter without having zeros. Parameters for the model $A_p(z)$ can be determined directly from the speech waveform by applying a least square criterion to Eq. (2.3).

Linear predictive coefficients, which minimize the mean square values of the predictive residual, are obtained by the following processing.

A mean squared error of residual is

$$\overline{(\sigma\, e(n))^2} = \overline{\left(\sum_{i=1}^{p} \alpha_i\, x(n-i)\right)^2} = \overline{(x(n)+\alpha_1 x(n-1)+ \cdots +\alpha_p x(n-p))^2} \tag{2.7}$$

Simultaneous equations can be obtained by differentiating both sides in Eq. (2.7).

$$\begin{bmatrix} V_0 & V_1 & \cdots & V_{p-1} \\ V_1 & V_0 & \cdots & V_{p-2} \\ \vdots & & \ddots & \vdots \\ V_{p-1} & V_{p-2} & \cdots & V_0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = - \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_p \end{bmatrix} \tag{2.8}$$

where V_i means the short time autocorrelation coefficients of speech waveform and the relation $V_i = V_{-i}$ is assumed.

Equation (2.8) is called regular equation or a Yule-Walker equation.

V_i can be calculated as the following equation.

$$V_i = \sum_{n=0}^{N-1} (x(n)\ x(n+i))/N \qquad (2.9)$$

$$i = 1,2,------,p$$

where N is the number of speech samples.

In Eq. (2.9), x(n) is usually replaced by x(n) weighted using a window function [33].

This solution method is called a "autocorrelation method". Another method, called "covariance method", is proposed. In this method, V_i is defined the following equation, instead of Eq. (2.9)

$$V_{ij} = \sum_{n=0}^{N-1} (x(i+n)\ x(j+n))/N \qquad (2.10)$$

2.2.2 Maximum Likelihood Estimation of Speech Spectrum

Another approach to estimating α parameters in the frequency domain was presented. It is assumed that the speech signal has the following characteristics.

(1) The speech production system can be represented as a transfer function which has only poles.

(2) The speech signal is assumed to be generated with adding a random signal into the system mentioned above. The averaged value of that signal is zero and its variance is assumed to be σ^2 .

Based on these assumptions, parameters, such as $\tilde{\omega}$ $(\sigma^2, \alpha_1, -----, \alpha_p)$, can be estimated from observation sequences X=(x(1),x(2),----,x(N)) in the speech signal.

The spectrum envelope is assumed as

$$\left| H(z) \right|^2 = S(\omega) = \sigma^2/(2\pi \quad \frac{1}{\left| \prod_{i=1}^{p} \left(1 - \frac{z}{z_i}\right) \right|^2} \quad)$$

$$= \sigma^2/(2\pi \quad \frac{1}{\left| \sum_{i=0}^{p} \alpha_i z^{-i} \right|^2} \quad) \qquad (2.11)$$

$$= \frac{\sigma^2}{2\pi \sum_{\tau=-p}^{p} A_\tau \cos \omega\tau}$$

where z_i represents roots of $z^p + \alpha_1 z^{p-1} + \cdots + \alpha_p = 0$ and

$$A_i = \sum_{j=1}^{p-|i|} \alpha_j \alpha_{j+|i|}, \quad i = 0, \pm 1, \pm 2, \cdots, \pm p$$

If the signal x(n) is stationary and e(n) is following a Gaussian distribution, a logarithmic likelihood $L(x \mid \tilde{\omega})$ of observation X=(x(1),x(2),----,x(N)) is approximated by

$$L(x \mid \tilde{\omega}) = \frac{N}{2} \left(\log 2\pi\sigma^2 + \frac{1}{\sigma^2} \sum_{\tau=-p}^{p} A_\tau V_\tau \right)$$

$$= -\frac{N}{2} \left[2 \log 2 + \frac{1}{2\pi} \int_{-\pi}^{\pi} (\log S(\omega) + T(\omega)/S(\omega)) d\omega \right] \qquad (2.12)$$

where

$$V_\tau = \frac{1}{N} \sum_{t=1}^{N-|\tau|} x_t \, x_{t+|\tau|} \tag{2.13}$$

$$T(\omega) = \frac{1}{2\pi} \sum_{\tau=-(N-1)}^{N-1} V_\tau \, e^{-j\omega\tau} \tag{2.14}$$

V_τ means short time autocorrelation function of samples x and $T(\omega)$ is short time power spectrum, which is obtained by discrete Fourier transforming of V_τ . From these equations, when X is given, the logarithmic likelihood ratio is represented by the only function of V_τ . The solution of $\tilde{\omega}$, which maximize $L(x|\tilde{\omega})$, is called maximum likelihood method.

$L(x|\tilde{\omega})$ is maximized with σ^2 , namely executing $\partial L(x|\tilde{\omega})/\partial\sigma^2 = 0$,

$$\sigma^2 = J(\alpha_1, \alpha_2, \cdots, \alpha_p) = \sum_{\tau=-p}^{p} A_\tau V_\tau \tag{2.15}$$

where

$$L(x|\tilde{\omega})\Big|_{\sigma^2=J} = -\frac{1}{2N}\left[\log 2\pi e \, J(\alpha_1, \alpha_2, \cdots, \alpha_p)\right]$$

Consequently, the maximization of $L(x|\tilde{\omega})$ is equivalent to the minimization of $J(\alpha_1, \alpha_2, \cdots, \alpha_p)$.

On the other hand

$$\begin{aligned} J(\alpha_1, \alpha_2, \cdots, \alpha_p) &= \sum_{\tau=-p}^{p} A_\tau V_\tau \\ &= \sum_{\tau=-p}^{p} \left(\sum_{j=1}^{p-|\tau|} \alpha_j \, \alpha_{j+|\tau|} \right) V_\tau \end{aligned} \tag{2.16}$$

From $\partial J(\alpha_1, \alpha_2, \cdots, \alpha_p)/\partial\alpha = 0$

$$\sum_{j=0}^{p} V_{j-i}\ \alpha_j = 0 \qquad (2.17)$$

$$i = 1, 2, \text{-----}p, \quad \alpha_0 = 1$$

This equation is the same as Eq. (2.8), which is derived by linear predictive analysis with minimizing the square of prediction residual.

Namely, the linear predictive model of speech signal and maximum likelihood estimation of speech spectrum are equivalent, as long as the speech signal is assumed to have all pole spectrum. Figure 2.1 shows a short time spectrum $T(\omega)$ and an estimated spectrum $S(\omega)$ using the all pole model. Figures 2.1(a) and (b) show acoustic speech waveform versus time and its autocorrelation coefficients normalized by V_0 .

2.3 Speech Analysis and Synthesis Using PARCOR Coefficients

2.3.1 Partial Autocorrelation Method

When a speech synthesis filter is based on α parameters, high accuracy is required to quantize α parameters in order to maintain the filter's stability. To solve this problem, a PARCOR lattice filter was invented to achieve the same transfer function. The autocovariance or the autocorrelation coefficients can be regarded as a measure of linear dependency for time shift, but a set of these parameters is still redundant, because there is significant dependency among them. The notion of partial autocorrelation is introduced to reduce the redundancy using linear prediction techniques.

$\tilde{x}(n)$ and $\tilde{x}(n-p)$ can be predicted from the same samples by

$$\tilde{x}(n) = -\sum_{i=1}^{p-1} \alpha_i^{(p-1)} x(n-i) \tag{2.18}$$

$$\tilde{x}(n-p) = -\sum_{i=1}^{p-1} \beta_i^{(p-1)} x(n-i) \tag{2.19}$$

where the prediction coefficients $\alpha_i^{(p-1)}$ and $\beta_i^{(p-1)}$ are decided to minimize the residual of the least squares forward prediction $E(x(n) - \tilde{x}(n))^2$ and that of backward prediction $E(x(n-p) - \tilde{x}(n-p))^2$.

A PARCOR coefficient k_p, between x(n) and x(n-p) is defined as the cross correlation coefficient between two residuals, as shown in Fig. 2.3.

$$f^{(p-1)}(n) = \sum_{i=0}^{p-1} \alpha_i^{(p-1)} x(n-i) \tag{2.20}$$

$$g^{(p-1)}(n) = \sum_{i=1}^{p} \beta_i^{(p-1)} x(n-i) \tag{2.21}$$

$$k_p = \frac{E\left[\left(f^{(p-1)}(n)\right)\left(g^{(p-1)}(n)\right)\right]}{\left[E\left\{\left(f^{(p-1)}(n)\right)^2\right\} E\left\{\left(g^{(p-1)}(n)\right)^2\right\}\right]^{1/2}} \tag{2.22}$$

These parameters physically correspond to reflection coefficients in the acoustical tube model of the vocal tract [2]. Thus, PARCOR coefficients are sometimes called reflection coefficients, too.

The following relation is obtained by substituting Eqs. (2.20) and (2.21) into Eq. (2.22).

$$k_m = \frac{\sum_{i=0}^{m-1} \alpha_i^{(m-1)} V_{m-i}}{\sum_{i=0}^{m-1} \alpha_i^{(m-1)} V_i}$$

$$= \frac{w_{m-1}}{u_{m-1}} \qquad (2.23)$$

$$m = 1, 2, \text{----}, p$$

On the other hand, prediction coefficients $\left\{\alpha_i^{(P-1)}\right\}_{i=1}^{P}$ and $\left\{\beta_i^{(P-1)}\right\}_{i=1}^{P}$ satisfy

$$\sum_{i=0}^{P-1} \alpha_i^{(P-1)} V_{i-j} = 0 , \quad \alpha_0^{(P-1)} = 1 \quad (j= 1, 2, \text{--} ,p-1) \qquad (2.24)$$

$$\sum_{i=1}^{P} \beta_i^{(P-1)} V_{i-j} = 0 , \quad \beta_P^{(P-1)} = 1 \quad (j= 1, 2, \text{--} ,p-1) \qquad (2.25)$$

From Eqs. (2.23) (2.24) and (2.25), the relations

$$\alpha_i^{(m)} = \alpha_i^{(m-1)} - k_m \beta_i^{(m-1)}, \quad \alpha_m^{(m-1)} = 0, \quad (i= 1,2, \text{--} ,m) \qquad (2.26)$$

$$\beta_i^{(m)} = \beta_{i-1}^{(m-1)} - k_m \alpha_{i-1}^{(m-1)}, \quad \beta_0^{(m-1)} = 0, \quad (i= 1,2, \text{--} ,m) \qquad (2.27)$$

$$u_m = u_{m-1} (1 - k_m^2) \qquad (2.28)$$

are obtained.

From Eqs. (2.24) and (2.25)

$$\beta_i^{(m-1)} = \alpha_{m-i}^{(m-1)} \tag{2.29}$$

is also obtained.

2.3.2 Direct PARCOR Coefficients Derivation

PARCOR coefficients are also derived directly from the speech signal. It is called a "lattice PARCOR analysis method"

PARCOR coefficients have already been defined as the crosscorrelation coefficient between the residual of the forward prediction and that of the backward prediction. Two linear operators, forward and backward residual operators, are introduced.

$$\begin{aligned} x(n) - \tilde{x}(n) &= f^{(P-1)}(n) \\ &= \sum_{i=0}^{P-1} \alpha_i^{(P-1)} x(n-i) \\ &= \left[\sum_{i=0}^{P-1} \alpha_i^{(P-1)} D^i \right] x(n) \\ &= A_{P-1}(D)\, x(n) \end{aligned} \tag{2.30}$$

$$\begin{aligned} x(n) - \tilde{x}(n-p) &= g^{(P-1)}(n) \\ &= \sum_{i=1}^{P} \beta_i^{(P-1)} x(n-i) \\ &= \left[\sum_{i=1}^{P} \beta_i^{(P-1)} D^i \right] x(n) \end{aligned}$$

$$= B_{P-1}(D)\, x(n) \qquad (2.31)$$

where D is the shift operator for unit time, $D^i x(n)=x(n-i)$, $\alpha_i^{(P-1)}$ and $\beta_i^{(P-1)}$ are weighting coefficients for prediction.

From Eqs. (2.26) and (2.27), the next recursions are obtained.

$$A_m(D) = A_{m-1}(D) - k_m B_{m-1}(D) \qquad A_0(D)=1 \qquad (2.32)$$

$$B_m(D) = D\,(B_{m-1}(D) - k_m A_{m-1}(D)) \qquad B_0(D)=D \qquad (2.33)$$

Then, it is found that residual operators $A_m(D)$ and $B_m(D)$ satisfy recursive relations (2.32) and (2.33). Noting initial conditions $A_0(D)=1$ and $B_0(D)=D$, $A_m(D)$ and $B_m(D)$ are recursively composed of lattice type digital filters, whose k_m coefficients are controlled by the output of the short time crosscorrelation shown in Fig. 2.4.

2.3.3 Extraction of Source Parameters Using Modified Autocorrelation Method

As the input signal passes through the PARCOR analyzer, autocorrelations between adjacent samples are gradually removed. If the number of sections, p, is chosen sufficiently large, the spectral envelope features of the input signal are extracted almost completely, and the spectrum envelope of the residual is nearly flattened. Thus, only characteristics related to the excitation source, such as signal amplitude, voicing and pitch period, are contained in it. The signal amplitude is the root mean square value of the residual.

In order to detect periodicity and to determine pitch period, autocorrelation coefficients for the residual are computed. Lag time T, indicating the maximum autocorrelation, is picked up.

Autocorrelation coefficients for the residual $f^{(P)}(n)$ are obtained by two methods. One is the signal $f^{(P)}(n)$ to input the autocorrelator directly. The other is to remove the spectral envelope component in the speech signal from the short-time autocorelation coefficients. The term R (m), the autocorrelation coefficient of $f^{(P)}(n)$, is expressed as Eq. (2.34).

$$R(m) = \frac{E\left[f^{(P)}(n)\, f^{(P)}(n+m) \right]}{E\left[\left(f^{(P)}(n) \right)^2 \right]} \tag{2.34}$$

$$\begin{aligned} &= \frac{1}{\mu_P} E\left[\sum_{i=0}^{P} \alpha_i^{(P)} x(n-i) \sum_{j=0}^{P} \alpha_j^{(P)} x(n+m-j) \right] \\ &= \frac{1}{\mu_P} \sum_{i=-P}^{P} A_i^{(P)} V_{i+m} \end{aligned} \tag{2.35}$$

where

$$A_i^{(P)} = \sum_{j=0}^{P-|i|} \alpha_j^{(P)} \alpha_{j+|i|}^{(P)} \quad i = 0, \pm 1, \text{----}, \pm p \tag{2.36}$$

$$\mu_i = E\left[f^{(P)}(n)^2 \right] = \sum_{i=0}^{P} \alpha_i^{(P)} V_i$$

That is, it is shown, from Eqs. (2.34) and (2.35), that the residual autocorrelation coefficient, R (m), is calculated by using a moving average technique, which is weighted to autocovariance V_i of speech signal x(n) by weighting function $\left[A_i^{(P)} \right]_{i=-P}^{P}$. Vocal cords vibration periodicity is detectable by judging the maximum

autocorrelation coefficients value. Periodicity is used to determine voiced parameter and unvoiced parameter, too.

If the maximum value is greater than a preset threshold, the input speech may be judged to be voiced sound with pitch period P. If not, it is considered to be unvoiced. A preset threshold was determined experimentally. This excitation source parameter extraction method is called a "modified autocorreation method", which has good accuracy.

2.3.4 Speech Synthesis from PARCOR Coefficients

Speech synthesis, from PARCOR coefficients and excitation source parameters, is an inverse process of speech analysis. The excitation source is generated by controlling the impulse generator and the white noise generator by pitch period, voicing and amplitude. The resulting signal with flat spectral envelope excites a time-varying filter, composed of lattice sections.

The filter transfer function is exactly inverse to the residual operator $A_m(D)$, which is shown in Fig. 2.5. If $1/A_i(D)$ and $B_i(D)$, which are the filter transfer functions, have already been realized, output signal y(n) is obtained by adding signal $A_m(D)\ y(n)$ to terminal a_i. Then, signal $B_m(D)\ y(n)$ is obtained at terminal b_i. Using recursive Eqs.(2.32) and (2.33), the transfer function is recursively constructed by cascaded lattice circuit connection in the same figure. Therefore, the output signal from the filter has nearly the same spectrum as the input speech signal. Figure 2.6 is derived by equivalent transformation of Fig.2.5.

It is proved that these PARCOR synthesis filters are always stable, as long as each k parameter is less than one [36].

2.3.5 PARCOR Speech Analysis Synthesis System Construction

A PARCOR speech analysis synthesis system blockdiagram is shown in Fig.2.7. It is composed of the speech analyzer, the feature parameter transmission and the speech synthesizer. PARCOR coefficients extracted from the analyzer are related to the spectral envelope parameters, and signal amplitude, voiced/unvoiced parameter V/UV and pitch period P related to the excitation source parameters. The total number of parameters is (p+3). Parameters are digitally coded and transmitted. Then, received parameters are decoded and the nearly original speech is reconstructed by controlling the speech synthesizer using transmitted parameters.

2.3.6 Impairment Factors in Synthesized Speech Quality

The synthesized speech quality impairment is caused by coding the parameters. Impairment factors in synthesized speech are p, which is the number of PARCOR coefficients, N and N_S, which are the numbers of quantization bits of PARCOR coefficients and excitation source parameters per frame, and T_f, which is the sampling period of the parameters. The number of PARCOR coefficients (p) determines the speech spectral envelope accuracy. Much spectral distortion is generated in the case of a fewer number of p, because details are not reconstructed. Fewer PARCOR coefficient quantization bits generate much irregular distortion in the spectrum. Moreover, fewer excitation source parameter quantization bits generate much speech wave irregular fluctuation. On the other hand, the frame period determines time resolution efficiency.

The longer frame period generates speech transitional characteristics that are not reconstructed accurately.

The relation between information rate and synthesized speech quality is supplementary. That is, a lower information rate generates much synthesized speech impairment. Thus, it is necessary to clarify the relation between information rate and transmission quality. Of course, optimum bit allocation is needed, when total information rate is given.

2.3.7 PARCOR Coefficients Characteristics

The distribution differs in accordance with the difference in coefficient degree. That is, k_1 is concentrated to nearly +1 and k_2 is inversely concentrated to nearly -1. Moreover, higher degree coefficients are distributed about zero. PARCOR coefficients are autocorrelation coefficients after removing the autocorrelations between adjacent samples, so that the lower degree PARCOR coefficients are more important. A study of small fluctuation in the values of PARCOR coefficients influencing the synthesized speech spectrum shows that lower degree coefficients are much influenced, but higher degree coefficients are little influenced. It also shows that the lower degree coefficients have influence on the low frequency component of the spectrum, but higher degree coefficients have influence on the flat frequency of the spectrum. When analysis frame period is long, PARCOR coefficients must be interpolated. However, interpolated values sometimes generate irregular spectra. These spectra influence the synthesized speech quality very adversely. These properties must be considered in coding PARCOR coefficients.

To avoid these characteristics, much effort has been expended in studying encoding PARCOR coefficients. Many quantization techniques, including non-uniform bit allocation and non-linear transform quantization of PARCOR coefficients, were proposed to improve the synthesized speech quality. Synthesized speech quality was verified from objective and subjective measurements. Several optimum coding techniques have been established to minimize the spectral distortion or the synthesized speech impairment with lower bit rate through many, many experiments [38],[39],[40].

2.4 Conclusions

(1) A speech production model, based on linear prediction, is explained. One of the most typical speech analysis synthesis methods, PARCOR, is introduced to assure understanding this thesis easily. The methods for extracting PARCOR coefficients and excitation source parameters were reviewed as an introduction to this thesis.

(2) Several problems are clarified in order to represent a speech signal at a low bit rate, using the PARCOR speech analysis synthesis method.

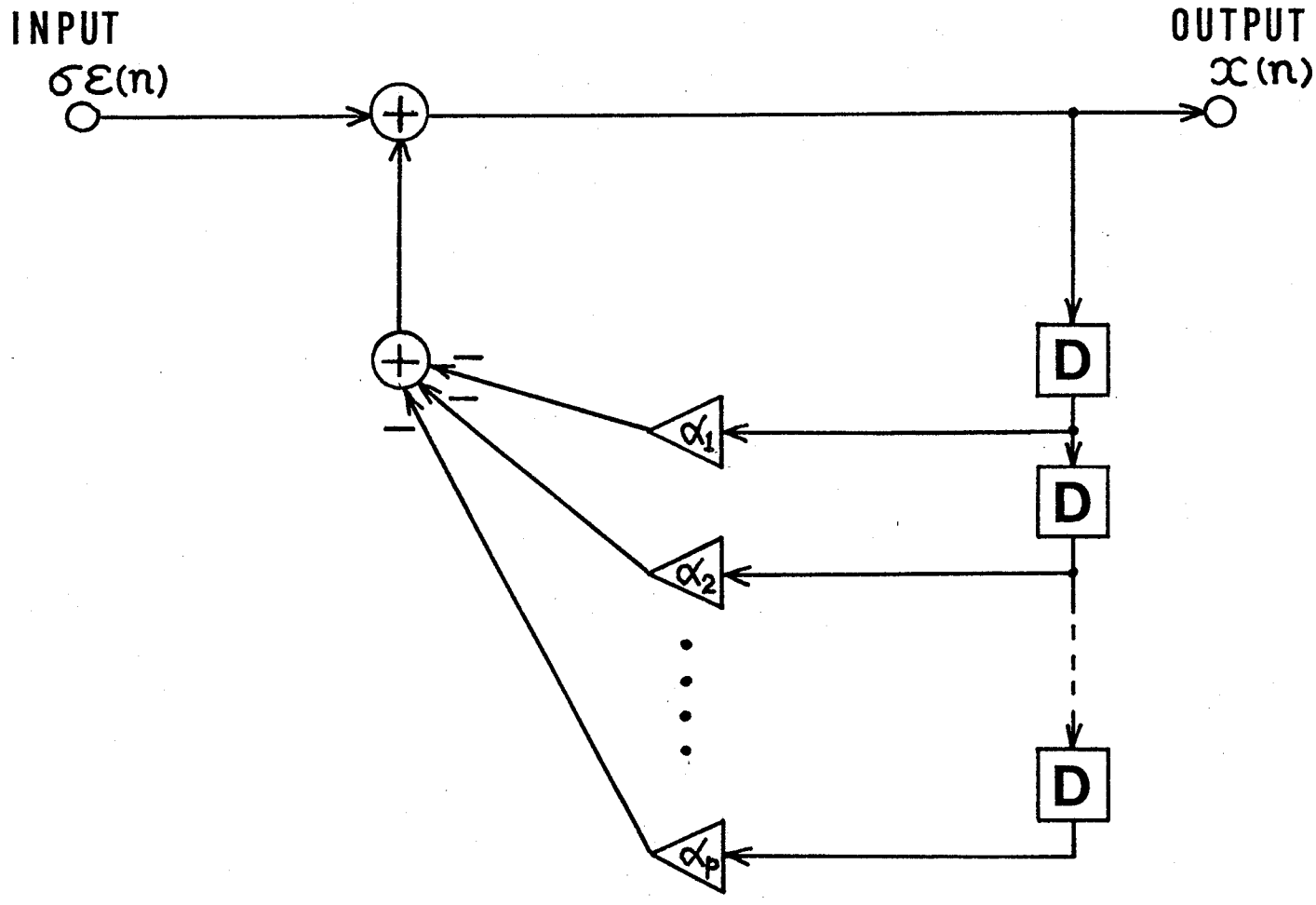
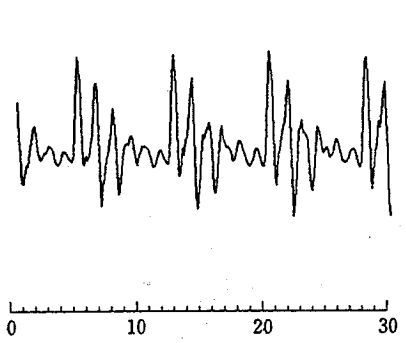
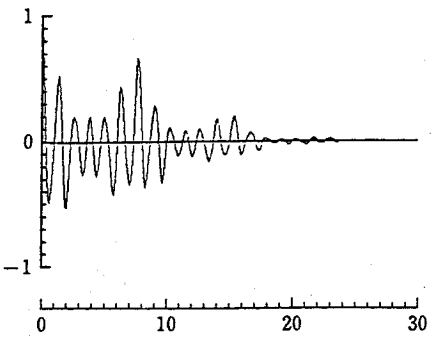


Fig. 2.1 Speech synthesis filter by linear predictive coefficients.

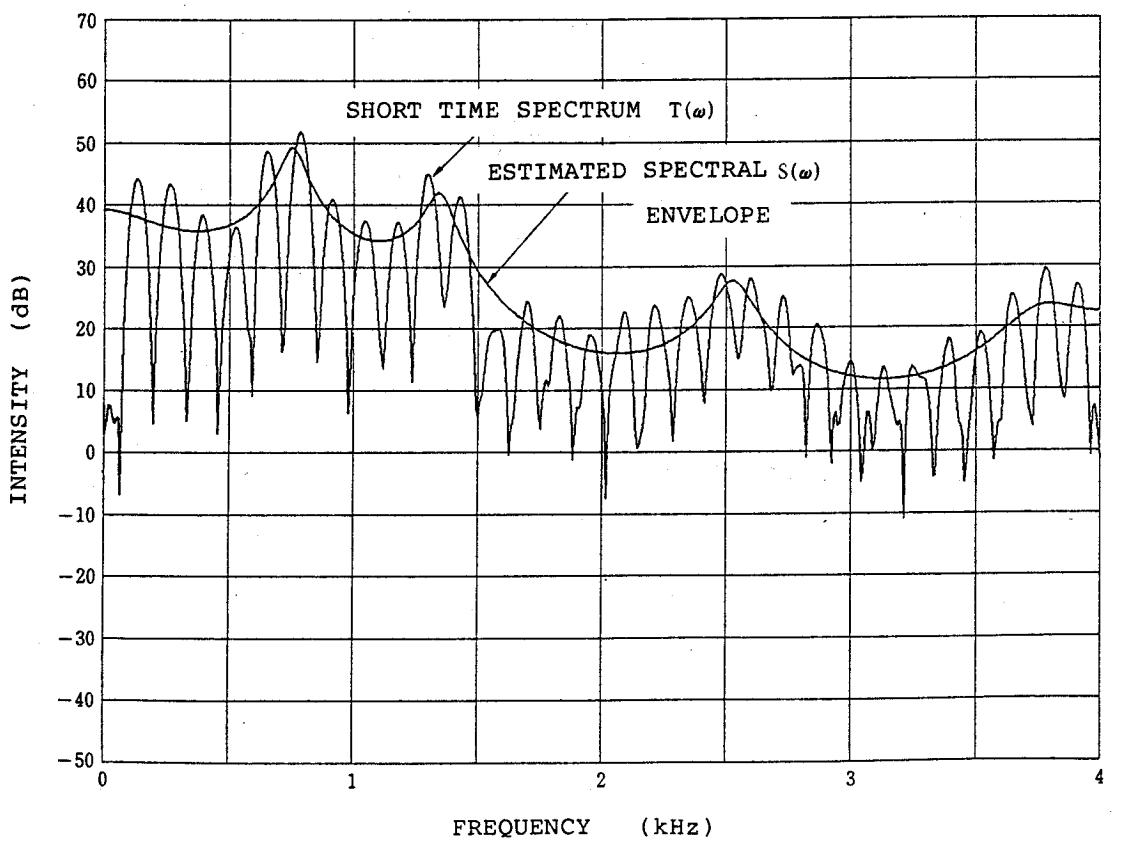


(a) SPEECH WAVEFORM X(n)



(b) AUTO-CORRELATION FUNCTION V_m

NORMALIZED V_0



(c) SHORT TIME SPECTRUM AND ESTIMATED SPECTRUM ENVELOPE WITH ALL-POLE FILTER

Fig. 2.2 Short time spectrum T(w) and estimated spectrum S(w) based on all pole model. This is a sample of the /a/ sound uttered by a male speaker.

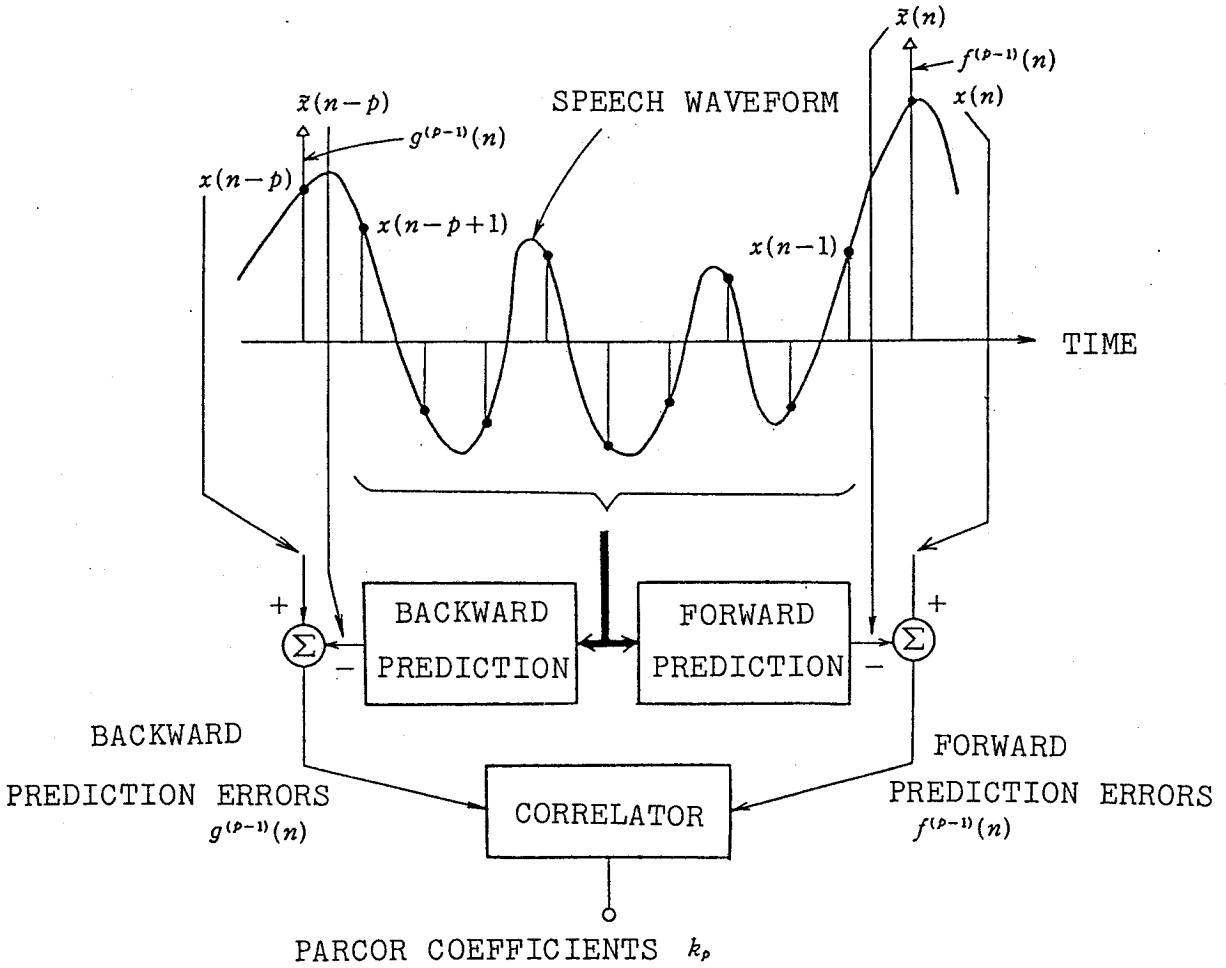


Fig. 2.3 Illustration showing samples used in both forward and backward predictions.

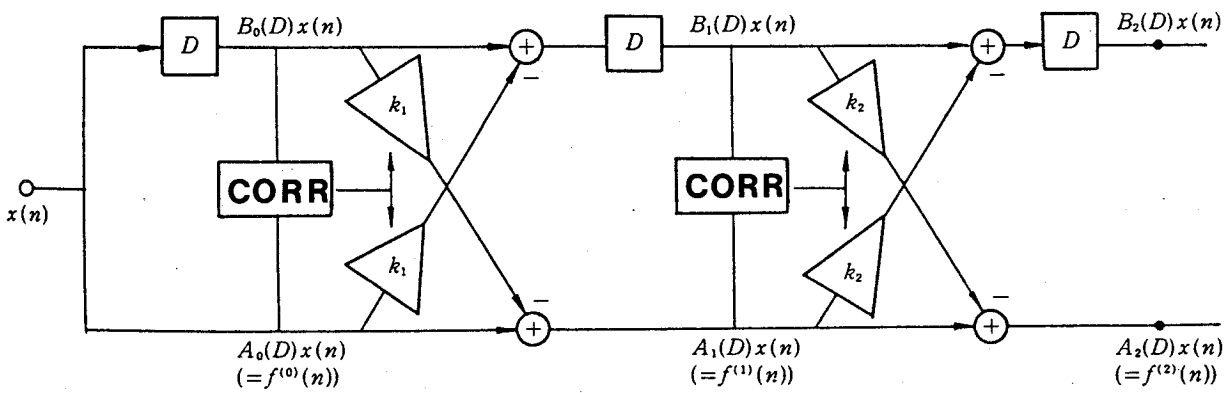


Fig. 2.4 Recursive construction of lattice type digital filter for PARCOR analysis.

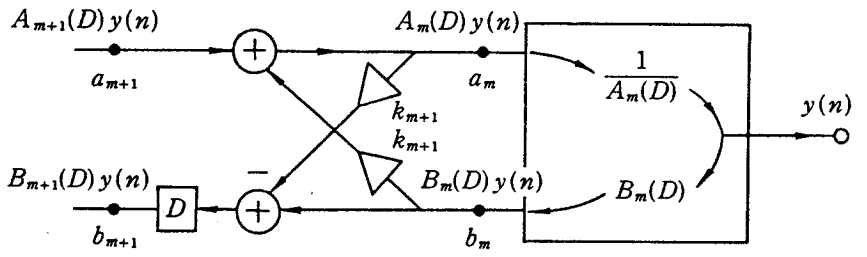


Fig. 2.5 Synthetic circuit transfer function recursive construction.

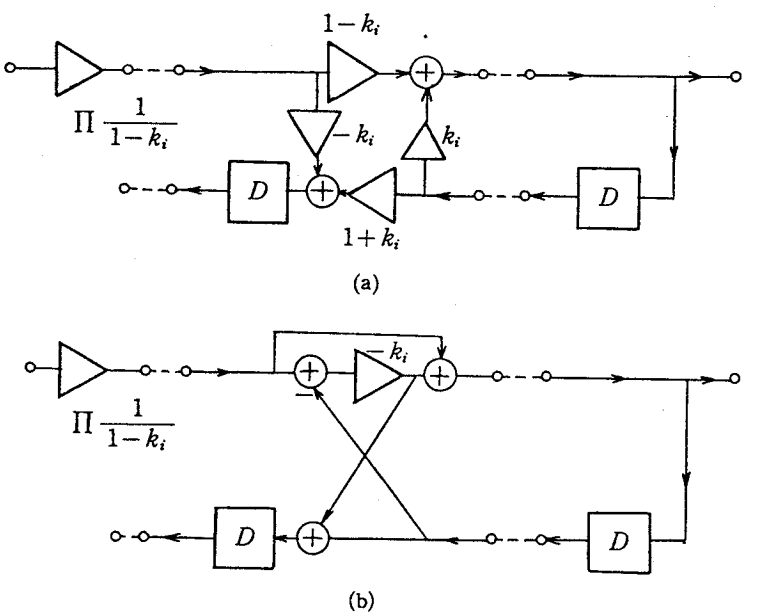


Fig. 2.6 Equivalent circuits for synthetic digital filter.

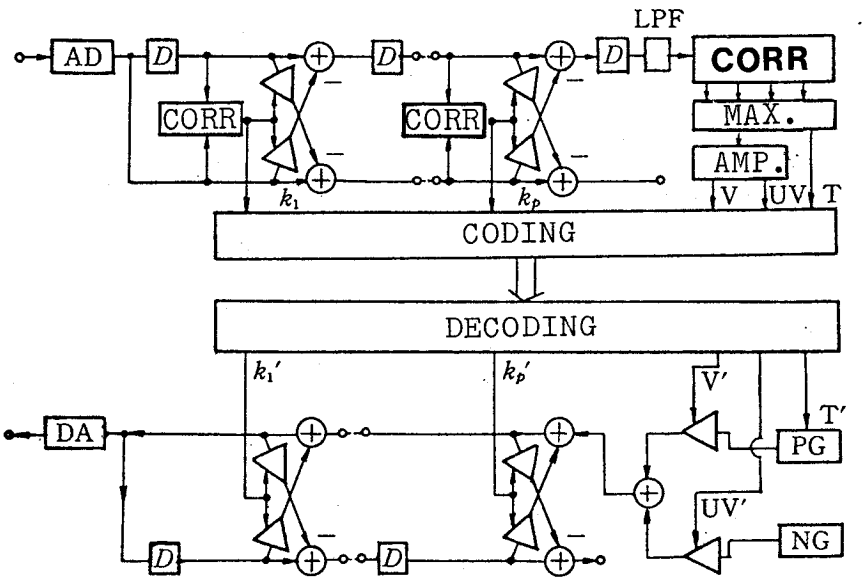


Fig. 2.7 PARCOR analysis synthesis system

CHAPTER 3 LINE SPECTRUM REPRESENTATION OF LINEAR PREDICTIVE COEFFICIENTS

3.1 Introduction

This chapter presents a new expression of spectral feature parameters, which are equivalent to α parameters or PARCOR coefficients. These parameters are defined as a set of line spectrum frequency and its amplitude. They are similar to formant parameters. The method of extracting these parameters from PARCOR coefficients is related. The properties of these parameters, in quantization and interpolation, are examined experimentally, comparing with PARCOR coefficients.

3.2 Line Spectrum Analysis [41],[42]

In the previous chapter, a typical LPC system, PARCOR vocoder, was reviewed. PARCOR coefficients are superior to α parameters in quantization property. Furthermore, a PARCOR synthesis filter is stable, as long as k_i is less than one for all i components [43]. PARCOR coefficients are one of the expressions in the time domain. On the other hand, historically, formant frequency and its bandwidth are well known to represent speech spectra with very few bits of information. These parameters are better than the others, because physical meanings of parameters are very clear. However, it is very difficult to extract true formant parameters. Thus, it is almost impossible to extract them fully automatically. On the other hand, if it is assumed that there is no loss at the glottis in the pseudo-vocal

tract model, determined by PARCOR coefficients, a transfer function of the vocal tract is represented as a line spectrum. These processes are executed fully automatically, without using parameters' dynamics and stochastic properties. Line spectrum parameters are similar to formant parameters in the frequency domain, but they are not exact formant parameters, as mentioned later.

3.2.1 Pseudo Vocal Tract Model by PARCOR Coefficients

In the speech analysis synthesis method by LPC parameters, spectral envelope is approximately represented as an all pole model, as mentioned in Chapter 2. The all pole transfer function is constructed as a lattice filter using PARCOR coefficients, as shown in Fig. 2.5. This lattice filter corresponds to a pseudo vocal tract, which is constructed by the cascade connection of lossless tubes with different cross sections [2]. Input termination corresponds to the glottis and the output corresponds to the lips, as shown in Fig. 3.1. PARCOR coefficients at a discontinuous point in the vocal tract correspond to a cross section, A_n, according to the following equation.

$$k_n = \frac{A_n - A_{n-1}}{A_n + A_{n-1}} \tag{3.1}$$

where A_n means the cross section of the n-th acoustic tube from the lips.

Therefore PARCOR coefficients are sometimes called reflection coefficients.

The transfer characteristics of the vocal tract depend on only the boundary conditions at the glottis and the lips. Boundary conditions at the glottis and the lips are generally assumed to be :

(1) A complete opening at the lips.

(2) A matched impedance termination at the glottis.

By these assumptions, these is no loss in the pseudo vocal tract, except the loss at the glottis. This lossless system can be examined using S matrix representation. In a lossless linear passive system, as shown in Fig. 3.2, the amplitudes of incidence wave and reflection wave at both terminals, 1 and 2, are represented as (a_1, b_1) and (a_2, b_2), respectively.

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{3.2}$$

If it is a lossless system, S matrix becomes a Unitary matrix and the following relations are valid.

$$\begin{aligned} |S_{11}|^2 + |S_{21}|^2 &= 1 & |S_{12}|^2 + |S_{22}|^2 &= 1 \\ \overline{S_{11}}\, S_{12} + \overline{S_{21}}\, S_{22} &= 0 & |S_{12}| &= |S_{21}| \end{aligned} \tag{3.3}$$

where $\overline{S_{ij}}$ represent the conjugate of S_{ij} .

Terminals 1 and 2 are regarded as the glottis and the lips, respectively. When signal b_2 is reflected completely and fedback to a_2 at the lips, the transfer function from the glottis to the lips and its power transfer function are

$$T_{21} = \frac{S_{21}}{1 - S_{22}} \tag{3.4}$$

$$|T_{21}|^2 = \frac{|S_{21}|^2}{1 + |S_{22}|^2 - (S_{22} + \overline{S_{22}})} \tag{3.5}$$

On the other hand, the immitance at the lips and its real part are,

$$R = \frac{1 + S_{22}}{1 - S_{22}} \tag{3.6}$$

$$Re(R) = \frac{1 - |S_{22}|^2}{1 + |S_{22}|^2 - (S_{22} + \overline{S_{22}})} \tag{3.7}$$

From Eqs. (3.5) and (3.7),

$$Re(R) = |T_{21}|^2 \tag{3.8}$$

is obtained.

This equation shows the power transfer function $|T_{21}|^2$ is equal to the real part of immitance function R at the lips. Furthermore, immitance function R can be determined by only an impulse response function, S_{22}, at the lips.

Consequently, the power transfer function is determined by only an impulse response function at the lips.

3.2.2 Extraction of the Immitance Function at the Lips from PARCOR coefficients

An equivalent circuit for an acoustic tube junction J_n is represented as shown in Fig.3.3 [2]. The S matrix in this figure is represented by

$$S = \begin{bmatrix} -k_n & \sqrt{1-k_n^2}\, z^{1/2} \\ \sqrt{1-k_n^2}\, z^{1/2} & k_n z \end{bmatrix} \tag{3.9}$$

Immitance function R is obtained by calculating the response at terminal A, when a unit impulse is given to the vocal tract from terminal B.

When a black box, whose impulse response is s, is connected to the left side of J_n , the impulse response $w_n(z,s)$ at the right side of J_n is

$$w_n(z,s) = k_n + \frac{(1-k_n^2)s}{(1+k_n s)} z \tag{3.10}$$

$$= \frac{k_n + s}{1 + k_n s} z$$

Similarly, J_n ,J_{n-1} ,------,J_1 are cascade connected in the same manner. When J_1 is terminated by the black box, whose impulse response is s, the impulse response at the right side of J_n is represented as $W_n(z,s)$. For example, when n=1 and n=2, W_1 and W_2 are

$$W_1(z,s) = \frac{k_1 + s}{1 + k_1 s} z \tag{3.11}$$

$$W_2(z,s) = \frac{k_1 + k_2 z + (k_1 K_2 + z)s}{(1+ k_1 k_2)z + (k_1 z + k_2)s} \tag{3.12}$$

A numerator and a denominator in Eq. (3.12) are represented as $C_n(z)$, $D_n(z)$, $E_n(z)$ and $F_n(z)$. These terms are represented as C_n ,D_n ,E_n and F_n for simplicity here.

$$W_n(z,s)/z = \frac{E_n - F_n s}{C_n - D_n s}$$

$$= \frac{E_{n-1} - F_{n-1} W_n(z,s)}{C_{n-1} - D_{n-1} W_n(z,s)} \tag{3.13}$$

Substituting Eq. (3.10) into Eq. (3.13),

$$W_n(z,s)/z = \frac{(E_{n-1} - k_n zF_{n-1}) - (zF_{n-1} - k_n E_{n-1})s}{(C_{n-1} - k_n zD_{n-1}) - (zD_{n-1} - k_n C_{n-1})s} \qquad (3.14)$$

Equations (3.13) and (3.14), C_n ,D_n ,E_n and F_n follow the next relations.

$$\begin{aligned} C_n &= C_{n-1} - k_n zD_{n-1} \quad C_1 = 1 \\ D_n &= zD_{n-1} - k_n C_{n-1} \quad D_1 = -k_1 \end{aligned} \qquad (3.15)$$

$$\begin{aligned} E_n &= E_{n-1} - k_n zF_{n-1} \quad E_1 = k_1 \\ F_n &= zF_{n-1} - k_n E_{n-1} \quad F_1 = -1 \end{aligned} \qquad (3.16)$$

The immitance function, $R_n(z)$, is obtained using Eqs. (3.6) and (3.14)

$$R_n(z) = \left.\frac{1 + W_n(z,s)}{1 - W_n(z,s)}\right]_{s=0}$$

$$= \frac{(C_{n-1} - k_n zD_{n-1}) + z(E_{n-1} - k_n zF_{n-1})}{(C_{n-1} - k_n zD_{n-1}) - z(E_{n-1} - k_n zF_{n-1})} \qquad (3.17)$$

Polynomials $A_{n-1}(z)$, $B_{n-1}(z)$, $A^*_{n-1}(z)$ and $B^*_{n-1}(z)$ are defined

as

$$
\begin{aligned}
A_{n-1}(z) &= C_{n-1} - z\,E_{n-1} \\
B_{n-1}(z) &= D_{n-1} - z\,F_{n-1} \\
A^{*}_{n-1}(z) &= C_{n-1} + z\,E_{n-1} \\
B^{*}_{n-1}(z) &= D_{n-1} + z\,F_{n-1}
\end{aligned}
\tag{3.18}
$$

The immitance function of $R_n(z)$ is,

$$
\begin{aligned}
R_n(z) &= \frac{A^{*}_{n-1}(z) - k_n z\, B^{*}_{n-1}(z)}{A_{n-1}(z) - k_n z\, B_{n-1}(z)} \\
&= \frac{A^{*}_{n}(z)}{A_{n}(z)}
\end{aligned}
\tag{3.19}
$$

where $A_n(z)$, $B_n(z)$, $A^{*}_n(z)$ and $B^{*}_n(z)$ follow

$$
\begin{aligned}
A_n(z) &= A_{n-1}(z) - k_n z\, B_{n-1}(z) \qquad A_0(z) = 1 \\
B_n(z) &= zB_{n-1}(z) - k_n A_{n-1}(z) \qquad B_0(z) = 1
\end{aligned}
\tag{3.20}
$$

$$
A^{*}_n(z) = A^{*}_{n-1}(z) - k_n z\, B^{*}_{n-1}(z) \qquad A^{*}_0(z) = 1
\tag{3.21}
$$

$$B_n^*(z) = zB_{n-1}^*(z) - k_n A_{n-1}^*(z) \qquad B_0^*(z) = -1$$

Polynomials $A_n(z)$ and $A_n^*(z)$ follow nearly the same formation and only the initial condition of $B_n(z)$ is different. Equations (3.20) and (3.21) are the same as Eqs. (2.32) and (2.33).

$R_n(z)$ is expanded into

$$R_n(z) = \sum_{i=1}^{n} \frac{1 + z/z_i}{1 - z/z_i} \mu_i \tag{3.22}$$

where z_i is a root of $A_n(z)=0$ and μ_i is obtained by

$$\mu_i = - \frac{A_n^*(z_i)}{2 z_i A_n'(z_i)}$$

$$\sum_{i=1}^{n} \mu_i = 1 \tag{3.23}$$

where $A_n'(z)$ means the differential calculus of $A_n(z)$ by z.

As explained in a previous section, the power transfer function, $S_n(\lambda)$, is a real part of $R_n(z)$.

$$S_n(\lambda) = \mathrm{Re}[\, Rn(e^{-j\lambda})]$$

$$= \sum_{i=1}^{n} \frac{(r_i^2 - 1)m_i + 2r_i \sin(\lambda - \lambda_i)n_i}{r_i^2 - 2r_i \cos(\lambda - \lambda_i) + 1} \tag{3.24}$$

where $\mu_i = m_i + jn_i$, $z_i = r_i e^{-j\lambda_i}$ $(r_i > 1)$

Each term of the summation in Eq. (3.24) is represented as the mono-resonance curve and its integration between $-\pi$ and π is a constant value of $2\pi m_i$ independent of r_i , λ_i and n_i . Furthermore, when r_i is nearly one, its resonance frequency is λ_i and its half band width is $2(r_i - 1)$.

From Eq. (3.23),

$$\sum_{i=1}^{n} m_i = 1 \tag{3.25}$$

The value of m_i is interpreted as the relative resonance intensity corresponding to z_i , when the total resonance power is assumed to be one.

3.2.3 Line Spectrum Representation of Power Transfer Function

The pseudo vocal tract in Fig. 3.1 has a power loss only at the glottis terminated by a matched impedance. If the boundary condition at the glottis is replaced by a complete closure or a complete opening, power loss becomes zero. According to this condition, all poles of the resulting vocal transfer function will move onto the unit circle in the z-plane. These conditions correspond to $k_{P+1} = -1$ or $k_{P+1} = +1$ respectively. Using this property, it is possible to represent the power transfer function as the sum of delta functions, such as

$$S^{(\pm)}(\lambda) = 2\pi \sum_{i=1}^{P+1} m_i^{(\pm)} \delta(\lambda - \lambda_i^{(\pm)}) \tag{3.26}$$

where the sign of each variable is corresponding to the two new boundary conditions at the glottis.

$\lambda_i^{(\pm)}$ and $m_i^{(\pm)}$ are the value when $A_n(z)$ is replaced by $A_{p+1}^{(\pm)}(z) = A_p(z) \pm zB_p(z)$ and $A_n^*(z)$ is replaced $A_{p+1}^{*(\pm)}(z) = A_p^*(z) \pm zB_p^*(z)$.

The immitance function, when $k_{p+1} = \pm 1$, is

$$R_{p+1}^{(\pm)}(z) = \frac{A_p^*(z) \pm zB_p^*(z)}{A_p(z) \pm zB_p(z)} \tag{3.27}$$

This function is a reactance function as to $|z| = 1$ and has the following properties.

(1) Poles and zeros of $R_{p+1}^{(\pm)}(z)$ lie on the unit circle and their order is one.

(2) μ_i corresponding to poles of $R_{p+1}^{(\pm)}(z)$ is not negative.

(3) Poles and zeros of $R_{p+1}^{(\pm)}(z)$ alternate with each other.

(4) Poles of $R_{p+1}^{(\pm)}(z)$ and $R_{p+1}^{(\pm)}(z)$ alternate with each other

(5) Zeros of $R_{p+1}^{(\pm)}(z)$ and $R_{p+1}^{(\pm)}(z)$ alternate with each other.

Representing the power transform function in an all-pole filter, using $m_i^{(\pm)}$ and $\lambda_i^{(\pm)}$, is named "line spectrum representation (LSR)". Parameters $m_i^{(\pm)}$ and $\lambda_i^{(\pm)}$ are called line spectrum amplitude and normalized line spectrum frequency. These parameters are called LSR parameters in this thesis.

From LSR properties, the poles and zeros displacement for $R_{p+1}^{(\pm)}(z)$ is classified into four categories, as shown in Table 3.1, according to

the sign of k_{p+1}. As the poles and zeros for $R^{(\pm)}_{p+1}(z)$ are conjugate, only poles and zero lying in $\lambda > 0$ are shown in this table.

3.2.4 Relationship Between Autocorrelation Coefficients and LSR Parameters

Autocorrelation coefficients are obtained by the Fourier transform of power transform function. Autocorrelation coefficients, corresponding to Eq. (3.26), are represented by

$$V'_{\tau} = \frac{1}{2\pi} \int_{-\pi}^{\pi} S^{(\pm)}(\lambda)\ e^{j\lambda\tau}\ d\lambda \qquad (3.28)$$

$$= \sum_{i=1}^{p+1} m_i\ e^{j\lambda_i\tau}$$

Using the complex conjugate property of λ_i ,

$$V'_{\tau} = m_0 + (-1)^{\tau} m_{n+1} + 2\sum_{i=1}^{n} m_i \cos\lambda_i\tau \qquad (3.29)$$

where m_0 and m_{n+1} are the amplitudes at $\lambda = 0$ and $\lambda = \pi$, respectively. Normalized line spectrum frequency λ_i is converted into f_i, whose dimension is real frequency (kHz), according to the following equation.

$$f_i = \lambda_i / 2\pi\ \Delta\ T \qquad (3.30)$$

This f_i is re-called line spectrum frequency. Autocorrelation function V_{τ}, corresponding to PARCOR coefficients $k_1, k_2, \text{----}, k_p$,

is proved to be the exact value of V_τ , namely,

$$V_\tau = V'_\tau \quad (\tau = 1,2,----,p) \tag{3.31}$$

This is evident, because only k_{p+1} is changed into +1 or -1 artificially.

To sum up, the flow-chart used in extracting line spectrum parameters from the input speech signal is represented as in Fig. 3.4.

Another LSR parameter extraction method was proposed, starting from the relation in Eq. (3.29) [44],[45].

3.3 Pseudo Formant Vocoder

LSR parameters can be converted into PARCOR coefficients mutually via autocorrelation coefficients using the relation in Eq. (3.29). Consequently, it is possible to construct a new vocoder system to transmit LSR parameters instead of formant parameters, or PARCOR parameters or α parameters etc. This new vocoder system was named "Pseudo formant vocoder", because line spectral parameters have nearly the same representation of formant parameters. However, in most cases, both values are not the same. In fact, the relationship between the first two line spectrum frequencies for five Japanese vowels, uttered by ten male and ten female speakers, is shown in Fig. 3.5. The mutual locations of the five vowels are nearly the same, in the case of formant frequency, but the frequency values are a little different from usual formant frequencies [46].

To construct the pseudo formant vocoder, two parts of line spectrum analysis and PARCOR coefficients reconstruction are added to

PARCOR vocoder system. The same excitation source parameters as those in PARCOR vocoder system are also used in the pseudo formant vocoder. This new vocoder system is shown in Fig. 3.6. Calculation amount in PARCOR is about 25000 multiplications and additions per frame, when using the 10-th order analysis. The increase calculation amount is 700 multiplications and additions, owing to line spectrum analysis and PARCOR coefficients reconstruction. The increase rate is under five percent, on the basis of the PARCOR system.

3.4 Statistical Quantization Characteristics [47]

In order to quantize LSR parameters efficiently, statistical properties were examined experimentally.

3.4.1 Parameter Distribution

Experimental conditions are listed in Table 3.2. The distribution ranges for line spectrum frequencies are shown in Figs. 3.7(a) and (b) as a function of the analysis order (denoted p). Figure 3.7(a) is for when k_{p+1}= +1 (complete opening at the glottis) and Figure 3.7(b) is for when k_{p+1}= -1 (complete closure at the glottis). The LSR parameters distributions are shown in Figs. 3.8(a) and (b), when p is ten and k_{p+1}=-1. The distribution range for each line spectrum frequency is from 1/4th to 1/8th of the full range. Efficient encoding can be achieved, considering the distribution ranges. Line spectrum amplitude are distributed in the wide range, from about 20 to 40 dB. The first line spectrum amplitude is rather larger than the other parameters. The LSR parameters fluctuation is shown in Fig. 3.9. This speech sample was

uttered by a female speaker. The lower point of each bar at each frame period indicates the line spectrum frequency location and the width of each bar is proportional to $\sqrt{m_i}$. As mentioned before, line spectrum frequencies never intersect over frame periods.

3.4.2 Parameter Quantization Characteristics

One synthesized speech impairment is caused by quantization of parameters. In this thesis, spectral distortion caused by parameter quantization is introduced as the objective assessment measure for synthesized speech quality [48],[49].

First, spectral sensitivity is introduced. Spectral distortion is defined as a log term mean square spectrum error between an original speech spectrum $S_0(\omega)$ and a synthesized speech spectrum $S(\omega)$ caused by quantization. This difference is illustrated by the hatched area in Fig. 3.10. Namely, spectral distortion ΔS is defined as,

$$\Delta S = \sqrt{\frac{1}{W}\int_0^W \left(S_0(\omega) - S(\omega) \right)^2 df} \qquad (3.32)$$

where W is the input speech signal bandwidth. ΔS is accumulated over a long time by

$$\overline{\Delta S} = \sqrt{\frac{1}{T}\int_0^T \left(\Delta S \right)^2 dt} \qquad (3.33)$$

$\overline{\Delta S}$ is defined as mean spectral distortion. Spectral sensitivity is defined as the ratio of mean spectral distortion to a small perturbation of only one of the LSR parameters, such as

$$C_{f_i} = \overline{\Delta S} / \Delta f_i \qquad \text{(dB/Hz)} \tag{3.34}$$

$$C_{m_i} = \overline{\Delta S} / \Delta m_i \qquad \text{(dB/dB)} \tag{3.35}$$

If an LSR parameter has a large spectral sensitivity, a spectral envelope is much distorted with small perturbation by quantization.

Spectral sensitivities for LSR parameters were examined experimentally, using the same speech samples in Table 3.2. Figures 3.11(a) and (b) show a spectral envelope displacement, when f_1 is perturbed by $\pm$ 15.625 Hz or $\pm$ 31.25 Hz and m_1 is $\pm$ 0.4 dB and $\pm$ 0.8 dB from real value. Spectral distortion by perturbing f_1 and m_1 are a little different versus the frequency axis. Spectral sensitivities are shown in Table 3.3, when Δf_i =15.625 and Δm_i =0.4 dB.

Spectral sensitivities are uniform and there is large difference between male and female voices.

3.5 Spectral Distortion Estimation by Quantization

Spectral distortion by quantization is estimated using spectral sensitivity without calculating the spectral envelope.

The influence on the spectral envelope by LSR perturbation is assumed to be independent and superimposed. Then, the spectral distortion is estimated as

$$\widetilde{\Delta S} = \sqrt{\sum_{i=1}^{5} \left(C_{f_i} \Delta f_i' \right)^2 + \sum_{i=1}^{6} \left(C_{m_i} \Delta m_i' \right)^2} \quad \text{(dB)} \tag{3.36}$$

where $\Delta f_i'$ and $\Delta m_i'$ are quantization errors for LSR parameters.

Using Eq. (3.36), spectral distortion, in the case of any bit

allocation, is estimated. If bit allocation to LSR parameters is given, quantization step size Δ_i is calculated using the distribution ranges. If the quantization error is distributed uniformly, the standard deviation in quantization error is $\Delta_i/\sqrt{12}$. Substituting spectral sensitivity and standard deviation into Eq. (3.36), mean distortion is computed. Estimated spectral distortion and real spectral distortion are shown in Fig. 3.12. In this experiment, each LSR parameter is quantized by the same bit, from two to five bits. The $\Delta\tilde{S}/\Delta S$ ratio is about 1.05. It proved that the spectral distortion can be estimated using spectral sensitivity without much difference. Of course, it can be possible to estimate spectral distortion for any other bit allocation in the same manner.

Next, when the acceptable spectral distortion is given, the optimum bit allocation to LSR parameters is proposed. First, the acceptable spectral distortion is given. This value is usually determined considering speech impairment. It is better to set this as the maximum value as long as synthesized speech impairment is not detectable for human beings. It was reported as about one dB experimentally [50]. It relates with difference limen (DL) for formant frequency and its amplitude [51],[52].

When the number of LSR parameters is n, the acceptable spectral distortion for each parameter is $\sqrt{1/n}$ (dB). Standard deviation in quantization error, $\Delta_i/\sqrt{12}$, is computed by $\sqrt{1/n}$ and spectral sensitivity. Consequently, quantization step size, Δ_i , is calculated and bit allocation for each LSR parameter is determined, using distribution ranges and step size. This results in the optimum bit allocation according to the spectral sensitivity.

Several examples of the relation between quantization bit and

spectral distortion are shown in Fig. 3.13. The quantization bit value for each LSR parameter is determined using this figure, when the acceptable spectral distortion is given. Finally, spectral distortions versus quantization bits per frame are summed up in Fig. 3.14. When DL for spectral distortion is assumed to be one dB, 38 bits are needed for male and 40 bits for female voices. These values are about 85 % of PARCOR parameters.

3.6 Interpolation Characteristics

In order to achieve bits saving, parameter transmission frame rate must be longer, for from 20 mseconds to 30 mseconds. However, when the frame rate is longer, the fidelity, achieved in reconstructing the original spectral transition, decreases. To avoid this, the untransmitted parameters must be interpolated, using adjacent transmitted frames at the receiver. In this processing, spectral distortion is caused by the parameter difference between interpolated value and real untransmitted value. To distinguish the spectral distortion by quantization, the spectral distortion caused by interpolation is called "time spectral distortion". In this section, the interpolation properties for LSR parameters are discussed, comparing with properties for PARCOR coefficients. When the fundamental analysis period is 10 mseconds, and the frame intervals are varied from 20 mseconds to 70 mseconds, spectral distortions in the untransmitted frame were calculated. The time spectral distortion are shown in Figs. 3.15 (a) and (b) versus frame period. The time spectral distortion for LSR parameters is about 80 % of PARCOR, at every frame period.

The spectral distortion correlation at the same frame is shown in

Fig. 3.16. A strong correlation is observed. Spectral distortion distributions in each frame are shown in Fig. 3.17. There are many frames which have a large spectral distortion in PARCOR coefficients.

Spectral pattern transitions, which are generated by linear interpolation of PARCOR coefficients or LSR parameters, are shown in Figs. 3.18 (b) and (c), respectively. Figure 3.17 (a) shows the original transition without interpolation. Analysis frame period was 10 mseconds and both parameters were linear interpolated every 50 mseconds. In Fig.3.18(b), there are several spectral peaks indicated by arrows which do not appear in original spectral patterns. Namely, the undesirable spectral envelopes are sometimes produced by interpolated PARCOR coefficients. These spectral patterns generate uncomfortable synthesized speech. One of the defects in the PARCOR vocoder lie in such an interpolation property. On the other hand, there are no undesirable spectral shapes, as shown in Fig.3.18(c), which is similar to the original spectral transition.

3.7 Conclusions

(1) A new representation of LPC parameters is introduced. These parameters are named LSR parameters and are similar to formant parameters in frequency domain.

(2) A new vocoder system, named "Pseudo formant vocoder", is introduced.

(3) Statistical properties of LSR parameters were examined. LSR spectral sensitivities were experimentally found to be uniform.

(4) Spectral distortion by quantization can be estimated using spectral sensitivity analysis.

(5) Optimum bits allocation in pseudo formant vocoder was proposed and the differences between LSR and PARCOR were examined. As a result, the information rate was 85 % that for the PARCOR system.

(6) LSR parameters have a good interpolation property. The time spectral distortion at the same frame period was 80 % that for the PARCOR system.

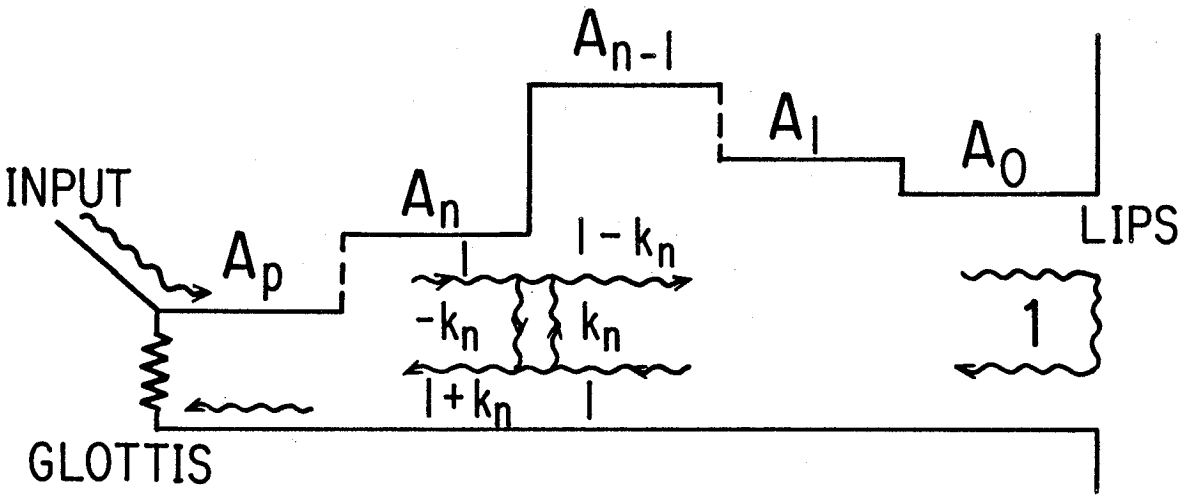


Fig. 3.1 Pseudo vocal tract model as a lossless acoustic tube terminated by a matched impedance at the glottis.

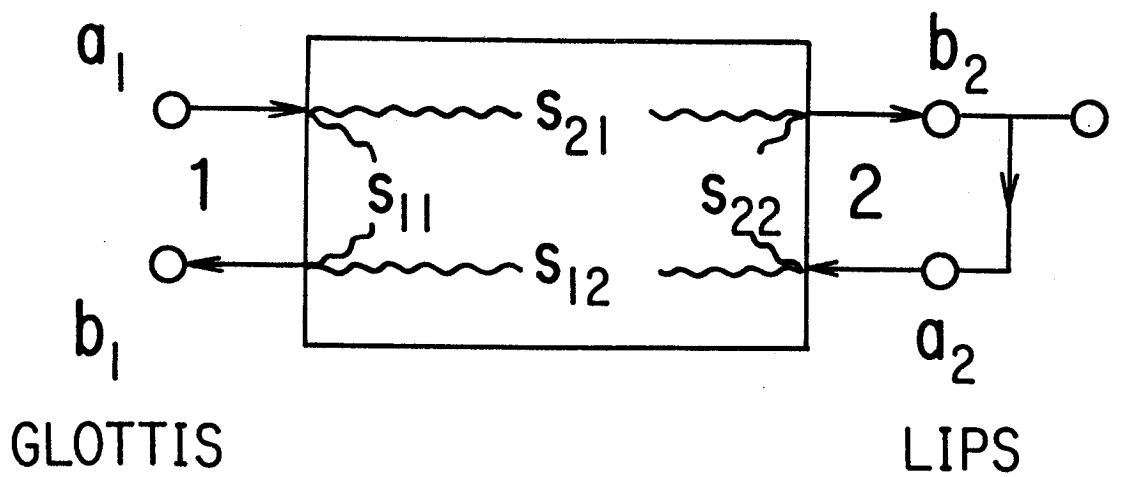


Fig. 3.2 S matrix representation of the lossless vocal tract.

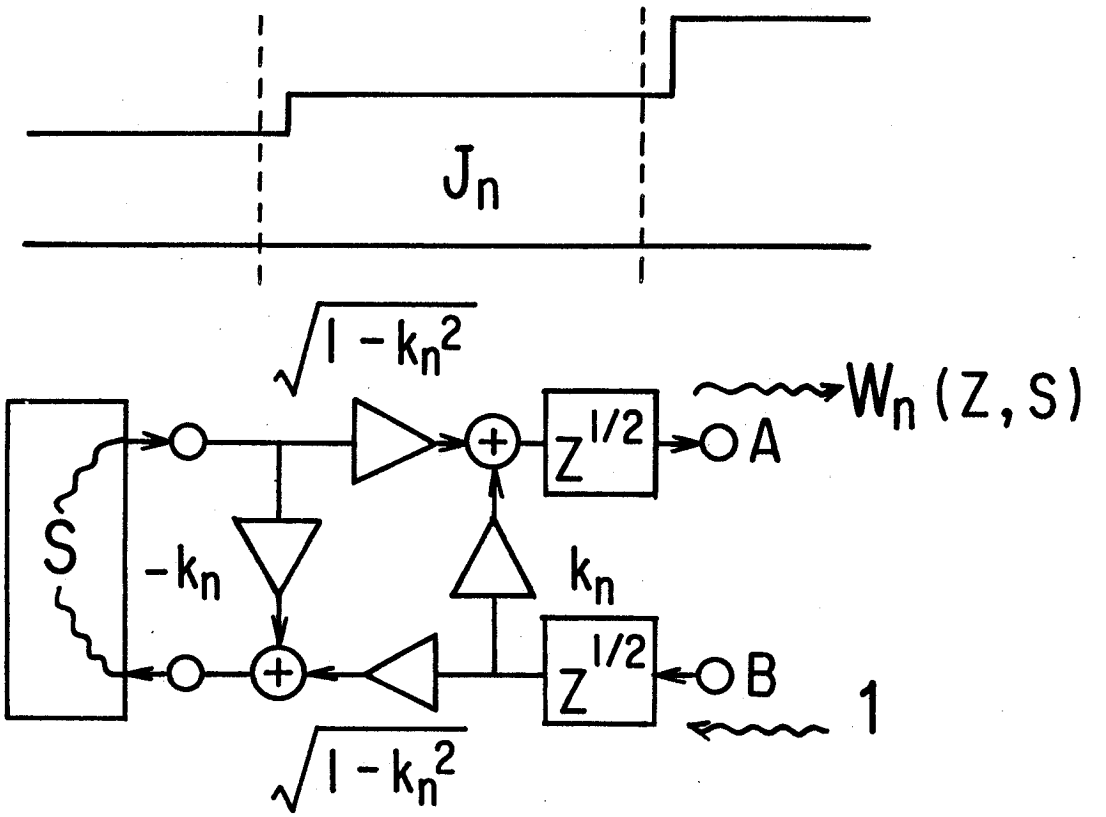


Fig. 3.3 Equivalent circuit of an acoustic tube junction using PARCOR coefficients.

Table 3.1 Poles and zeros displacement of immittance function.

p	k_{p+1}	POLE-ZERO PATTERN
		f=0 — f=4 kHz
2n+1	+ 1	x—o—x—o—x—o—x
2n−1	− 1	o—x—o—x—o—x—o
2n	+ 1	x—o—x—o—x——o
2n	− 1	o—x—o—x—o—x

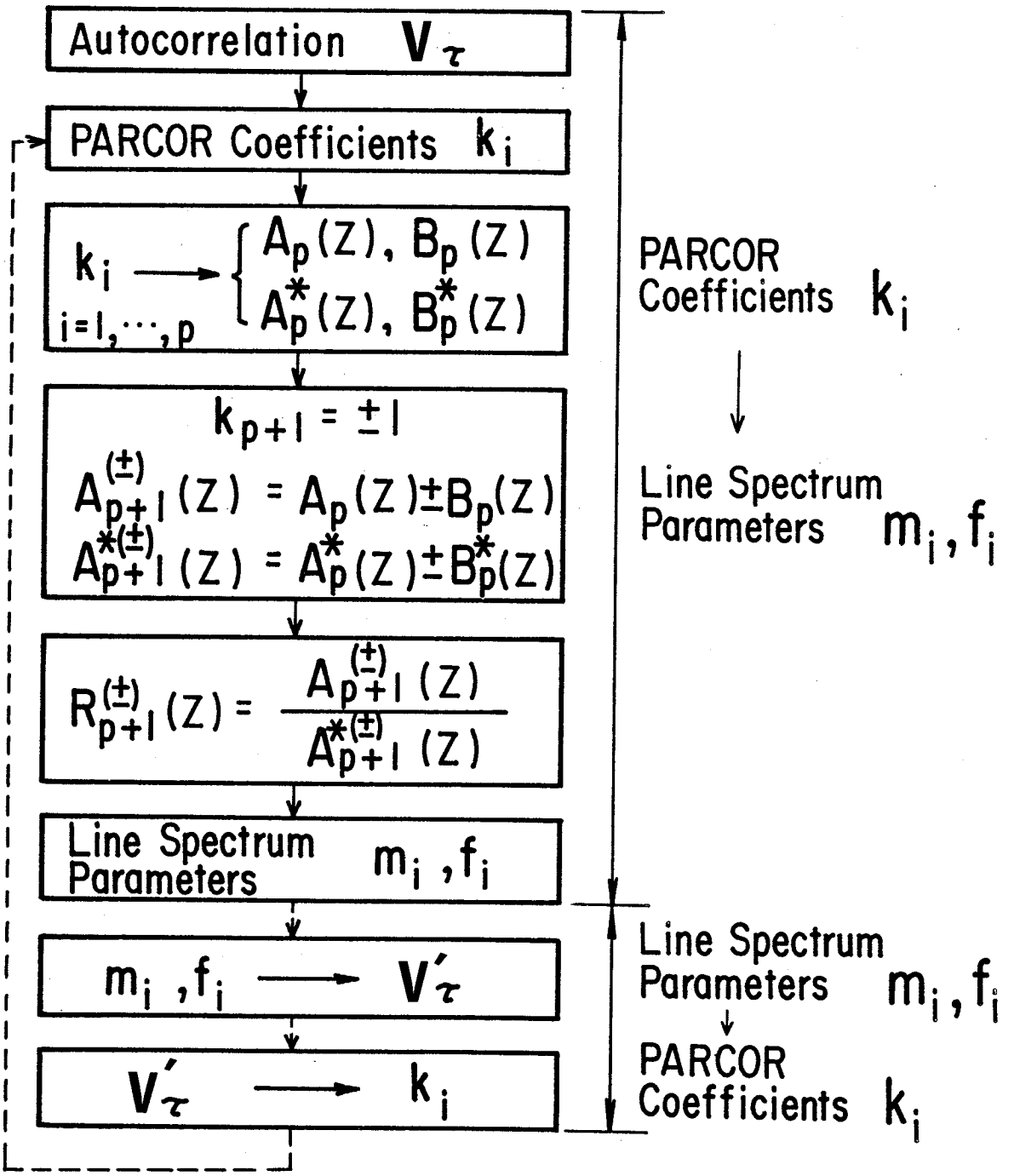


Fig. 3.4 Procedure for computing line spectrum parameters from input speech signal via PARCOR coefficients.

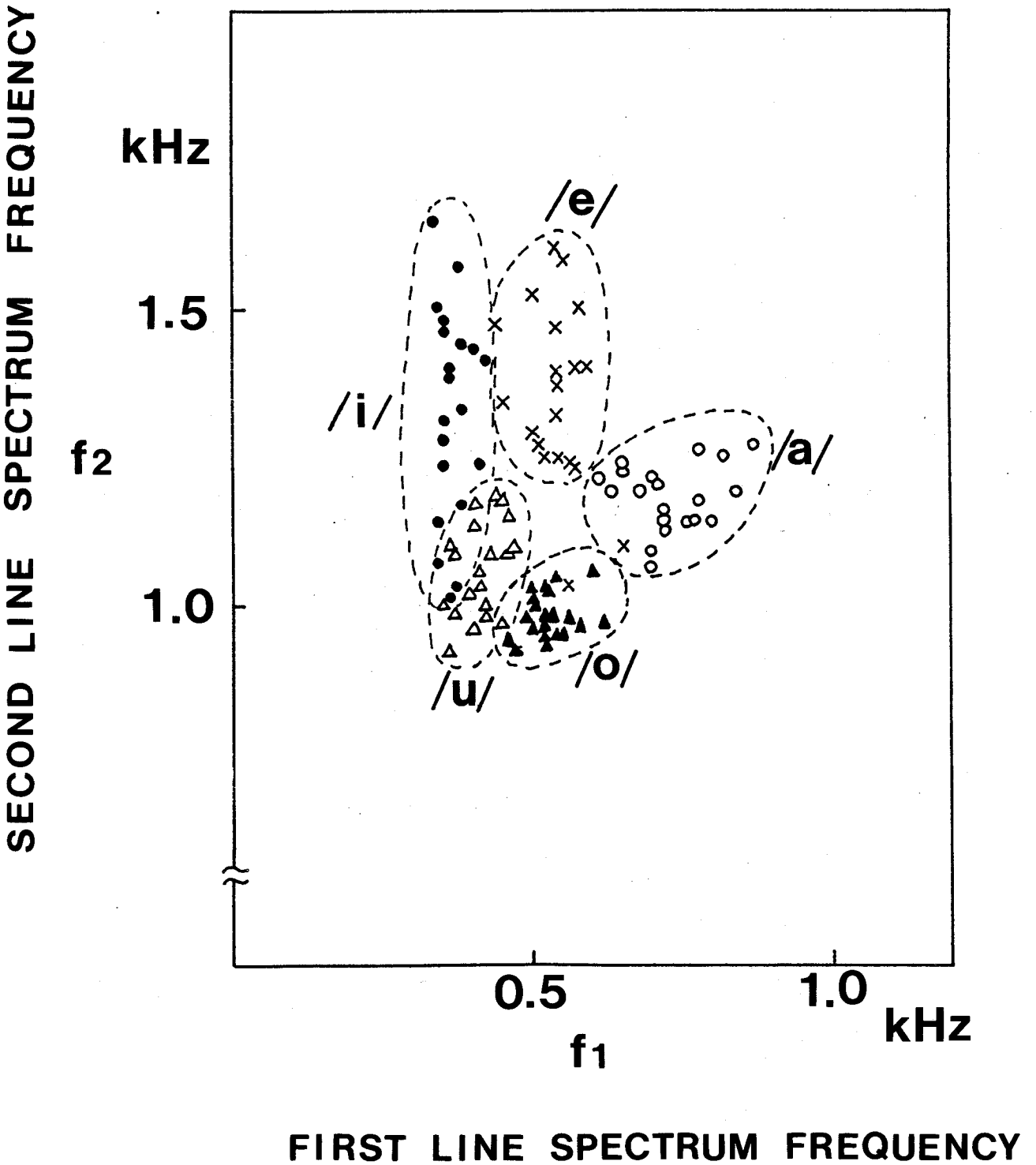


Fig. 3.5 Distribution of the first two line spectrum frequencies for five Japanese vowels.

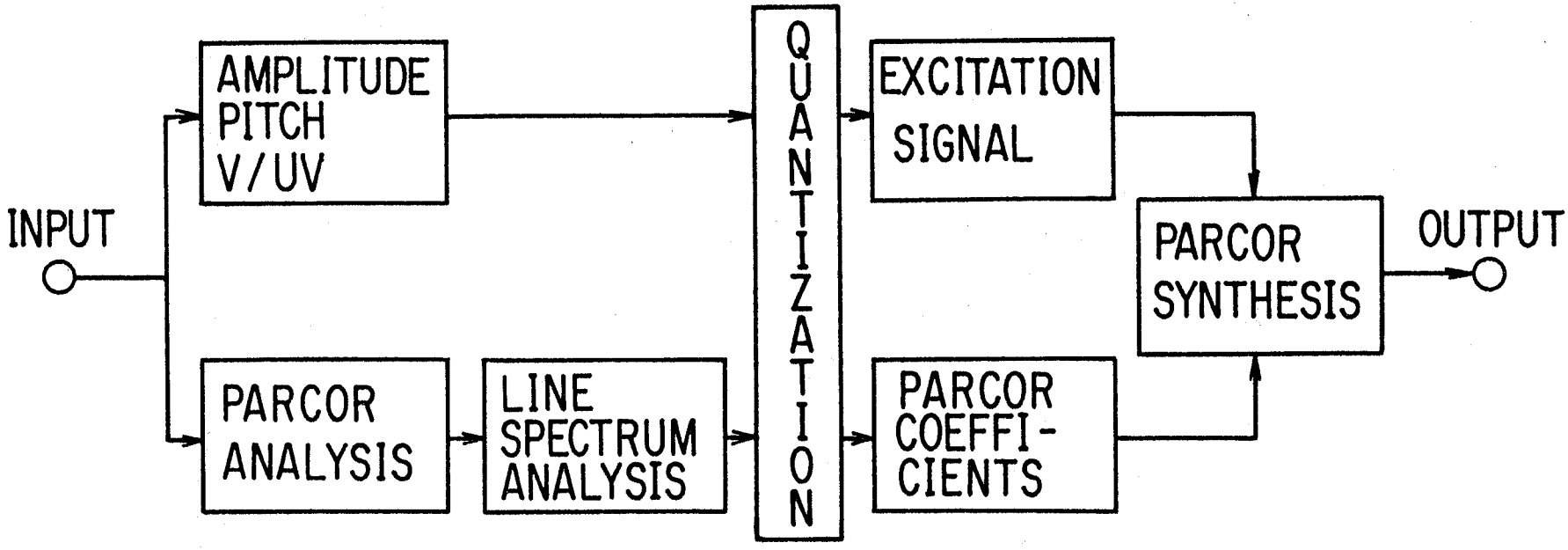


Fig. 3.6 Speech analysis synthesis system by LSR parameters, (Pseudo formant vocoder)

Table 3.2 Experimental conditions for extracting LSR parameters.

SPEECH DATA	
Contents	' bakuoNga giNsekaino koogeNni hirogaru '
Speakers	10 male speakers (2724 frames) 10 female speakers (2683 frames)
Sampling frequency	8 kHz
Cut off frequency for low pass filter	3.4 kHz
AD conversion accuracy	12 bits
L S R ANALYSIS	
Frame period	10 mseconds
Window	20 mseconds Hamming window
Prediction order	10
Boundary condition at the glottis	$k_{p+1} = -1$ (complete opening)

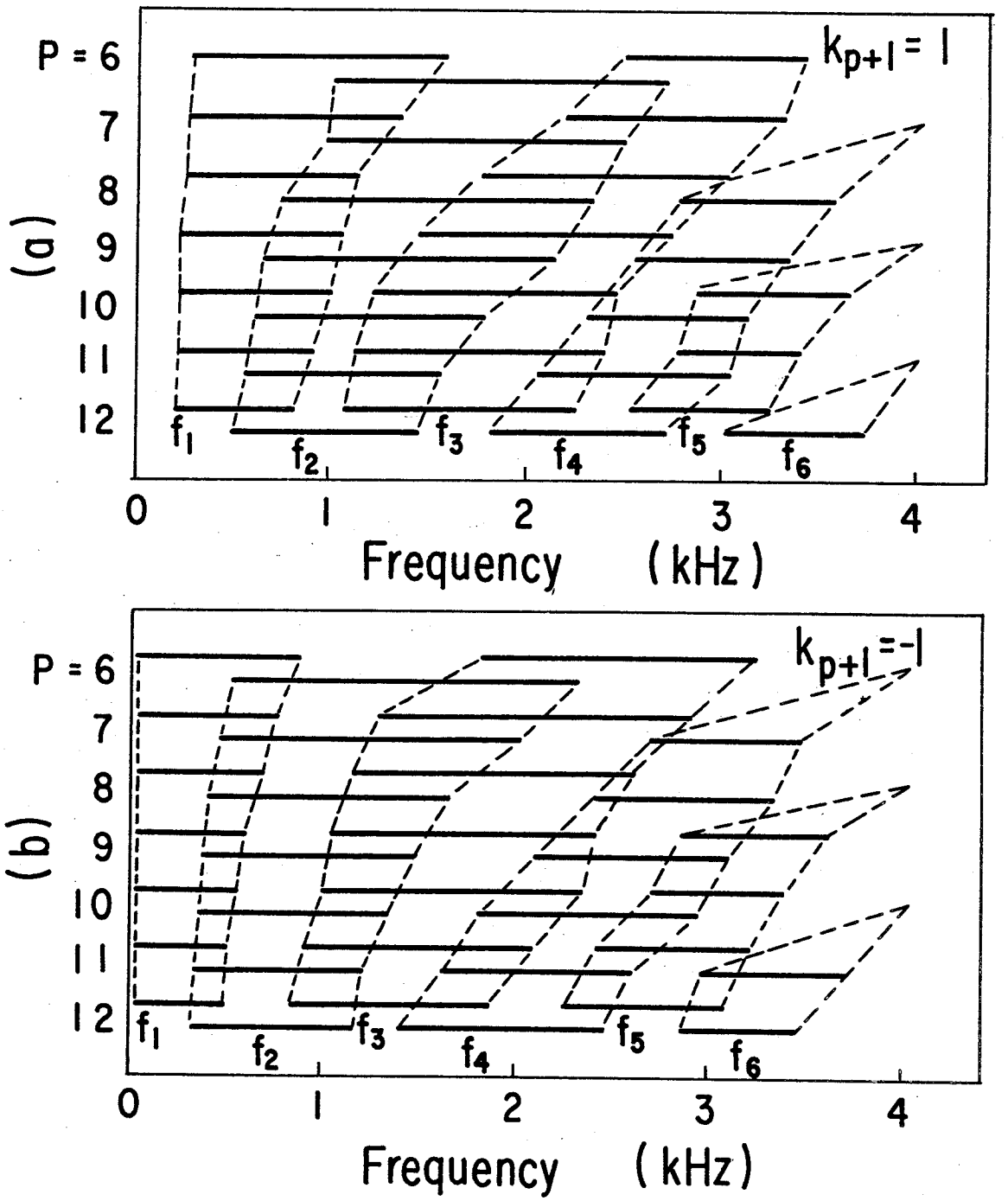


Fig. 3.7 Distribution ranges for line spectrum frequencies.

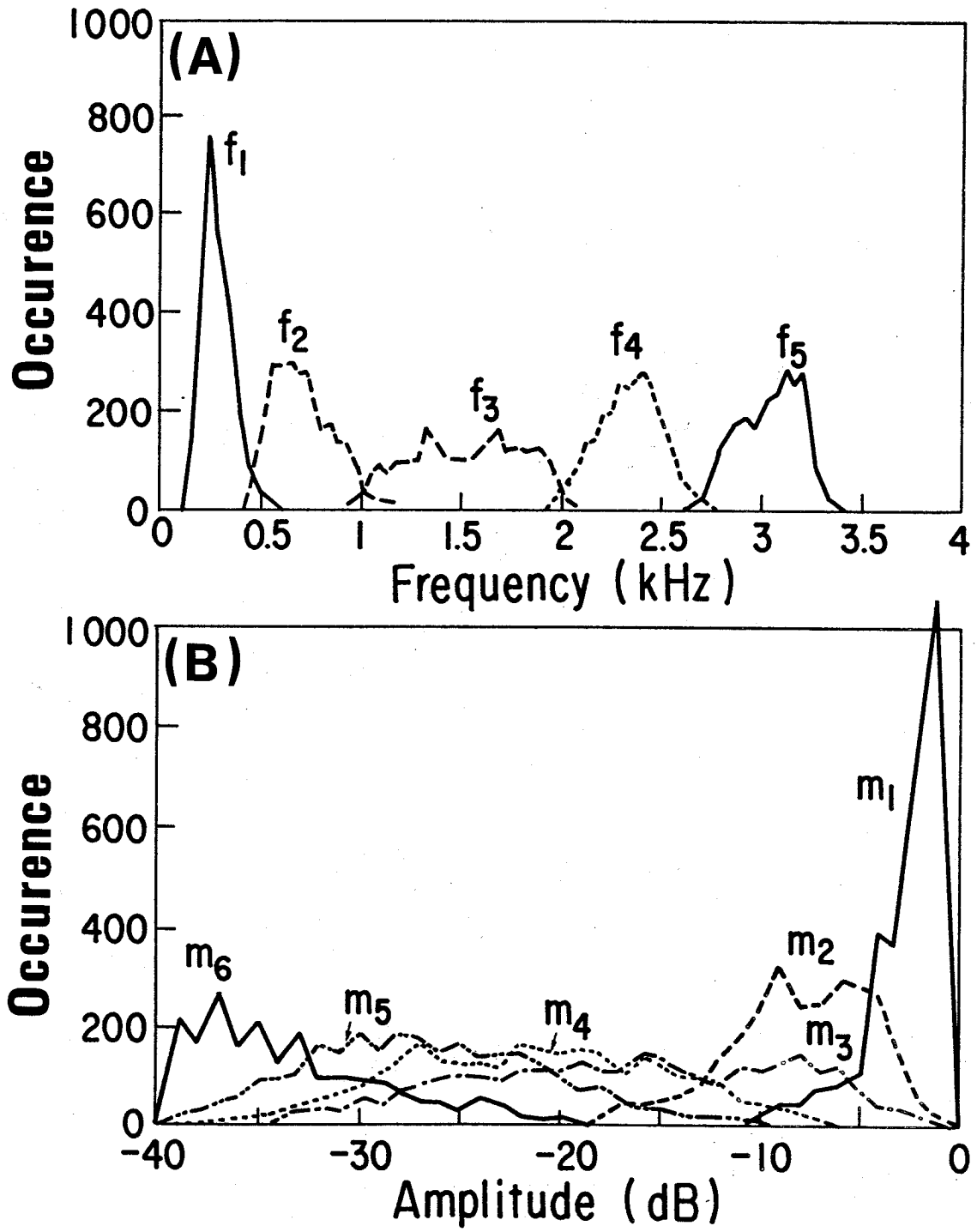


Fig. 3.8 LSR parameters occurences.

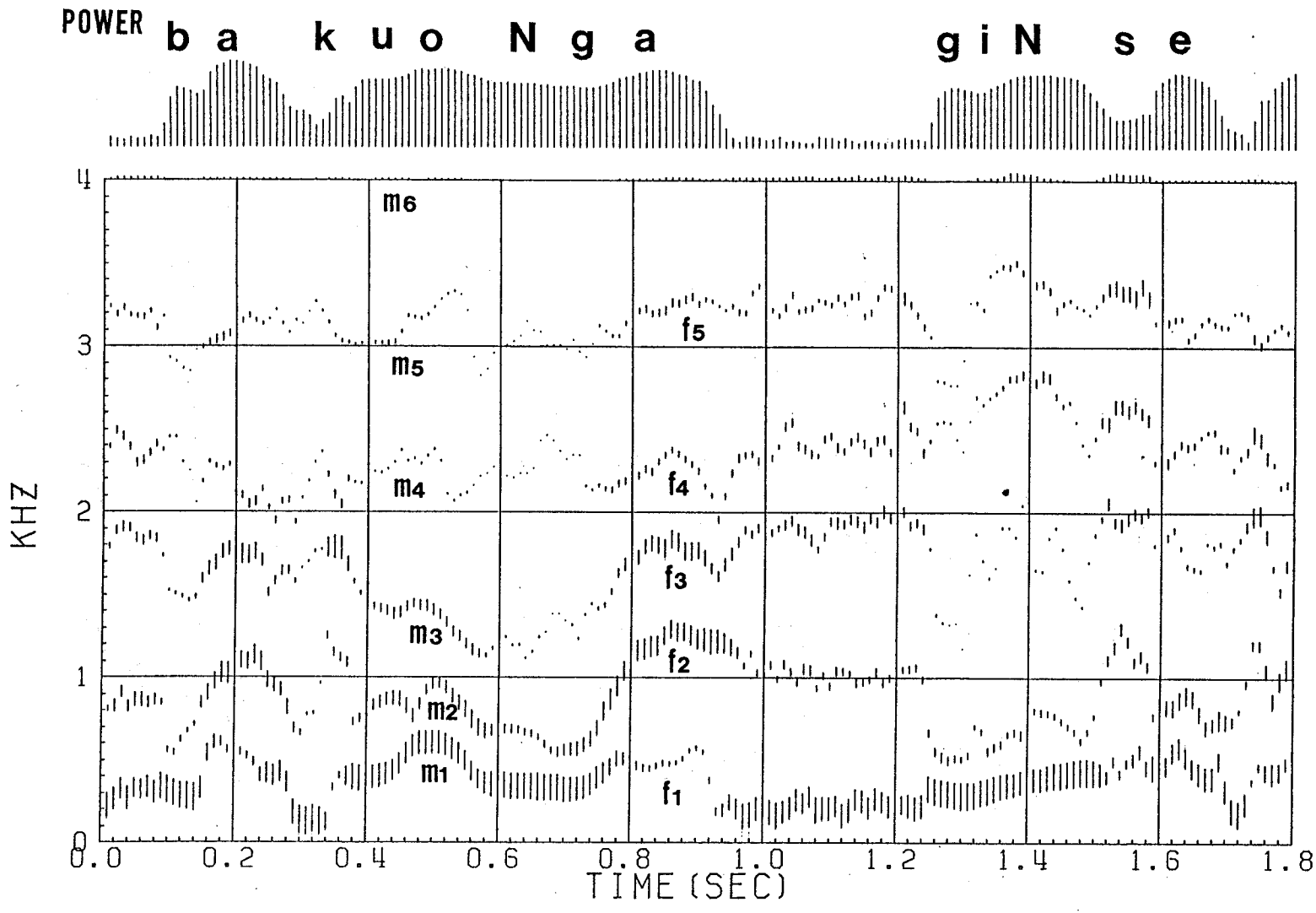


Fig. 3.9 Fluctuation in LSR parameters f_i and m_i . Speech sample is /bakuoNga giNse/ uttered by a female speaker. The lower point in each bar indicates the loaction in line spectrum frequency. The length of each bar is proportional to $\sqrt{m_i}$.

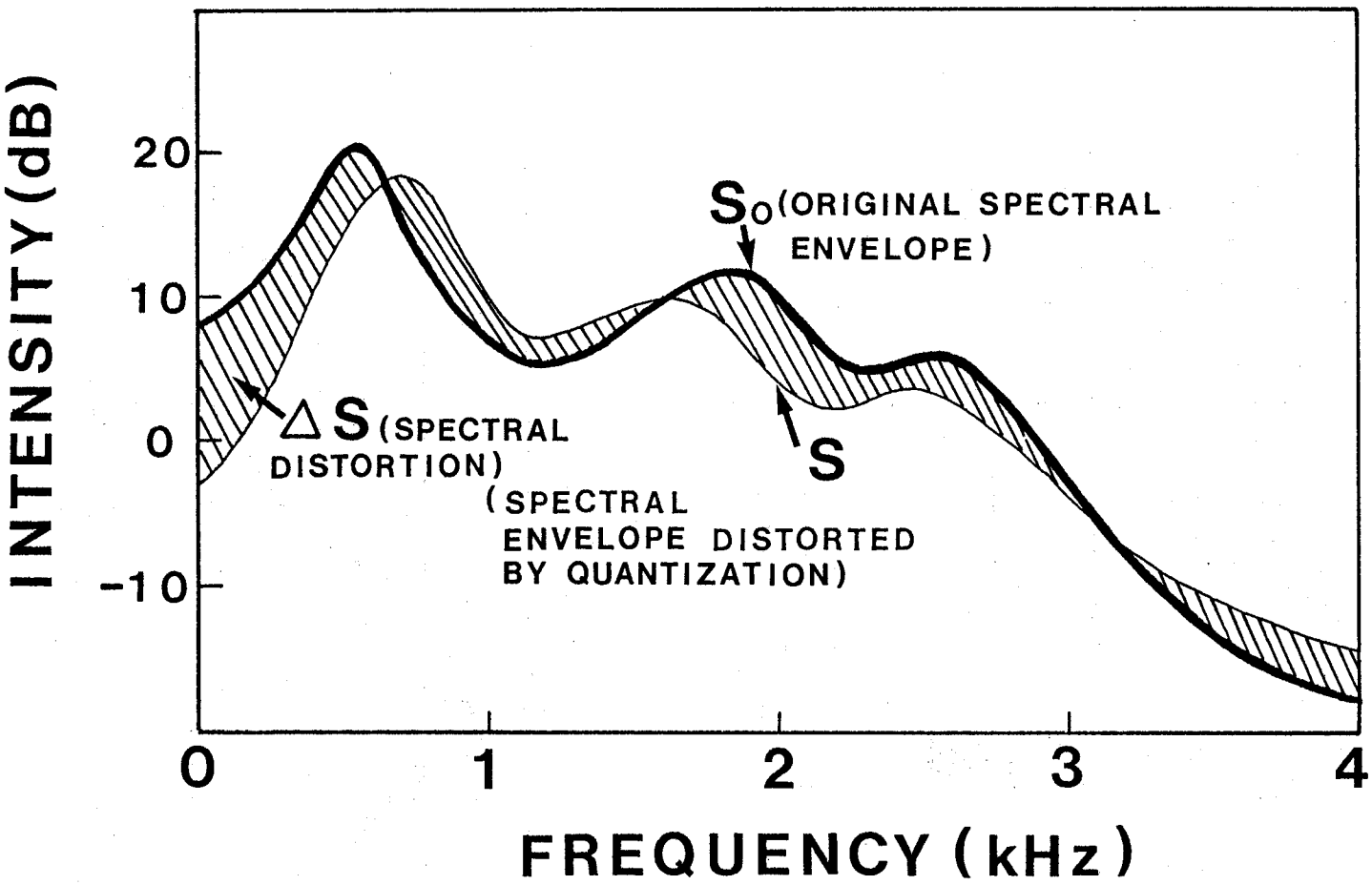


Fig. 3.10 Spectral envelope displacement by LSR parameters purturbation.

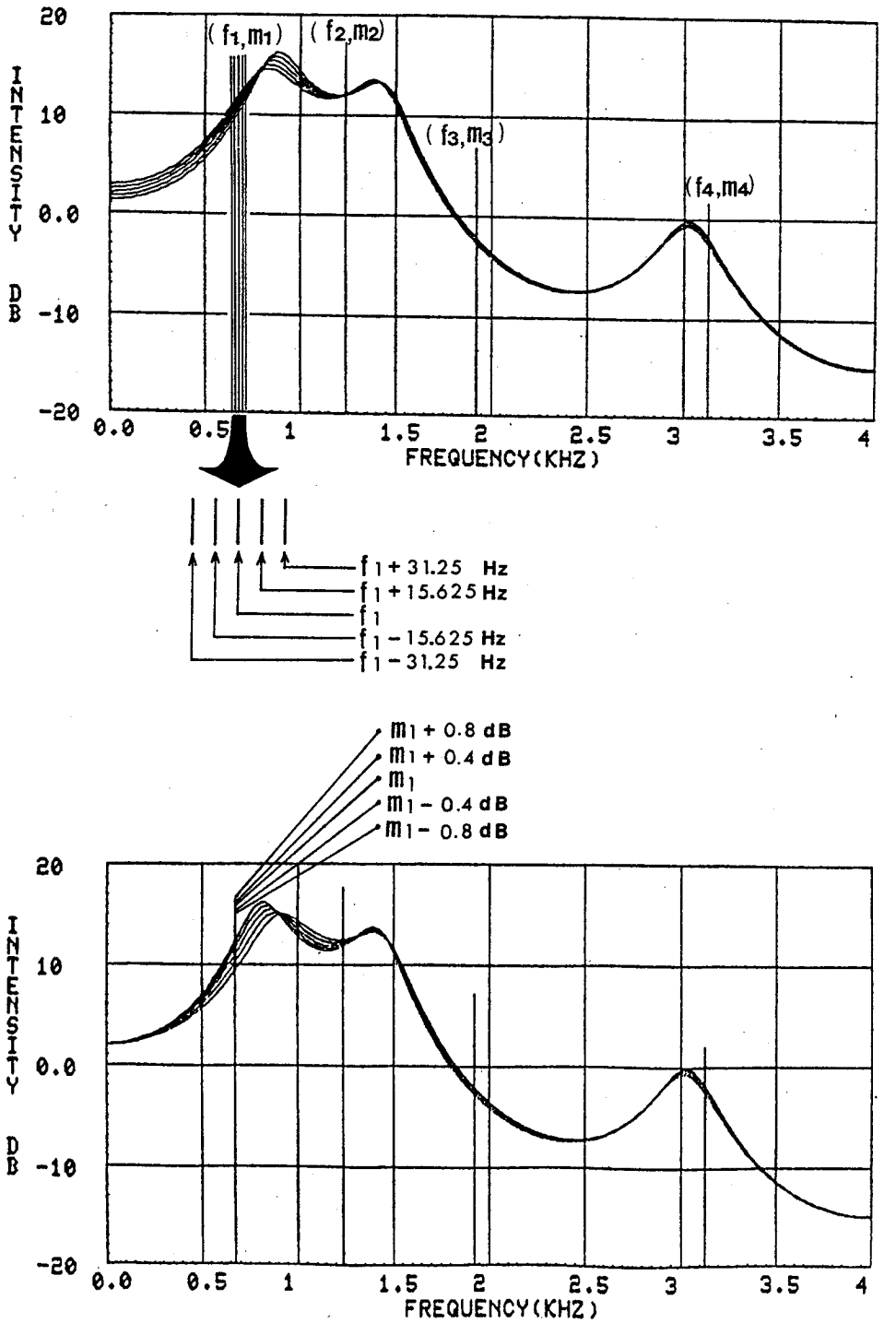


Fig. 3.11 Spectral envelope displacement by LSR parameters' small perturbation.

Table 3.3 Spectral sensitivities for LSR parameters.

	Male	Female
f_1	0.018 dB/Hz	0.022 dB/Hz
f_2	0.024	0.025
f_3	0.023	0.020
f_4	0.018	0.017
f_5	0.017	0.018
m_1	0.51 dB/dB	0.43 dB/dB
m_2	0.58	0.51
m_3	0.53	0.55
m_4	0.51	0.61
m_5	0.54	0.52
m_6	0.51	0.52

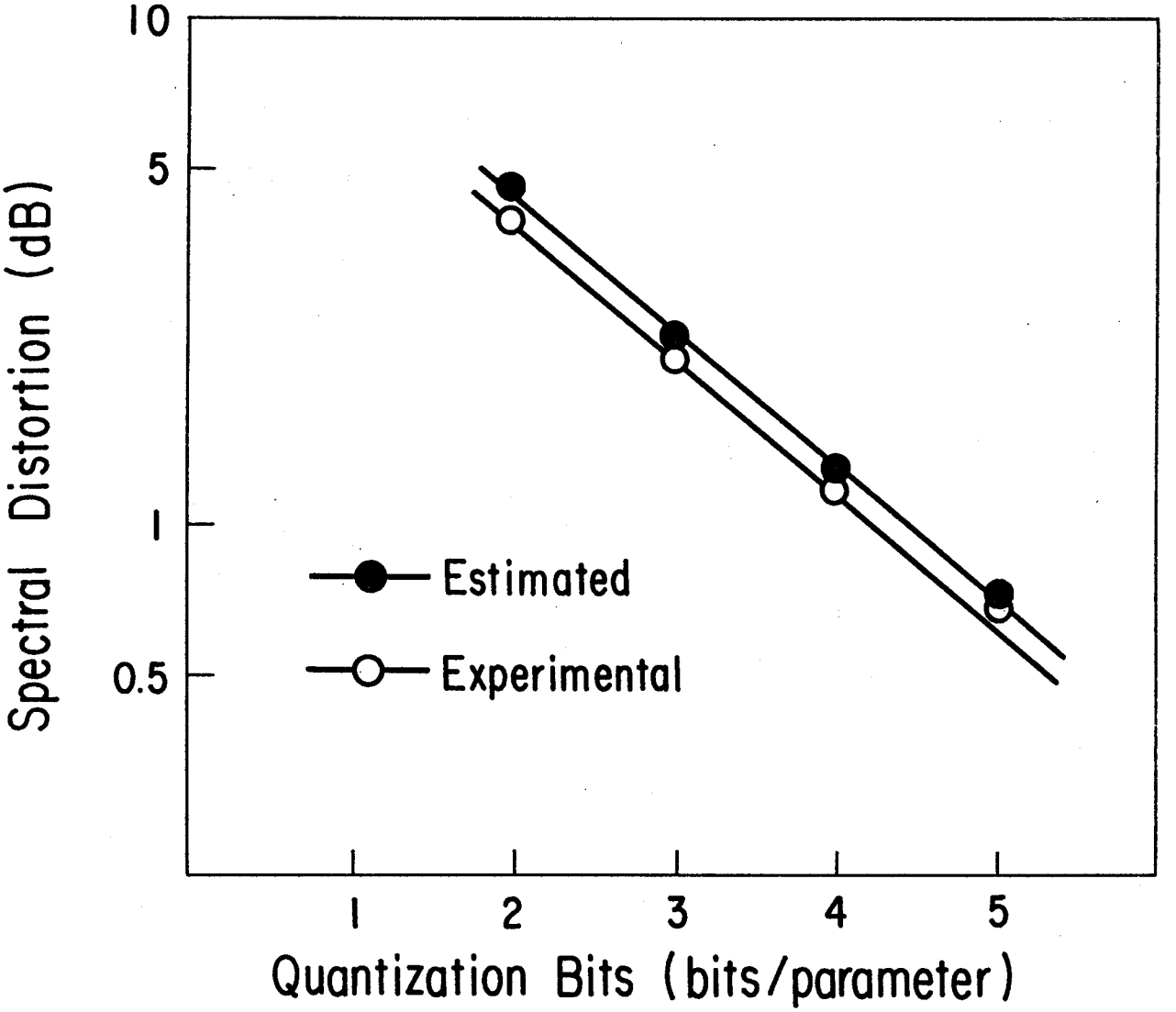


Fig. 3.12 Comparison between estimated spectral distortion and experimental result.

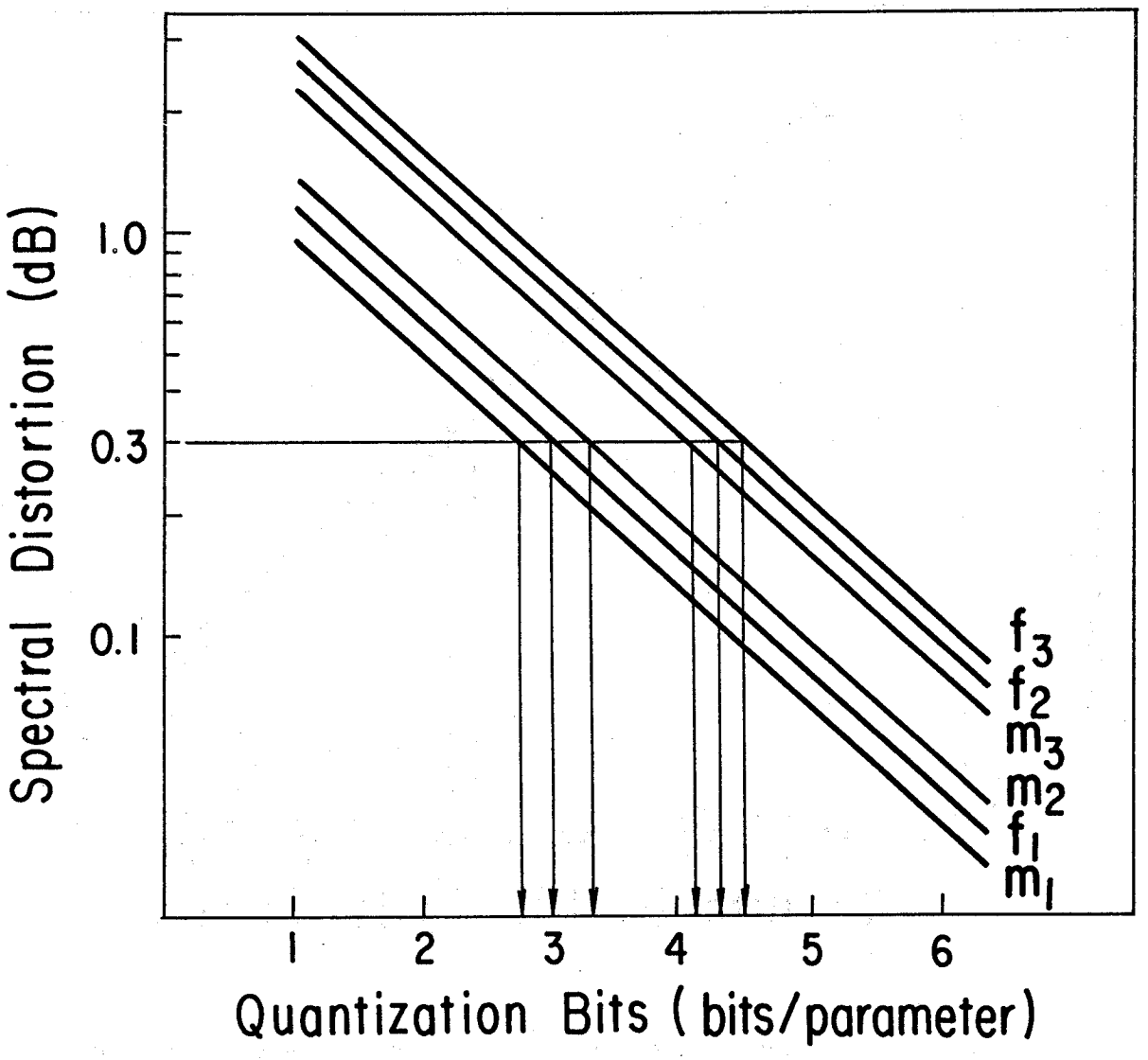


Fig. 3.13 Relation between spectrum distortion and quantization bits of each LSR parameter.

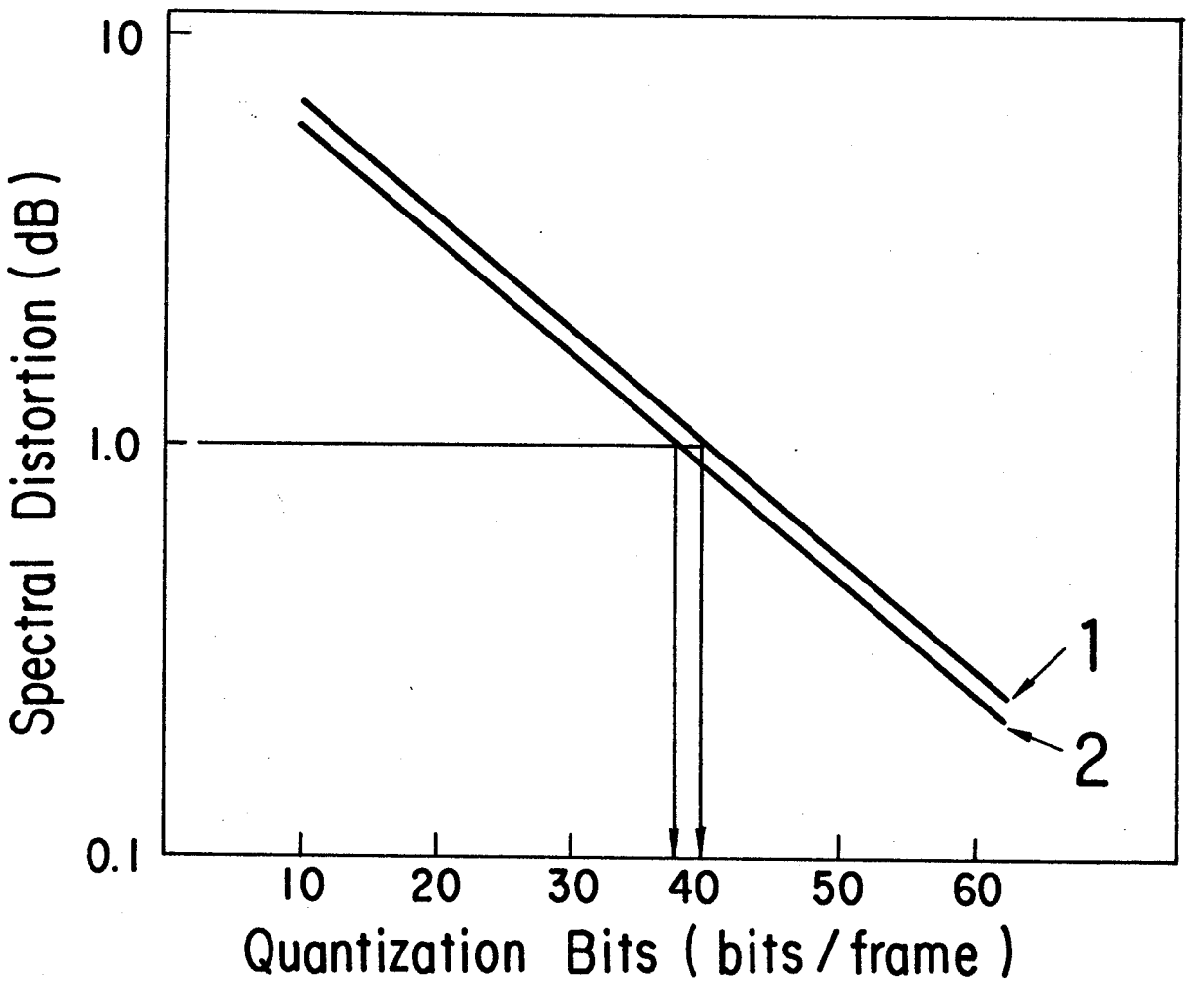


Fig. 3.14 Spectral distortion per frame versus quantization bits.
1:female 2:male

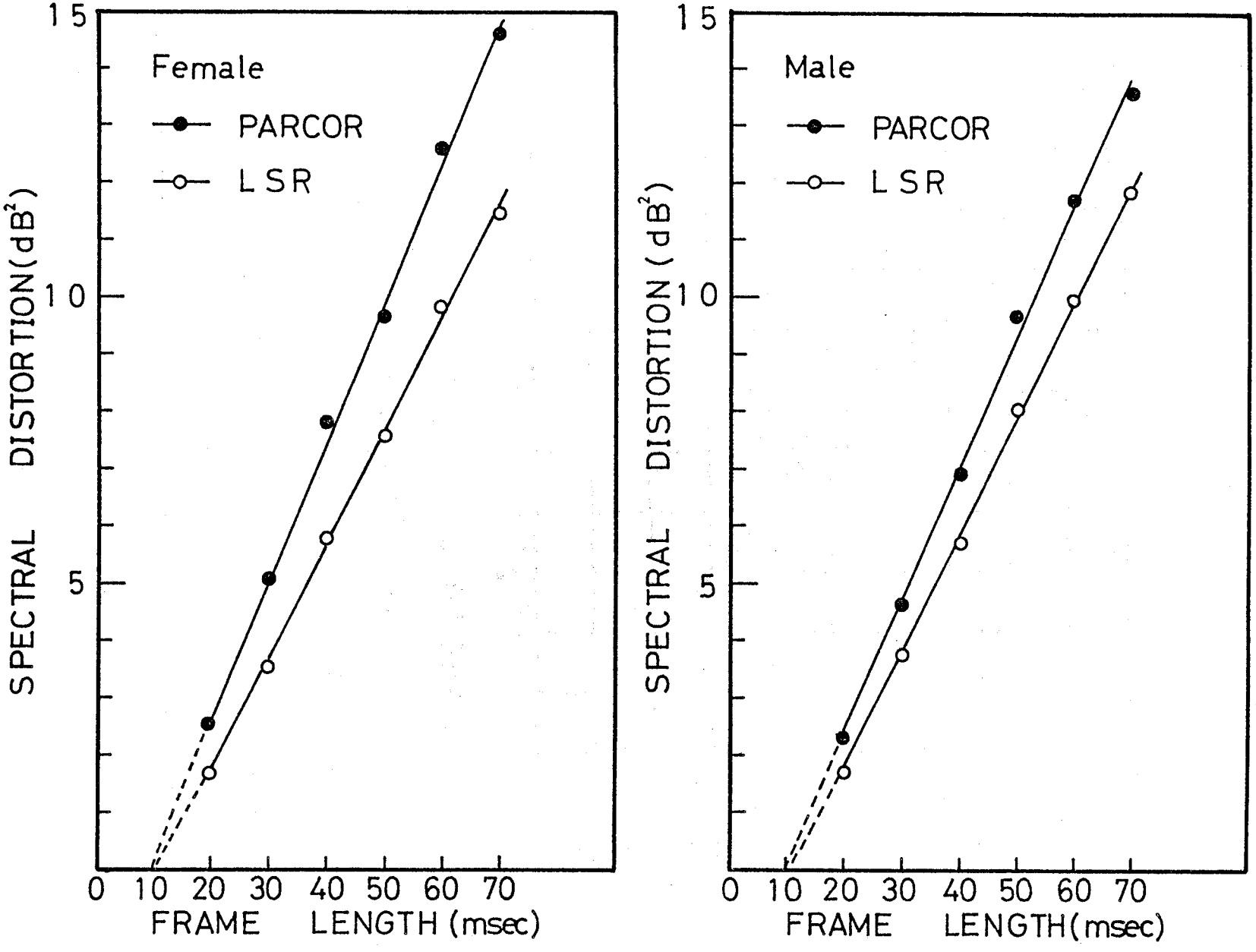


Fig. 3.15 Time spectral distortion versus frame period.
(a) male (b) female

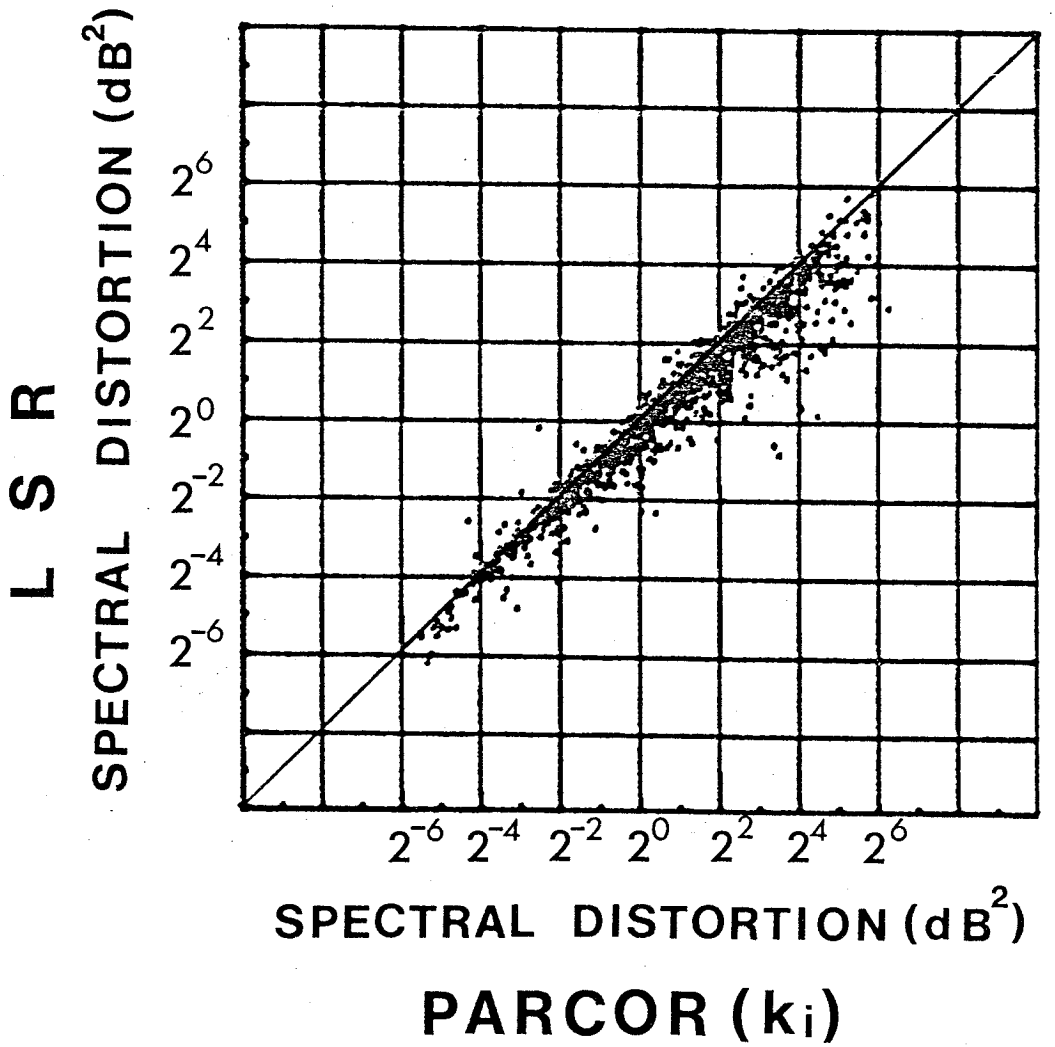


Fig. 3.16 Spectral distortion correlation at the same frame.

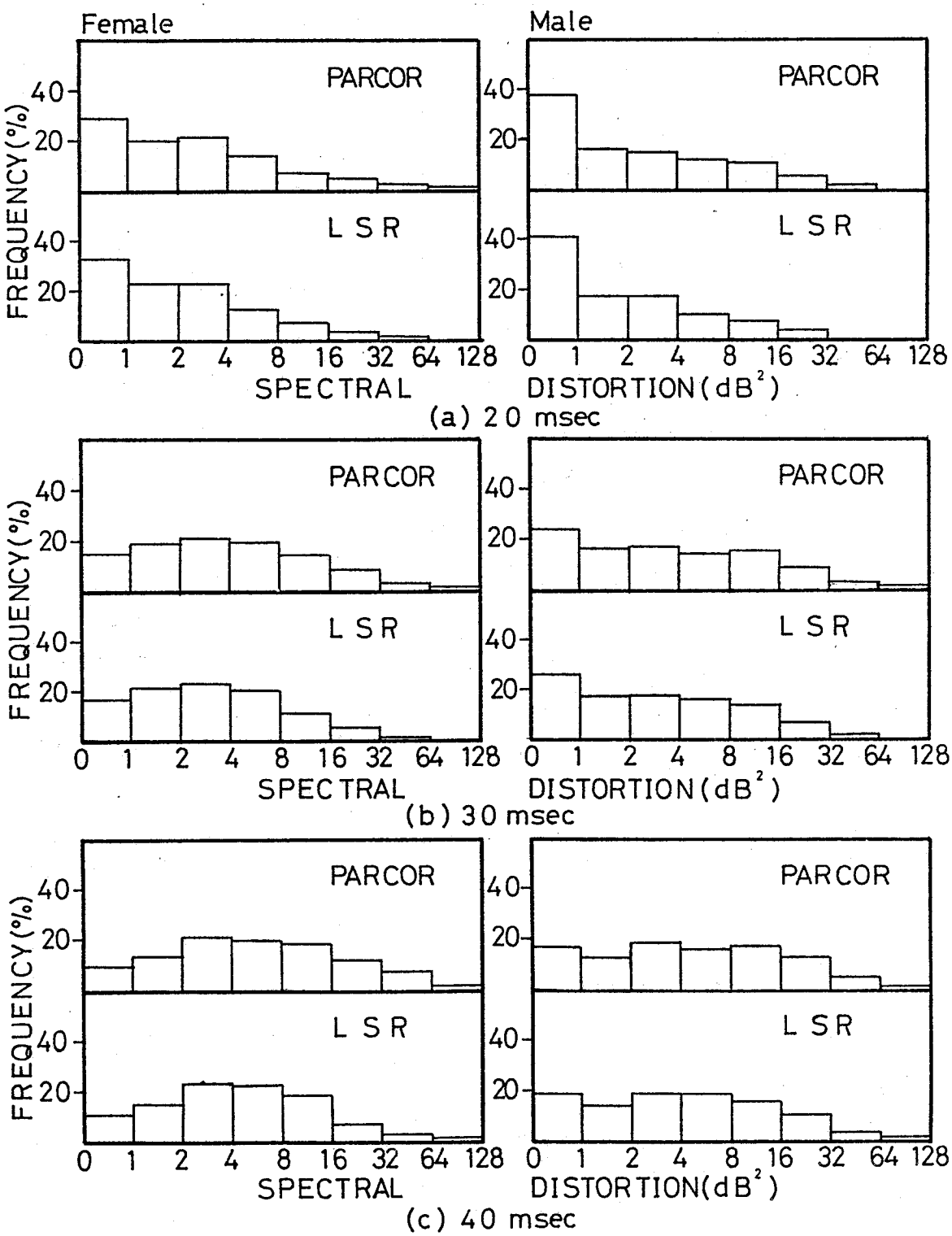


Fig. 3.17 Time spectral distortion distribution for each interpolated frame.

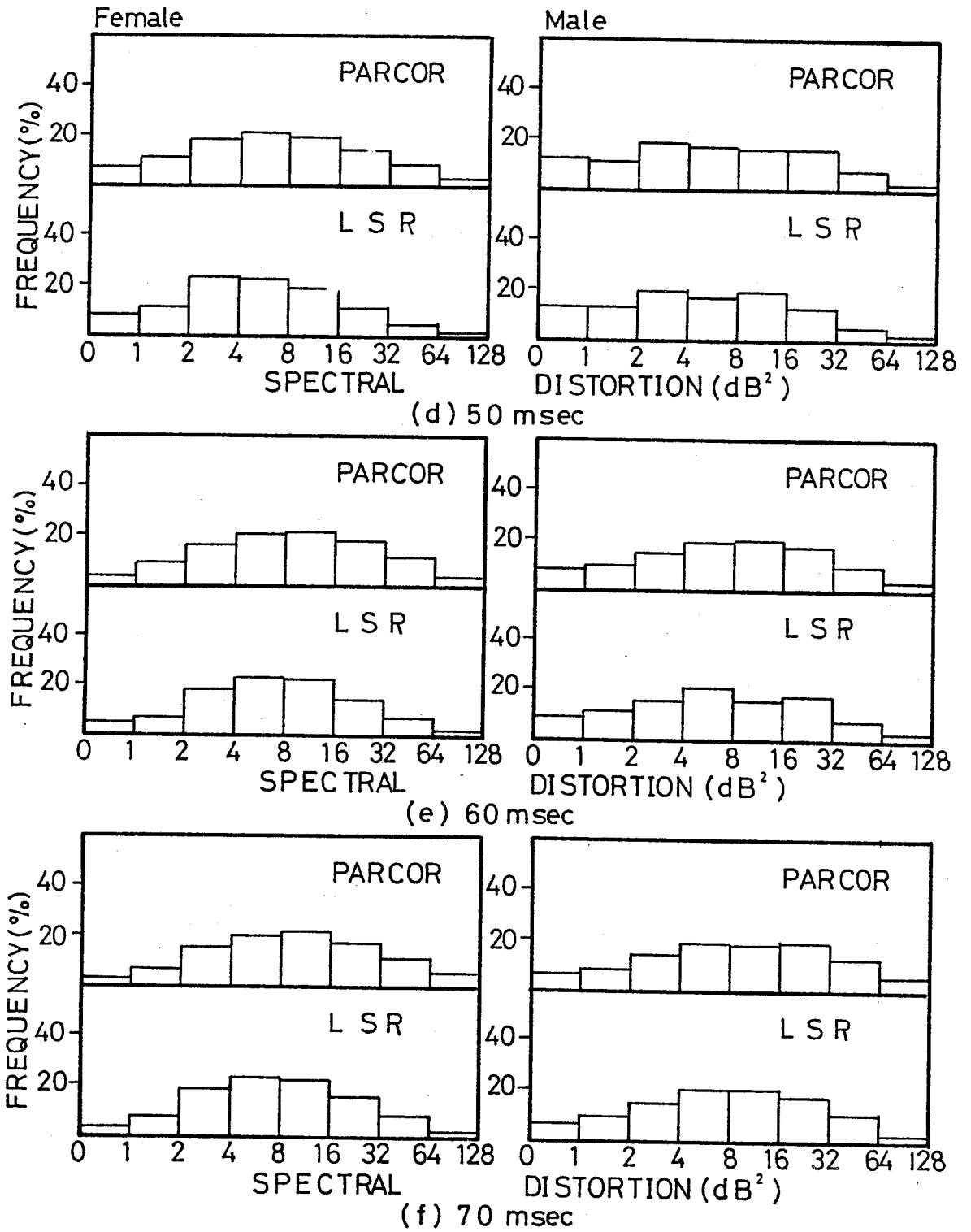
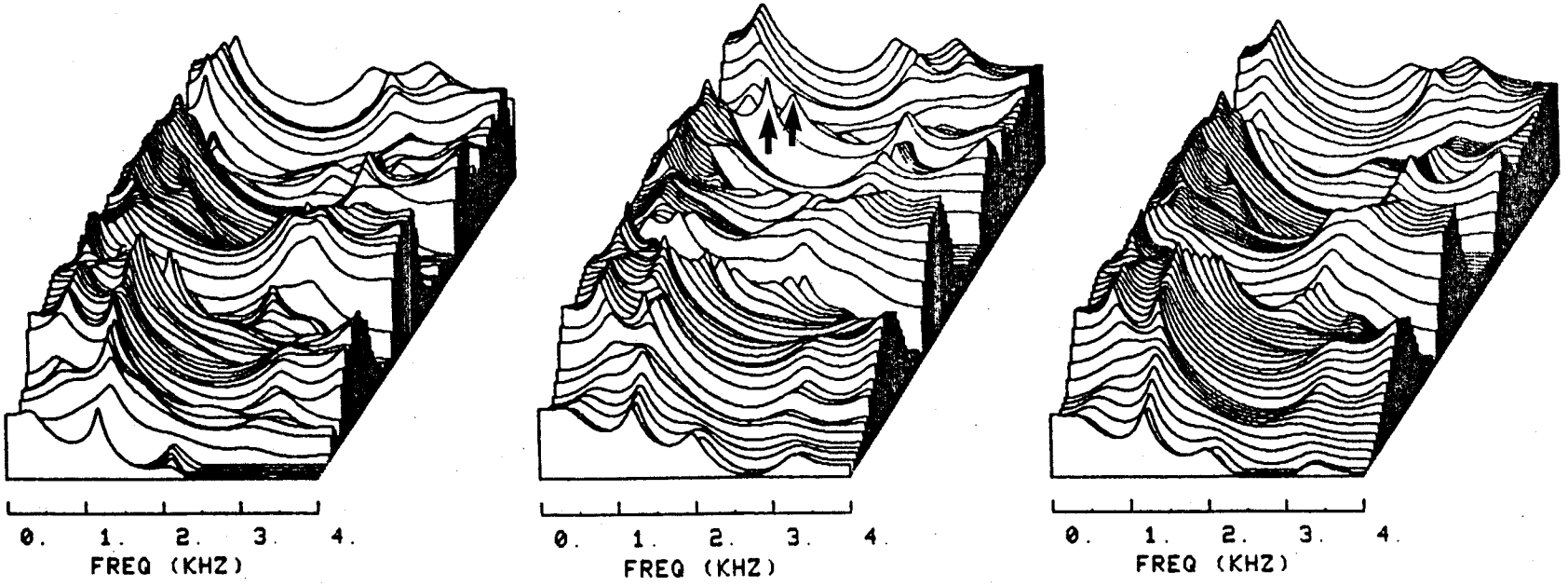


Fig. 3.17 (continuation)



(A) ORIGINAL (B) PARCOR (C) L S R

Fig. 3.18 Spectral envelope transition by interpolated parameters.

CHAPTER 4 LINE SPECTRUM PAIR (LSP) SPEECH ANALYSIS-SYNTHESIS SYSTEM

4.1 Introduction

The PARCOR speech analysis synthesis system is one of several efficient methods to compress a speech signal. However, the PARCOR system has a limitation in regard to data compression. The PARCOR synthetic speech quality rapidly becomes poor at rates under 4.8 kbps. There are two main reasons: first, in parameter quantization, four to eight bits are required for PARCOR coefficients [38]. Second, the synthesized speech time spectral distortion rapidly increases as the parameter refreshing period is lengthened.

On the other hand, the LSR parameters explained in the previous section have good interpolation property. In this system, however, it is necessary to calculate PARCOR coefficients again at the receiver from LSR parameters. It is impossible to synthesis a speech signal from LSR parameters directly.

This chapter describes a new approach to speech analysis synthesis which covers the defects in the PARCOR system. The approach, called the line spectrum pair (LSP) method, also exploits the all pole modeling of speech. LSP parameters are interpreted as one of the LPC parameters in the frequency domain, such as LSR parameters. This approach involves the excellent LSR parameter properties. Moreover, it can synthesize speech directly.

4.2 LSP Speech Analysis [53],[54]

A parameter set $(\omega_1, \theta_1, \omega_2, \theta_2, \text{ --- }, \omega_{p/2}, \theta_{p/2})$ in the frequency domain is introduced as another equivalent set to a set of linear predictive coefficients $(\alpha_1, \alpha_2, \text{ ------ }, \alpha_p)$. The new parameters are referred to as "line spectrum pair (LSP) " parameters.

Let $1/A_p(z)$ be a stable all pole digital filter in Eq. (2.4), and define the following (p+1)-th order polynomials. $A_p(z)$ follows the relations (2.32) and (2.33) and also follows, (3.20) and (3.21).

$$P(z) = A_p(z) - z^{p+1} A_p(1/z)$$

$$= 1+(\alpha_1 - \alpha_p)z+ \text{ ----- } +(\alpha_p - \alpha_1)z^p - z^{p+1} \qquad (4.1)$$

$$Q(z) = A_p(z) + z^{p+1} A_p(1/z)$$

$$= 1+(\alpha_1 + \alpha_p)z+ \text{ ----- } +(\alpha_p + \alpha_1)z^p + z^{p+1} \qquad (4.2)$$

These polynomials hold the following theorem.

(Theorem)

All zeros of P (z) and Q (z) lie on the unit circle and they alternate. This theorem is proved by introducing a rational function of z. (Appendix 1) It was proved that this property relates the LSP synthesis filter stability, too [55].

Both P (z) and Q (z) have only real coefficients. Therefore, if $e^{j\omega}$ is one of their zeros, $e^{-j\omega}$ is also one of their zeros. Then, we get an expression of P (z) and Q (z) using parameters (ω_i, θ_i).

(1) In case p is even

$$P\ (z) = (\ 1 - z\ ) \prod_{i=1}^{P/2} (1 - 2\cos\omega_i z + z^2\ ) \qquad (4.3)$$

$$Q\ (z) = (\ 1 + z\ ) \prod_{i=1}^{P/2} (1 - 2\cos\theta_i z + z^2\ ) \qquad (4.4)$$

(2) In case p is odd

$$P\ (z) = (\ 1 - z^2\ ) \prod_{i=1}^{(P-1)/2} (1 - 2\cos\omega_i z + z^2\ ) \qquad (4.5)$$

$$Q\ (z) = \prod_{i=1}^{(P+1)/2} (1 - 2\cos\theta_i z + z^2\ ) \qquad (4.6)$$

From Eqs. (4.1) and (4.2), $A_p(z)$ is expressed by $P\ (z)$ and $Q\ (z)$

$$A_p(z) = (P\ (z) + Q\ (z))/2 \qquad (4.7)$$

As stated above, consequently a set of parameters $(\omega_i\ ,\ \theta_i\ )$ is derived from $A_p(z)$ and $A_p(z)$ is also reconstructed in terms of $(\omega_i\ ,\ \theta_i\ )$. Then the set $(\alpha_i\ )$ is also equivalent to a set $(\omega_i\ ,\ \theta_i\ )$, similarly to $(\ k_i\ )$.

It is referred to LSP analysis of speech signal to analyze speech signal and extract LSP parameters from it. A considerable amount of computation is reduced by an elegant algorithm to get roots of $P\ (z)$ and $Q\ (z)$, exploiting the fact that both have complex conjugate roots. Parameter loci of $(\omega_i\ ,\ \theta_i\ )$ are shown in Table 4.1, according to the number of the analysis order p.

LSP parameters, ω_i and θ_i, are converted to f_i and g_i, whose dimension is in kHz, via the relations, $f_i = \omega_i / 2\pi\Delta T$ and $g_i = \theta_i / 2\pi\Delta T$.

4.3 LSP Speech Synthesis

The LSP all pole filter is a digital filter, whose transfer function is identical to that for the LPC or PARCOR all pole filter $1/A_p(z)$. LSP speech synthesis is performed by replacing the LPC all pole filter with this filter. The all pole digital filter is realized, as shown in Fig. 2.1, by a digital filter with a negative feedback loop, whose transfer function is $A_p(z)=1$.

$$H(z) = \frac{\sigma}{A_p(z)} = \frac{\sigma}{1 + (A_p(z) - 1)} \tag{4.8}$$

Identical transfer function $A_p(z)-1$ is also realized directly using LSP parameters,

(1) In case p is even

$$\begin{aligned} A_p(z)-1 &= [\,(P\ (z)-1) + (Q\ (z)-1)\,]/2 \\ &= z\,[(a_1+z) + \sum_{i=1}^{P/2-1} (a_{i+1}+z) \prod_{j=1}^{i} (1 + a_j z + z^2) \\ &\quad - \prod_{j=1}^{P/2} (1 + a_j z + z^2) \\ &\quad + z\,[(b_1+z) + \sum_{i=1}^{P/2-1} (b_{i+1}+z) \prod_{j=0}^{i} (1 + b_j z + z^2) \\ &\quad + \prod_{j=1}^{P/2} (1 + b_j z + z^2) \end{aligned} \tag{4.9}$$

(2) In case p is odd

$$
\begin{aligned}
A_p(z)-1 &= [\,(P\ (z)-1) + (Q\ (z)-1)\,]/2 \\
&= z\,[(a_1+z) + \sum_{i=1}^{(P-3)/2} (a_{i+1}+z) \prod_{j=1}^{i} (1 + a_j z + z^2) \\
&\quad - z \prod_{j=1}^{(P-1)/2} (1 + a_j z + z^2) \\
&\quad + z\,[(b_1+z) + \sum_{i=1}^{(P-2)/2} (b_{i+1}+z) \prod_{j=1}^{i} (1 + b_j z + z^2)
\end{aligned}
\tag{4.10}
$$

where $a_i = -2\cos\omega_i$, $b_i = -2\cos\theta_i$.

An example of an all pole digital filter, which has the above transfer function in the negative feedback loop, is constructed as shown in Figs. 4.1 and 4.2.

Figures 4.1(b) and 4.1(d) are equivalent transformations of Fig. 4.1(a) and 4.1(c), respectively. Generation of one sample output requires p multiplications and (3p+1) additions and subtractions.

4.4 Physical Meaning of LSP

LSP parameters have a clear physical meaning. If the vocal tract characteristics can be expressed by $1/A_p(z)$, the vocal tract is modeled as a non-uniform section area acoustic tube, consisting of p equally long sections mentioned in Chapter 3. The acoustic tube is open at the terminal correspondent to the lips, and each section is numbered from the lips. Mismatching between adjacent sections n and n+1 causes wave propagation reflection. The reflection coefficients are equal to

the n-th PARCOR coefficients k_n . Section p+1, corresponding to the glottis, is terminated by matched impedance. The signal excitation applied to the glottis drives the acoustic tube.

The PARCOR lattice filter is regarded as an equivalent digital filter to this acoustical model, so any speech signal can be modeled according to the acoustic tube model through PARCOR analysis. In ordinary PARCOR analysis, the boundary condition at the glottis is impedance matched. Now consider a pair of artificial boundary conditions, where the acoustic tube is completely closed or open at the glottis. These conditions correspond to a pair of extreme values of artificially extended PARCOR coefficients, namely $k_{p+1} = 1$ and $k_{p+1} = -1$, respectively. Under these conditions, $A_p(z)$ should be noticed to be identical to $P(z)$ or $Q(z)$, respectively. The acoustic tube becomes lossless, and then the transfer function displays line spectrum structure at frequencies $\omega_1, \theta_1, \omega_2, \cdots\cdots, \omega_{p/2}, \theta_{p/2}$.

Consequently, LSP parameters means frequencies at which the acoustic tube shows line spectrum structure under a pair of extreme artificial boundary conditions: complete opening and complete closure at the glottis. The term LSP (line spectrum pair) is presented reflecting such a physical interpretation. Moreover, LSP is interpreted as a kind of pulse density modulation of the power spectrum. Using Eqs. (4.8), (4.9) and (4.10), the following representation of the power spectrum $S(\omega)$ is produced by the LSP all pole filter.

(1) In case p is even

$$S(\omega) = \sigma^2 / \left| A_p(e^{-j\omega}) \right|^2$$

$$= 2^{P+2}\sigma^2/(\cos^2\frac{\omega}{2}\prod_{i=1}^{P/2}(\cos\omega-\cos\theta_i) + \sin^2\frac{\omega}{2}\prod_{i=1}^{P/2}(\cos\omega-\cos\omega_i)) \quad (4.11)$$

(2) In case p is odd

$$S(\omega) = 2^{P+2}\sigma^2/(\prod_{i=1}^{(P+1)/2}(\cos\omega-\cos\theta_i) + \sin^2\frac{\omega}{2}\prod_{i=1}^{(P-1)/2}(\cos\omega-\cos\omega_i)) \quad (4.12)$$

Assume that ω_i and θ_i are close, and that ω is near to them. Then, $S(\omega)$ takes a large value. Conversely, the necessary condition under which $S(\omega)$ has strong resonance at a certain frequency is that more than two LSP parameters concentrate near them. Therefore, LSP is interpreted as one representation of an all pole filter by means of density location of p discrete frequencies, instead of line spectrum amplitude in pseudo formant vocoder.

Figure 4.2 shows the relationship between LPC spectrum envelope and LSP parameters location of sound /a/ in continuous speech, as shown Fig. 4.2(a). This is a typical spectral pattern for Japanese vowel /a/ sound uttered by a female speaker.

Figures 4.3 and 4.4 show sonograms of female and male speech signals ' bakuon ga ginsekaino koogen ni hirogaru' (phoneme-balanced meaningless sentence). LSP parameters are plotted over them. If there are strong resonances, two or three LSP parameters concentrate. If there are no strong resonances, for example, during a silence duration period, LSP parameters are replaced at nearly equal intervals versus the frequency axis.

4.5 Statistical Properties of LSP Parameters

In a speech analysis synthesis system using LSP parameters, synthesized speech quality impairment is caused by coding the parameters. In this section, parameter properties in relation to parameter distribution, quantization and interpolation are examined statistically. These results are compared with those for PARCOR and LSR parameters.

4.5.1 Parameter Distribution

Speech samples and analysis condition used in the experiments are listed in Table 4.2. LSP parameters distributed as shown in Fig. 4.5. The distribution range for each LSP parameter is from 500 Hz to 1000 Hz. These values are from 1/8th to 1/4th of the full range. At any frame period, LSP parameter are ordered according to the theorem, but statistically adjacent parameters are overlapped.

4.5.2 Quantization Characteristics

In speech data compression using LSP parameters, it is very important to achieve optimum quantization for bits saving.

The LSP analysis synthesis system is based on speech spectrum preservation principle, such as the PARCOR system. Therefore, the transmission distortion is to be discussed as a function of spectral distortion. To clarify the relation between information rate and

spectral distortion by quantization, a spectral sensitivity analysis technique is applied to LSP parameters to realize optimum quantization.

A spectral sensitivity for each LSP parameter is defined in the same manner as mentioned in Chapter 3. In order to get spectral sensitivity, speech samples uttered by ten speakers, five male and five female, were used in the experiment. Each sentence uttered is a phonetically balanced Japanese sentence with four second duration.

Experimental results of spectral sensitivities are summed up in Table 4.4. Individual spectral sensitivities are uniform and the ratio of minimum value to maximum is small, about two times, in comparison with that for PARCOR coefficients.

Spectral distortion by quantization for each parameter is shown in Fig. 4.6.

Two quantization methods for LSP parameters are evaluated by spectral distortion measure in order to compare PARCOR coefficients and LSR parameters. Spectral distortion between a non-distorted spectrum and a distorted spectrum due to quantization is defined by a mean spectral distance on a dB^2 scale. Spectral distortion is represented by the averaged distortion value obtained every frame period. Two quantization methods are :

(1) Quantization -A: Linear quantization of f_i and k_i over the range (min 1 % tail value, max 1 % tail value) and uniform bits allocation to each parameter.

(2) Quantization -B: Linear quantization of f_i and k_i over the range (min 1 % tail value, max 1 % tail value) and non-uniform bits allocation to each parameter, taking into account spectral sensitivity.

The parameter quantization range is selected, excluding very few occurences to quantize effectively.

Experiment results are shown in Fig. 4.7. In this figure, lines A, A', B, B', C and C' represent :

A ; PARCOR, Quantization-A

A'; PARCOR, Quantization-B

B ; LSR , Quantization-A

B'; LSR , Quantization-B

C ; LSP , Quantization-A

C'; LSP , Quantization-B

In Quantization-A, the information rate per frame to restrict the mean spectral distortion under one dB, needs to be 50 bps in the PARCOR system and 35 bps in the LSP system.

The difference in these information rates is caused by wide ranges and spectral sensitivity non-uniformity of PARCOR coefficients. That is, PARCOR coefficients are distributed more widely than LSP parameters and lower degree coefficients are concentrated at nearly +1 or -1. These spectral sensitivity values are much higher than that for higher degree coefficients. In the case of LSP, each parameter is quantized at four bits per frame, the mean spectral distortion is less than one dB. On the other hand, using Quantization-B, the quantization bits decrease from six to nine bits per frame were reduced in PARCOR coefficients, but only one or two bits per frame reduced in LSP parameters, in comparison with Quantization-A at one dB spectral distortion.

This results mean that uniform bit allocation is nearly optimum in LSP parameters. Quantization property for LSP parameters is also superior to that for LSR parameters.

The characteristics of LSP parameters quantization properties are summarized as :

(1) The ratio of information rate (LSP/PARCOR) to suppress spectral distortion under one dB is 0.6 (35/50) in the uniform bit allocation and 0.8(34/42) in the non-uniform bit allocation.

(2) For LSP parameters, if each parameter is quantized four bits equally, the spectral distortion is less than one dB on an average. It does not need the non-uniform bit allocation and parameter non-linear conversion, such as for PARCOR coefficients.

4.5.3 Interpolation Characteristics

Synthesized speech using the LPC parameter generates a kind of spectral distortion, called time spectral distortion, by sampling LPC parameters at more than 10 mseconds intervals for information rate saving. Therefore, in the speech analysis synthesis system, parameter interpolation is effective in smoothing synthesized speech roughness. In this section, interpolation properties are discussed, using time spectral distortion between an interpolated spectrum and an actual spectrum.

In the speech analysis synthesis system, spectral distortion is caused by quantization and interpolation. However, in this section, to clarify only interpolation properties, parameter quantization is not considered. The interpolated parameter selected is f_i in the case of LSP and is $\tanh^{-1} k_i$ in the case of PARCOR coefficients and linear interpolation is used. Speech samples used in this experiment are the same as those used in examining the quantization properties.

The analysis frame period is 10 mseconds. Parameter interpolation

periods are chosen as from 20 mseconds to 70 mseconds at every 10 msecond interval. Each LSP parameter is linearly interpolated from 20 mseconds to 70 mseconds and spectral distortion is calculated every 10 msecond period. The relation between spectral time distortion and parameter sampling frame period is shown in Fig.4.8. As is evident from this figure, the LSP parameters interpolation property is superior to that for PARCOR coefficients, as well as for LSR parameters. At the same spectral distortion, the sampling rate for LSP parameters is 75 % of that for PARCOR coefficients.

4.5.4 Comparison of Statistical Properties

LSP parameters statistical properties are summarized in Table 4.4, comparing PARCOR coefficients and LSR parameters. The information rate in the LSP system is proved to be about 60 % of that for the PARCOR system, through spectral distortion analysis.

4.6 Speech Quality Assessment for LSP Synthesized Speech [56]

As mentioned before, LSP parameters are superior to PARCOR coefficients in respect to quantization and interpolation properties as a function of spectral distortion. In order to clarify this experimental result subjectively, synthesized speech quality assessment using pair comparison was studied. The eight kinds of synthesized speech used for this test are shown in Table 4.5. Four of them are synthesized by PARCOR coefficients and the others are synthesized by LSP parameters. Speech data, bits allocation and other conditions for synthesized speech are shown in Table 4.6. To clarify only the

difference in spectral parameters (PARCOR or LSP), both synthesized speech categories have the same excitation parameters with the same bit allocations. Experimental results are shown in Fig. 4.9 and Fig. 4.10. Figure 4.9 shows the preference score for each sample and Fig. 4.10 shows the selective ratio at each information rate. The preference score for LSP synthesized speech is higher than that for PARCOR synthesized speech at the same information rate. Especially, the LSP to PARCOR preference score ratios are 7:3 at 4.8 kbps and 9:1 at 2.4 kbps. At 1.6kbps, the ratio becomes smaller than that at other information rates. In this case, the parameter sampling period is too rough for rapid speech spectral transition. Other parameter sampling techniques, for example, variable frame rate sampling (VFR) to be mentioned in Chapter 5 or vector quantization to be mentioned in Chapter 6, needed to be applied to achieve speech data compression with a lower bit rate.

4.7 Conclusions

(1) A new approach to speech analysis synthesis, based on line spectrum pair (LSP) representation of an all pole digital filter was introduced. Its principle and physical meaning were related.

(2) Characteristics for LSP parameters were examined experimentally as a function of spectral distortion. LSP parameters have better quantization and interpolation properties than PARCOR coefficients, or even LSR parameters.

(3) Optimum bit allocation was proposed through spectral sensitivity analysis. Spectral sensitivities for LSP parameters are uniform. Consequently uniform bit allocation is nearly optimum to quantize LSP parameters.

Four bits to each parameter is enough to compress the spectral distortion to under one dB. Moreover it only need 34 bps to achieve spectral distortion to under one dB, considering spectral sensitivity and distribution ranges for LSP parameters. This value is about 85 % of that for PARCOR coefficients.

(4) LSP parameters have good interpolation characteristics, better than those for LSR parameters. Consequently, the frame period can be expanded 1.25 times that for PARCOR.

(5) According to better quantization and interpolation properties, the 60 % information rate, on the basis of the PARCOR system, is enough to achieve the same spectral distortion as that in the PARCOR system.

(6) The fact that the LSP system achieves equivalent quality at 60 % bit rate, on the basis of the PARCOR system, was proved subjectively, too.

Table 4.1 Locations of zeros in polynomials P(z) and Q(z).

p	Polynomial	No. of Root		Roots Pattern 0 π
		Real	Complex	
Even	P	1	p/2	
	Q	1	p/2	
Odd	P	2	(p-1)/2	
	Q	0	(p+1)/2	

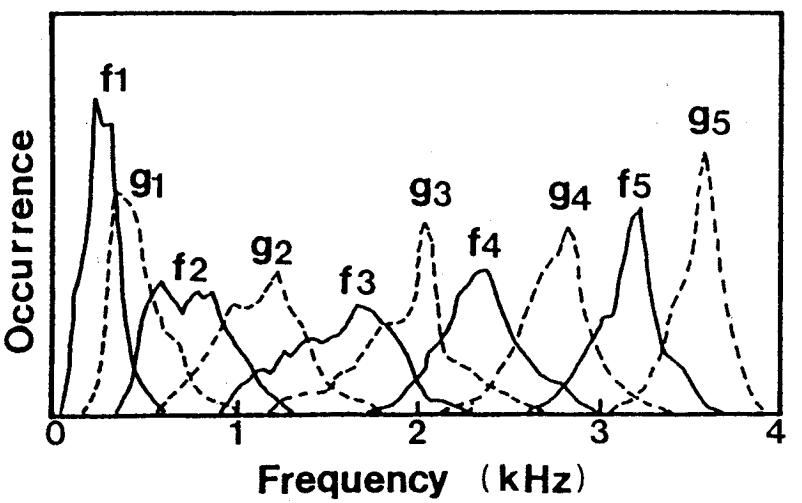
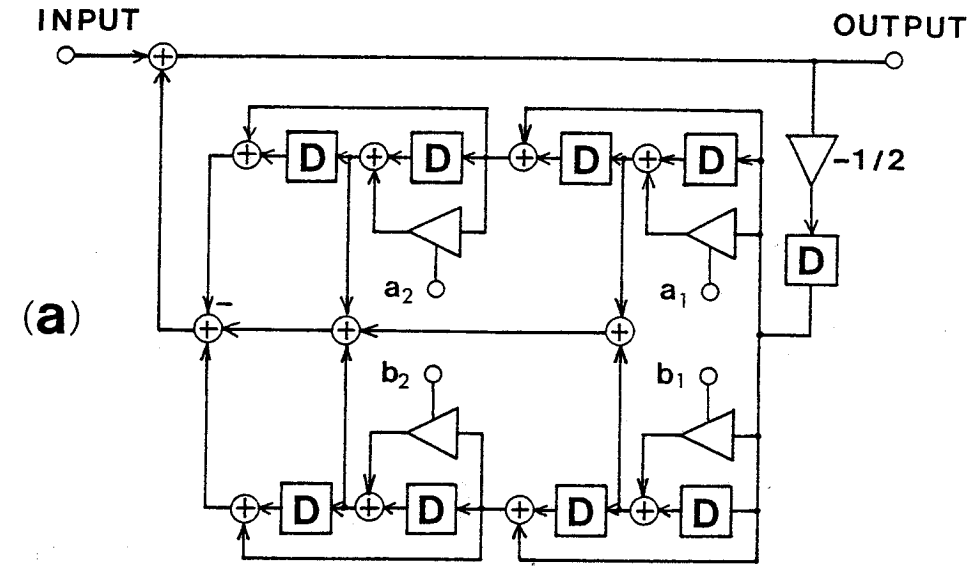
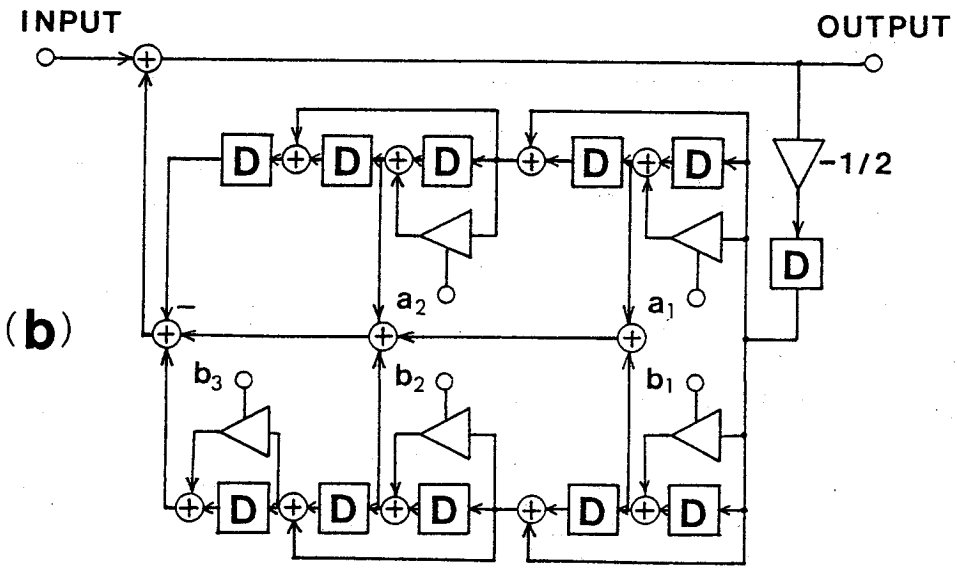


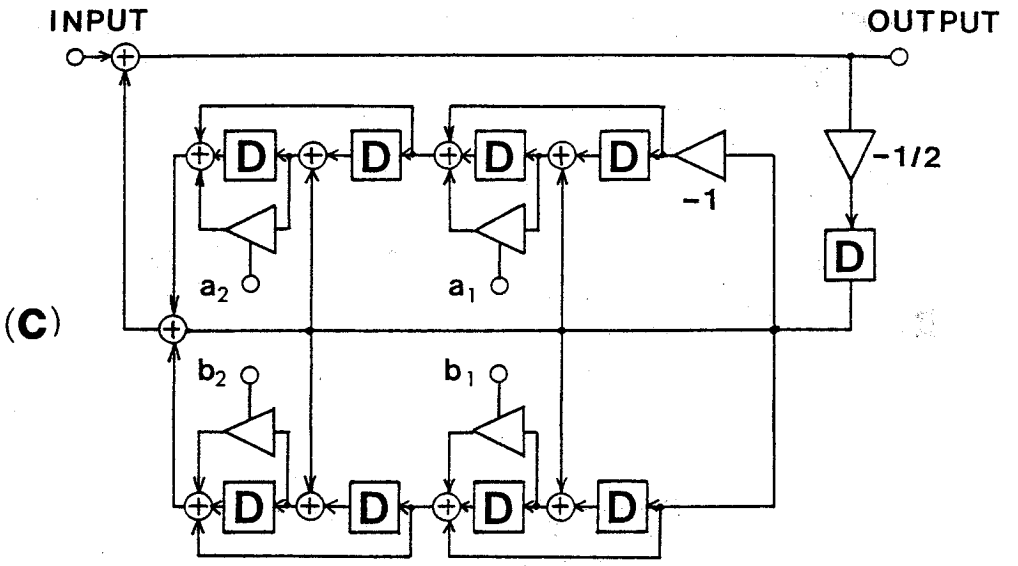
Fig. 4.5 LSP parameters distribution.

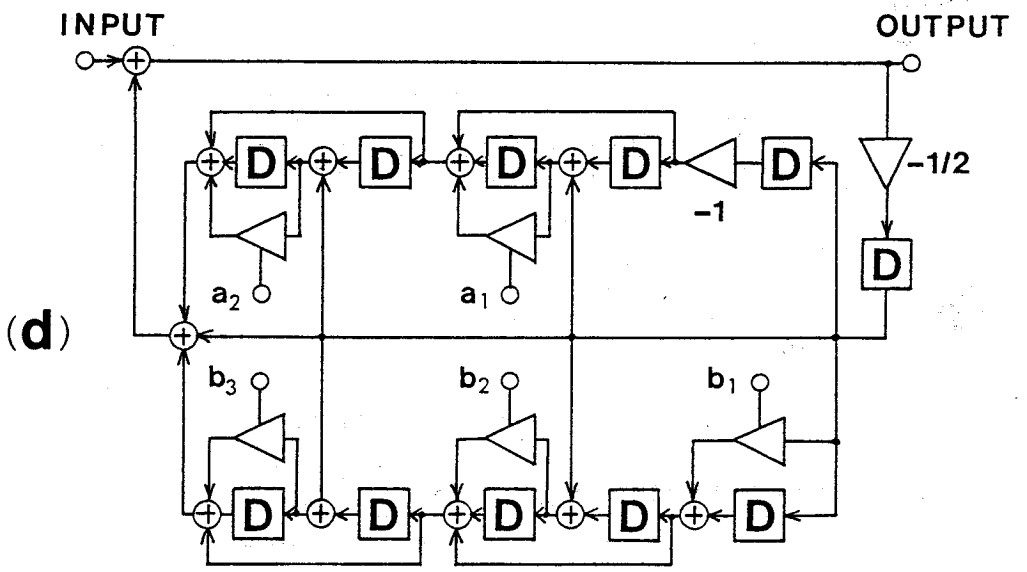




p is even

Fig. 4.1 LSP speech synthesis filters.





p is odd

Fig. 4.1 LSP speech synthesis filters (continuation).

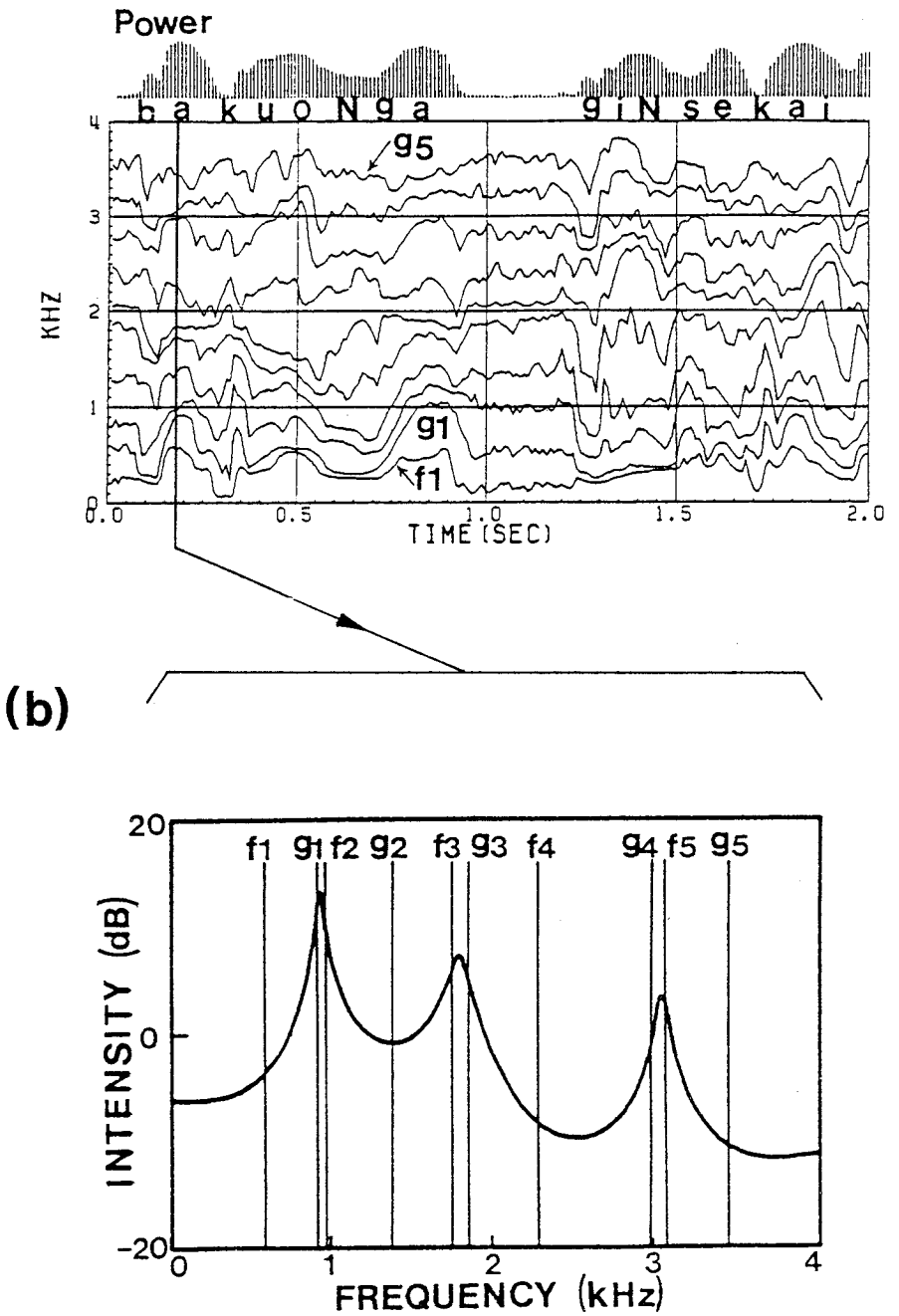


Fig. 4.2 Spectral envelope and LSP parameters locations.
(a) Point a extracted frame in utterance
(b) LSP parameters locations in the frame

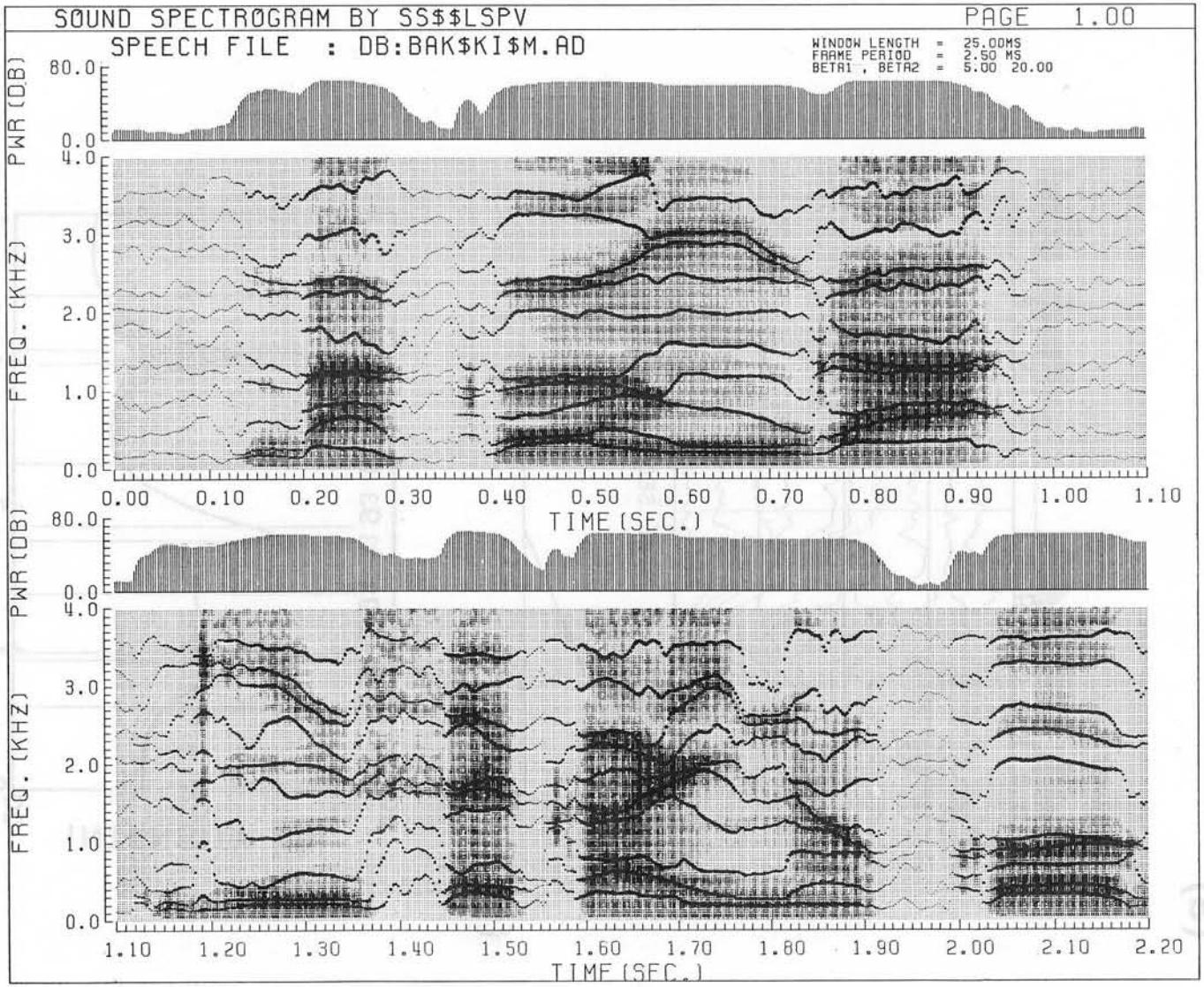


Fig. 4.3 Sonograms over plotted LSP parameters. Speech sample is uttered by a female speaker.

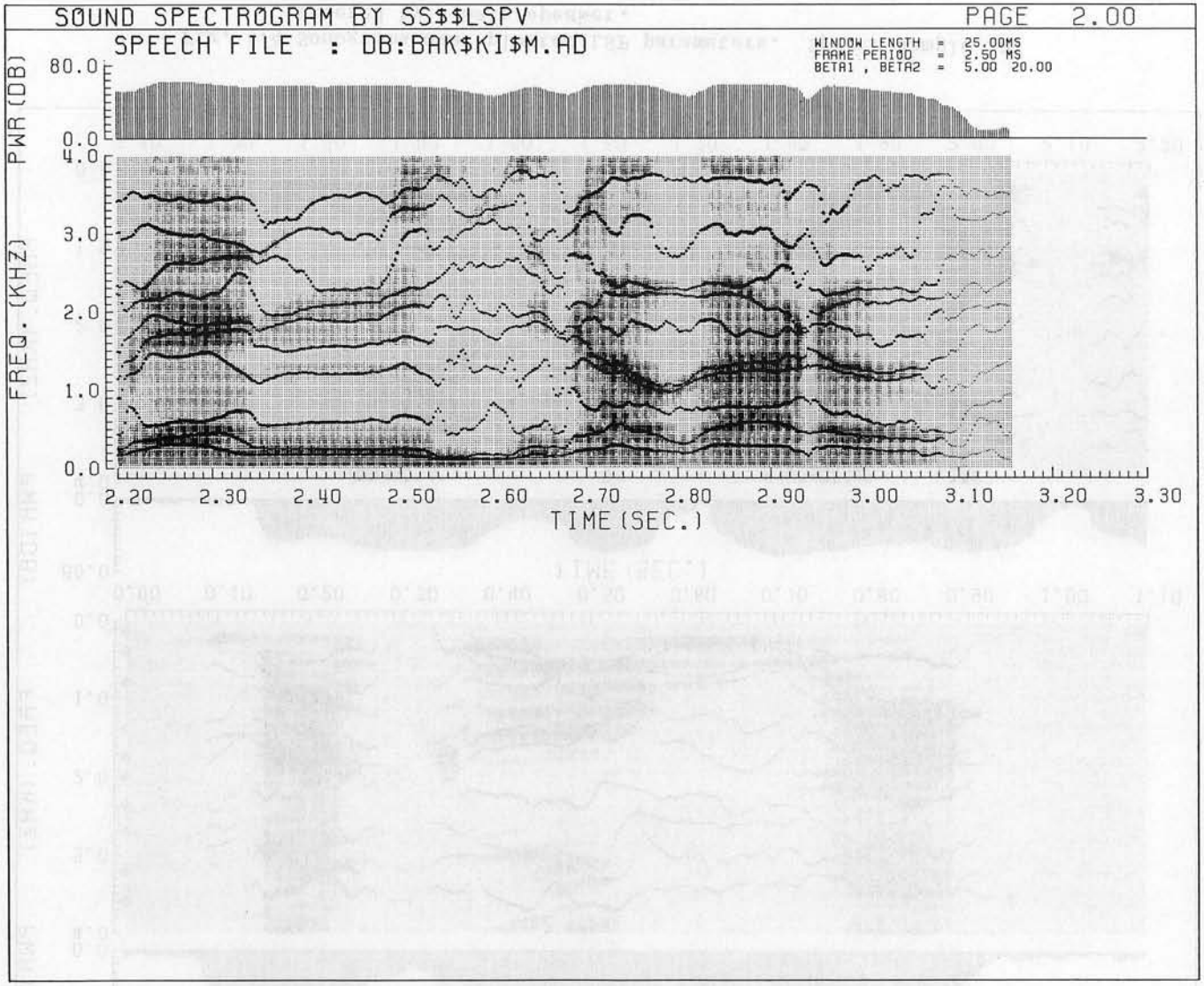


Fig. 4.3 (continuation)

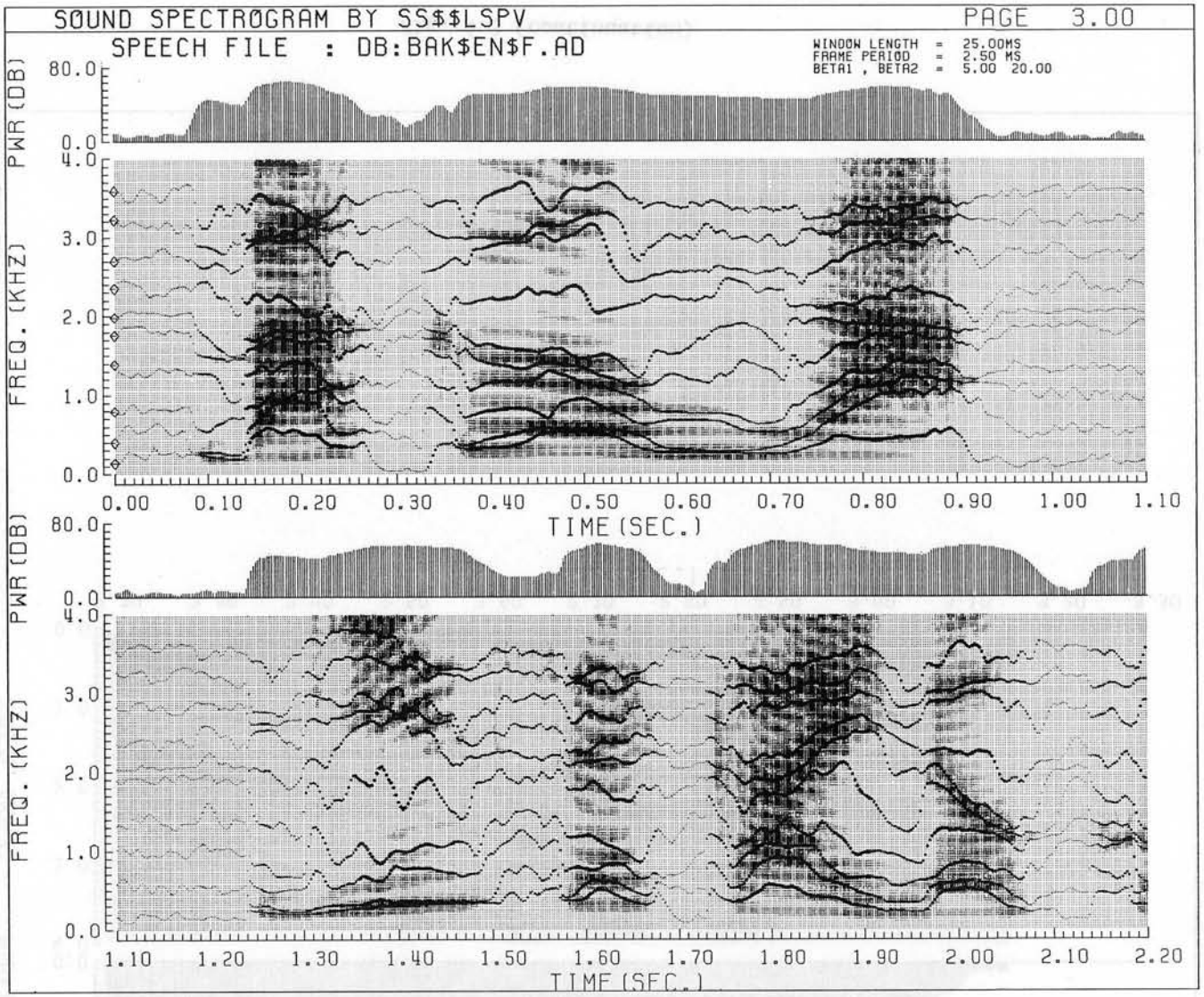


Fig. 4.4 Sonograms over plotted LSP parameters. Speech sample is uttered by a male speaker.

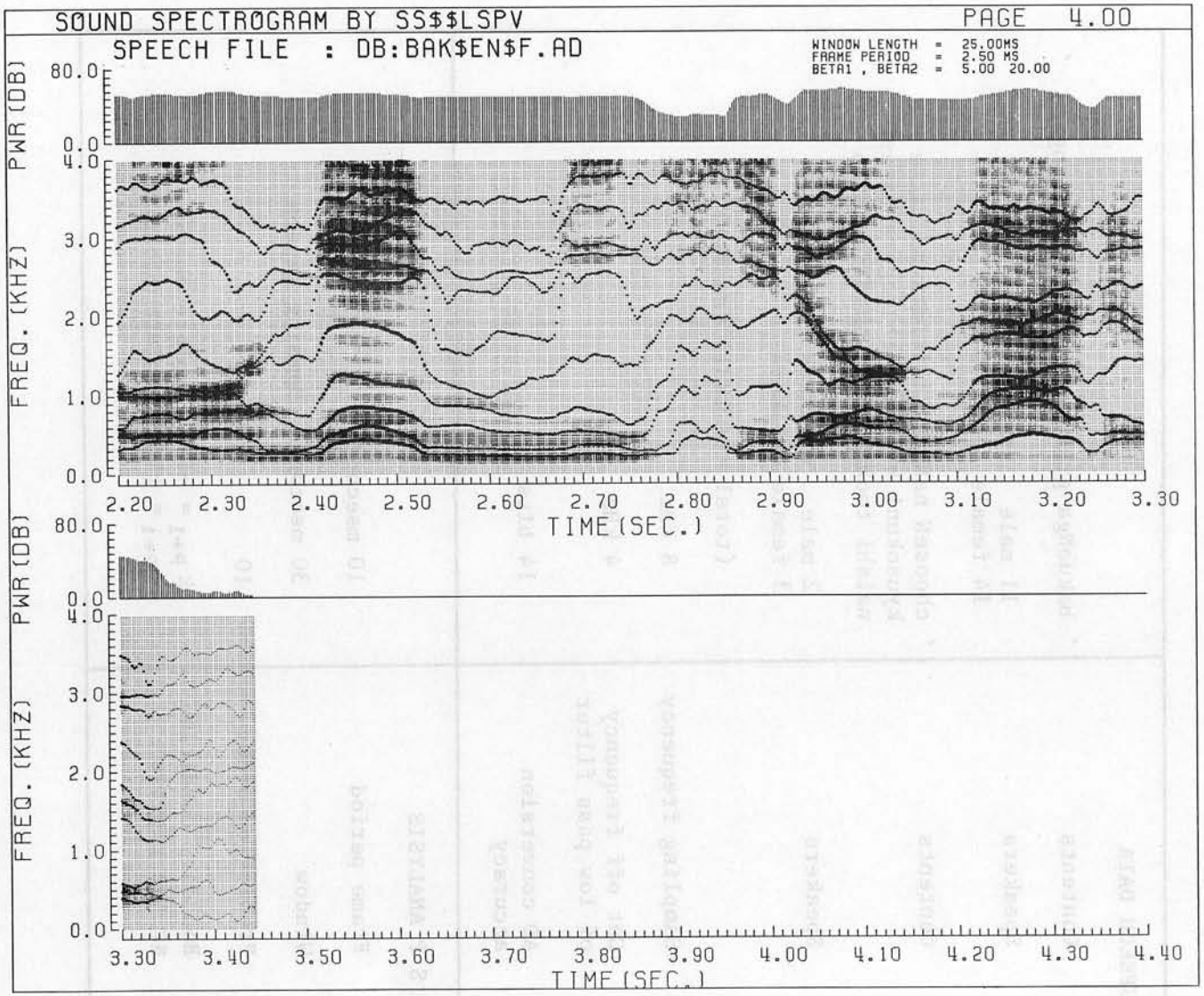


Fig. 4.4 (continuation)

Table 4.2 Experimental conditions.

SPEECH DATA	
Contents	' bakuoNga giNsekaino koogeNni hirogaru '
Speakers	11 male speakers 14 female speakers
Contents	' chooseN naNbuni teikiatsuga atte kyusokuni hattatsu shinagara higashi naishi toohokutooni susuNde imasu '
Speakers	2 male speakers 3 female speakers (total --- 12311 frames)
Sampling frequency	8 kHz
Cut off frequency of low pass filter	4 kHz
AD conversion accuracy	14 bits
L S P ANALYSIS	
Frame period	10 mseconds
Window	30 mseconds Hamming window
Prediction order	10
Boundary condition at the glottis	$k_{p+1} = -1$ (complete opening) $k_{p+1} = +1$ (complete closure)

Table 4.3 Spectral sensitivities for LSP parameters.

LSP Parameter	Spectral Sensitivity (dB/Hz)	LSP Parameter	Spectral Sensitivity (dB/Hz)
f1	0.022	g1	0.020
f2	0.013	g2	0.014
f3	0.015	g3	0.014
f4	0.012	g4	0.014
f5	0.012	g5	0.011

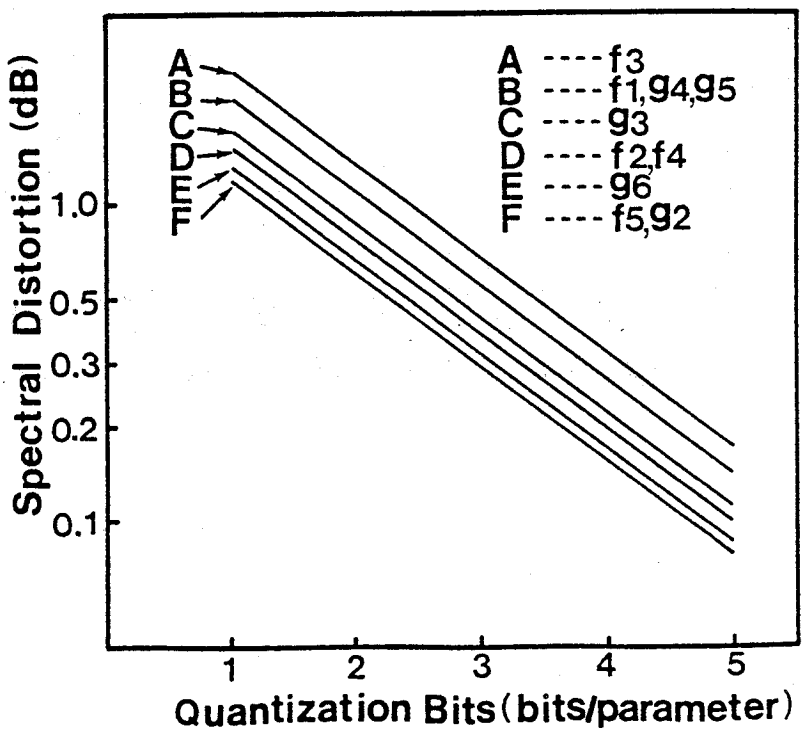


Fig. 4.6 Relationship between quantization spectral distortion and quantization bits per frame.

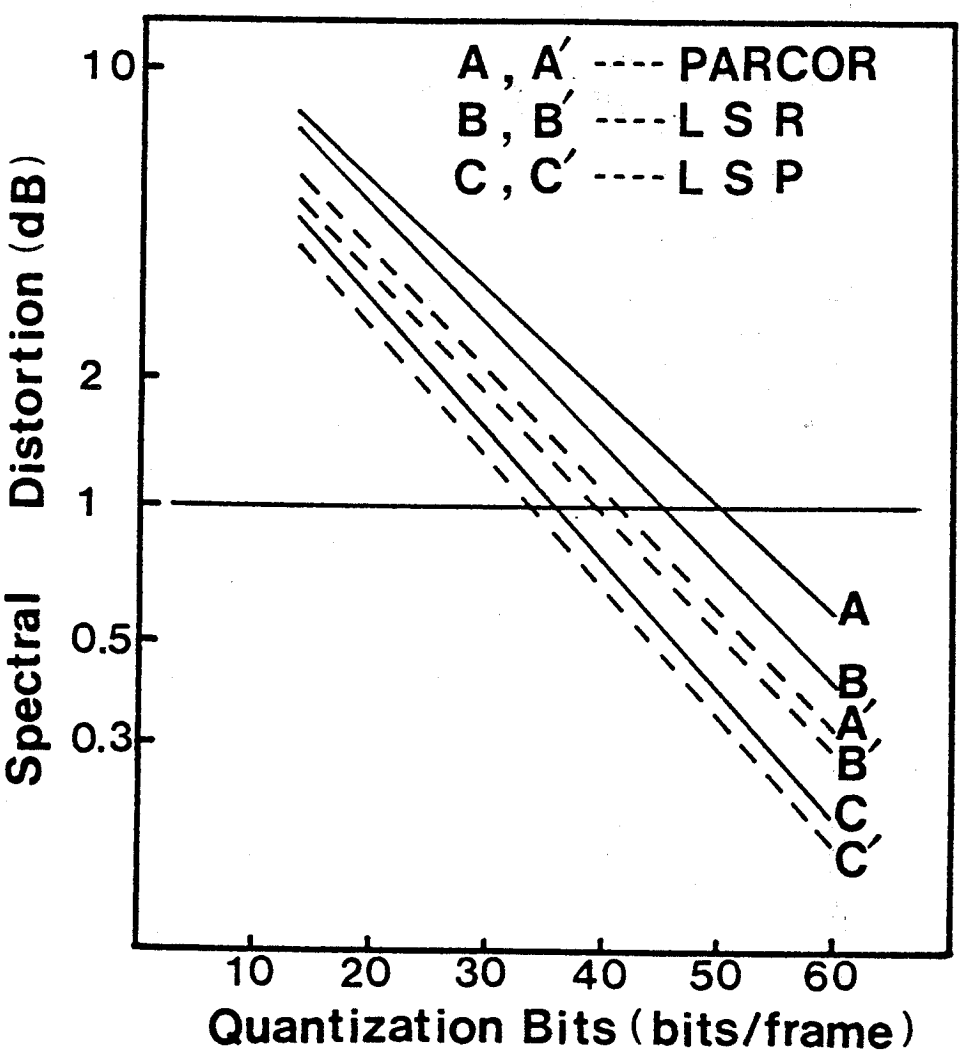


Fig. 4.7 Spectral distortion versus bit rate per frame.

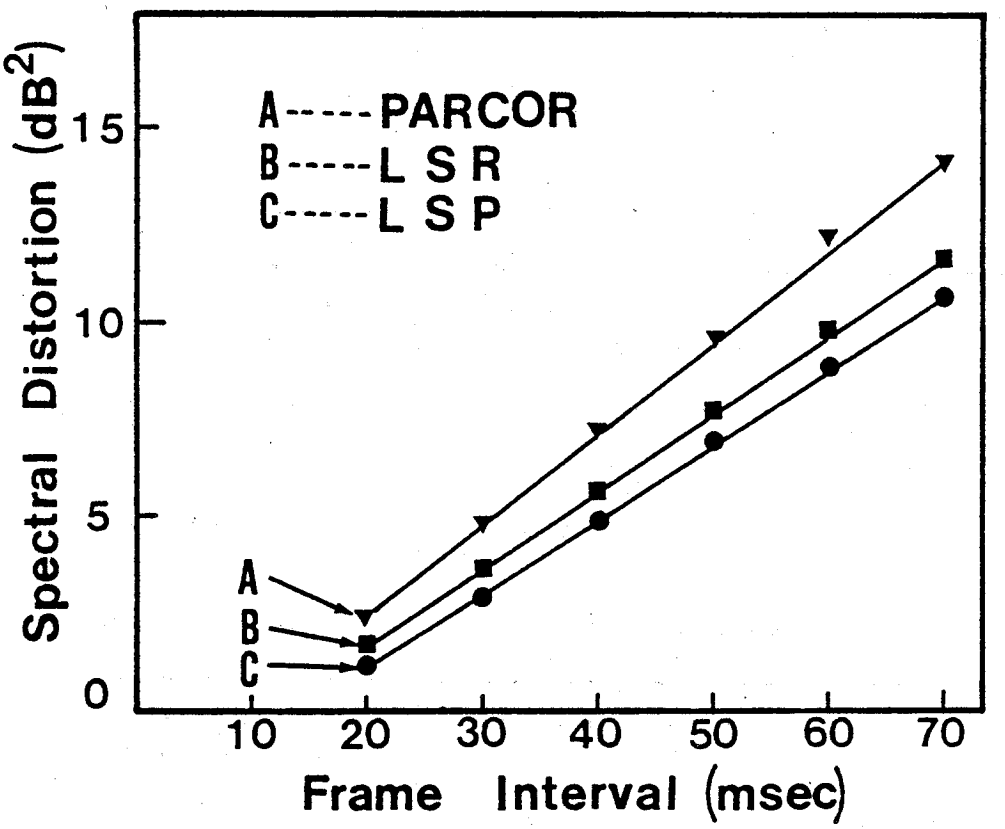


Fig. 4.8 Comparison between interpolation characteristics.

Table 4.4 Comparison of features of LSP, LSR and PARCOR parameters.

PARAMETER	MERIT	DEMERIT
PARCOR (k_i)	Stability condition $\leftrightarrow \lvert k_i \rvert < 1$ Direct extraction by lattice filter k_i is independent of analysis order Similar to formant parameters	Bad interpolation property Large varity in spectral sensitivity Parameter conversion is needed for synthesis
L S R (f_i ,m_i)	Uniform spectral sensitivity Good interpolation property Similar to formant frequency Uniform spectral sensitivity	Large calculation amount for interpolation f_i and m_i values depend on analysis order LSP parameters disorder occurs due to quantization with very few quantization bits
L S P (f_i ,g_i)	Good interpolation property Direct synthesis without parameter conversion Stability condition ↕ $f_1 < g_2 < \cdots < f_n < g_n$	f_i and g_i values depend on analysis order

Table 4.5 Speech samples for a pair comparison test.

No.	Synthesis Technique	Frame Period	Quantization	Bit Rate
1	L S P	5 msec	48 bits/frame (table 2)	9.6 kbits
2		10		4.8
3		20		2.4
4		30		1.6
5	PARCOR	5		9.6
6		10		4.8
7		20		2.4
8		30		1.6

Table 4.6 Speech synthesis conditions.

Speech Data		female speaker / bakuonga ginsekaino ··· /	
Sampling Frequency		8 kHz	
Synthesis Order		10	
Bits Allocation		Spectral Envelope	Excitation
	L S P	f_1 f_2 f_3 f_4 f_5 g_1 g_2 g_3 g_4 g_5 4 4 4 4 3 4 4 4 3 3	Pitch 6 Amplitude 5
	PARCOR	k_1 k_2 k_3 k_4 k_5 k_6 k_7 k_8 k_9 k_{10} 7 5 4 3 3 3 3 3 3 3	

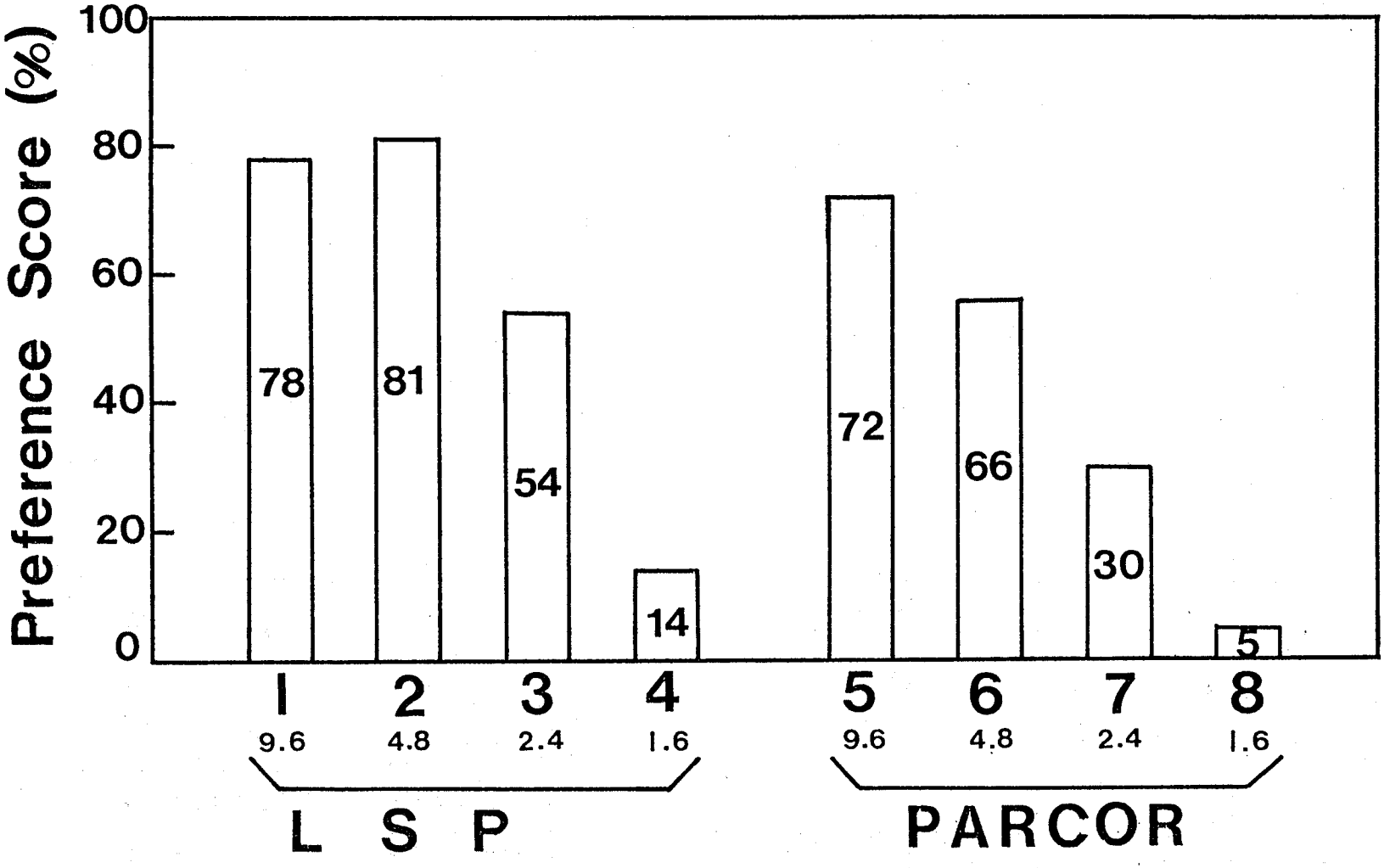


Fig. 4.9 Preference scores for each synthesized speech.

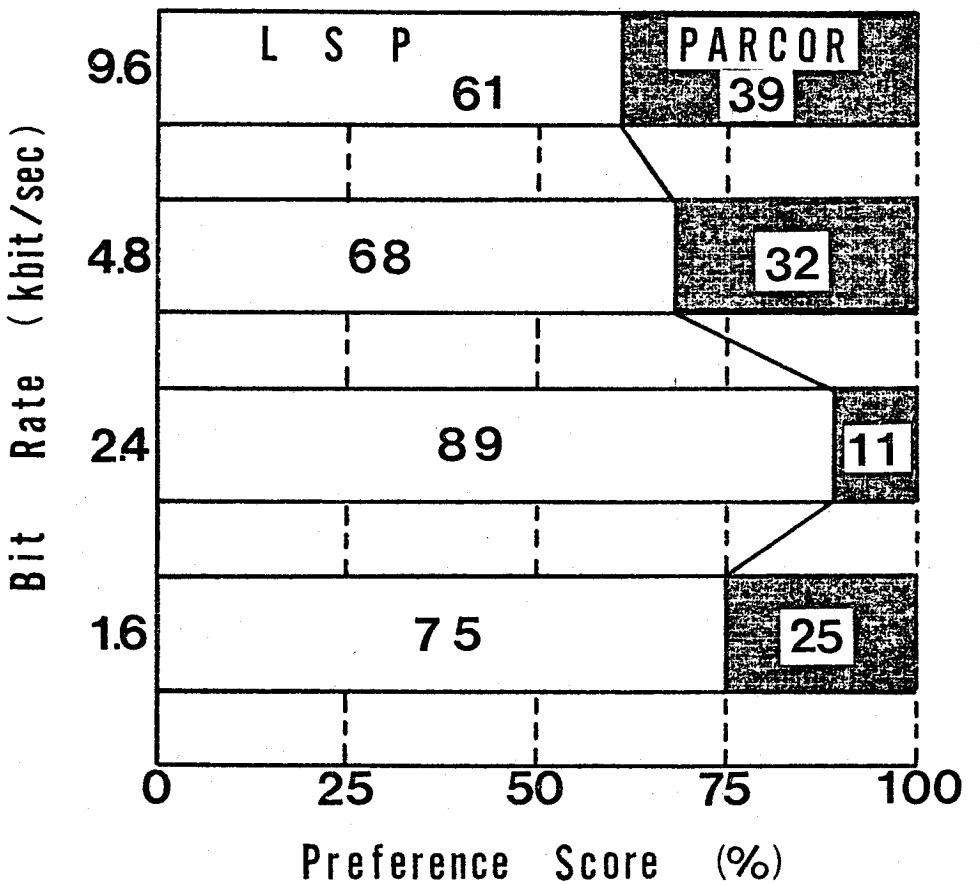


Fig. 4.10 Preference scores for each synthesized speech at the same information rate.

CHAPTER 5 VARIABLE FRAME RATE CODING USING LSR PARAMETERS

5.1 Introduction

Speech events generally occur at non-uniformly spaced time intervals, according to the utterance contents or articulator movements. Therefore, uniform speech parameters sampling is not efficient. Sampling at a variable rate in accordance with the changing speech signal characteristics is more efficient. In this chapter, parameter coding at non-uniformity spaced time intervals is presented. This technique has been called "Variable Frame Rate Transmission (VFR)" [25],[57],[58],[59]. The VFR transmission scheme in the pseudo formant vocoder is related to reducing the average transmission rate without speech quality degradation. Two VFR techniques are discussed as a function of time spectral distortion. Synthesized speech qualities are assessed by subjective measurement in comparison with conventional uniform sampling.

5.2 Comparison of Several Interpolation Methods

It is very important to deal with the speech signal redundancy in the time domain quantitatively. First, the time spectral distortion by several interpolation method is discussed.

The speech signal is analyzed every 10 mseconds and LSR parameters are extracted. Frame length was set from 20 mseconds to 70 mseconds, every 10 msecond period. The spectral time distortion was computed between real values and interpolated values.

Four interpolation methods are shown in Table 5.1. In Table 5.1,

x_i means line spectrum parameters and every parameter is not quantized, to clarify only the interpolation property.

The spectral distortion variety, using various interpolation methods, is shown in Fig. 5.1. The time spectral distortion in case (a) is very large. The spectral distortion in case (c) is nearly equal to that in case (b).

5.3 Variable Frame Rate (VFR) Scheme by Dynamic Programming

In a VFR scheme, feature parameters are transmitted only when a change in speech signal is observed, in comparison with the preceeding transmitted parameters. The untransmitted parameters are interpolated using parameters from the two adjacent transmitted frames at the receiver.

First, optimum piece-wise linear interpolation by dynamic programming (DP) is proposed. It is assumed that T is the speech signal duration. The problem is to determine an optimum division of T with N segments as a function of time spectral distortion.

Each block has eleven LSR parameters (five frequencies and six amplitudes).

LSR parameters are linearly interpolated in each segment and the spectral distortion between interpolated and real values is computed. This procedure is expressed in the following equations.

$$F(N,T) = \text{Min} \sum_{i=1}^{N} S(t_{n-1}, t_n) \qquad (5.1)$$

$$(0=t_0 < t_1 < t_2 \text{ -----} < t_N = T)$$

$$S(t_{n-1},t_n) = \sum_{\ell=t_{n-1}}^{t_n} (\Delta S_\ell)^2 \qquad (5.2)$$

where N is the number of divisions, Δ S is the spectral distortion in a frame and S is the spectral distortion in a segment. $S(t_{n-1},t_n)$ is approximately calculated by the following equations, as mentioned in Chapter 3.

$$\tilde{S}(t_{n-1},t_n) = \sum_{\ell=t_{n-1}}^{t_n} \left[\sum_{i=1}^{5} (C_{f_i} \Delta f_{i\ell})^2 + \sum_{i=1}^{6} (C_{m_i} \Delta m_{i\ell})^2 \right] \qquad (5.3)$$

where C_{f_i} and C_{m_i} mean spectral sensitivity, as shown in Table 3.3. Terms $\Delta f_{i\ell}$ and $\Delta m_{i\ell}$ are parameter differences between untransmitted value and interpolated value.

Dynamic programming technique can be applied to determine the optimum division $(t_1, t_2, \cdots, t_{N-1})$

When T and N are given, this procedure is shown as the following equations.

$$\begin{aligned}
&k = 1 \quad && G_1(t_1) = S(t_0,t_1) && t_0 < t_1 \leqq T-N+1 \\
&k = 2 \quad && G_2(t_1,t_2) = \text{Min}\,[\,S(t_0,t_1) + S(t_1,t_2)\,] && \\
& && && t_0 < t_1 < t_2 \leqq T-N+2 \\
&k = 3 \quad && G_3(t_2,t_3) = \text{Min}\,[\,G_2(t_1,t_2) + S(t_2,t_3)\,] && \\
& && && t_0 < t_1 < t_2 < t_3 \leqq T-N+3 \\
& \vdots && \vdots && \vdots \\
&k = N-1 \quad && G_{N-1}(t_{N-2},t_{N-1}) = \text{Min}\,[G_{N-2}(t_{N-3},t_{N-2}) + &&
\end{aligned} \qquad (5.4)$$

$$+S(t_{N-2}, t_{N-1}) + S(t_{N-1}, t_N)]$$

$$t_0 < t_1 < t_2 < \cdots < t_N = T$$

This algorithm gives the minimum time distortion, when T and N are given. The difference between uniform and non-uniform sampling with DP is examined using real speech samples. It is better to choose the long duration of T in using piecewise optimum approximation. The combination between T and N in the experiment is determined as shown in Table 5.2.

Spectral distortions by both sampling methods are plotted in Fig. 5.2. It is possible to achieve about 35 % spectral distortion reduction by DP in comparison with uniform sampling. The spectral distortion distribution in each frame is shown in Figs. 5.3(a)-(f). It is observed that frames which have a large spectral distortion are reduced by VFR sampling. The duration distribution for each segment by VFR are shown in Figs. 5.4 (a)-(f). Segments which exceeds 100 mseconds duration time mostly appear in silence or stable portion in vowel sound. Relationships between the spectral distortion and utterance speed rate are listed in Tables 5.3 (a) and (b). The spectral distortion increases according to the utterance speed rate. This result means that piecewise optimum division T and N are not sufficient to represent the rapid transition precisely.

In the VFR system, the most important factor is how to select the frame to be transmitted. The VFR scheme using DP has the following problems.

(1) It is better to use long speech duration as much as possible to get optimum division. However, in this case, the time delay for transmission becomes long.

(2) An extended calculation procedure is needed in getting the

optimum division.

5.4 Fully Automatic Straightforward VFR Scheme

To solve several problems in the VFR scheme using DP, an automatic straightforward VFR scheme for selecting the frame to be transmitted is introduced.

Assume that frame n is to be transmitted, and frame (n+m) (m=1,2,-----) is under consideration, for each of the (m-1) frames that lie between frames n and (n+m). Spectral distortion appears between the actual value and the value obtained from linear interpolation between frames n and (n+m). These (m-1) spectral distortion are accumulated and the average spectral distortion over these intervals is calculated. This average spectral distortion is compared against a spectral threshold fixed previously. If the spectral distortion exceeds the threshold value, the frame (n+m-1) is transmitted. If not, the same procedure is repeated until the spectral distortion exceeds the threshold value. Figure 5.5 shows the above procedure.

Using this method, the point can be achieved wherein the average spectral distortion in each section is under the threshold value. However, it should be pointed out that this method requires a buffer to maintain several frames for comparison and the frame interval information must be transmitted. This threshold value is set to be DL of spectral distortion mentioned in Chapter 3. Of course, several modifications can be considered. It is possible to replace average spectral distortion with the maximum spectral distortion.

Using this method, the spectral distortion decreases in comparison with uniform sampling. The spectral distortions caused by two VFR

schemes are compared in Fig. 5.6. The spectral distortion caused by the automatic VFR scheme is reduced to 65 % of that in uniform sampling. In a real system, parameters must be quantized properly. Therefore, the VFR scheme mentioned above takes parameter quantization into consideration.

LSR parameter fluctuations for a female speaker are shown in Figs. 5.7 (a), (b) and (c). Figure 5.7 (a) shows the original parameter fluctuation. Figure 5.7 (b) shows parameter fluctuation with uniform sampling, when frame length is 50 mseconds. Figure 5.7 (c) shows the piecewise linear approximation proposed above, when averaged frame length is 50 mseconds. In this figure, LSR parameters are transmitted less often during steady speech state, and are transmitted more often during rapid speech transition.

5.5 Speech Quality Assessment by Subjective Measurement

In order to verify the synthesized speech quality by VFR method mentioned above, listening experiments were carried out. In the experiment, speech quality between synthesized speech and speech samples with added noise were compared under several conditions.

5.5.1 Experimental Conditions

Three kinds of sentence(A,B,C) listed in Table 5.4 were uttered by two male and two female speakers. The speakers were professional announcers. The speakers and sentences combination is determined as shown in Table 5.4. Three sets, T_1 ,T_2 and T_3 in Table 5.4, were used in the pair comparison test.

5.5.2 Kinds of Synthesized Speech

Speech samples were synthesized by LSR parameters with uniform and variable frame rates. Average frame length was set from 30 mseconds to 50 mseconds. Consequently, information rates ranging from 1200 bps to 1800 bps and 13 speech samples were used for test totally. Bit allocations to LSR and excitation source parameters are listed in Table 5.5. Ranges for LSR parameters quantization are shown in Table 5.6.

5.5.3 Generation of Reference Speech Samples

Reference speech samples were generated for comparison with synthesized speech by the VFR technique. These samples were generated by adding multiplicative random noise to speech signal [60].

Noise signal n(t) is defined by

$$n(t) = k \; s(t) \; n_0(t) \tag{5.5}$$

where k is noise signal gain factor, s(t) is speech signal and $n_0(t)$ is random number between -1 and +1 with uniform distribution.

The signal to noise (S/N) ratio is represented as

$$S/N = 20 \log (1/k) - 10 \log \sigma^2 \tag{5.6}$$

where σ^2 is variance in $n_0(t)$ and is 1/3 in the case of uniform distribution.

Then, Eq. (5.6) is rewritten as

$$S/N = -20 \log k + 10 \log 3 \qquad (5.7)$$

The signal to noise ratio (S/N) is set freely, setting an available number as k in the above equation. S/N was set at thirteen different values, 0, 3, 7, 10, 13, 15, 17, 20, 23, 27, 30, 40, ∞ (without noise) for a speaker.

5.5.4 Listening Rating Test

In the listening rating measurement, many untrained subjects were asked to listen to various speech. Forty subjects took each test. A total of 120 subjects were used.

Subjects judge the speech quality on the following 5 point category scale.

0----Unsatisfactory

1----Poor

2----Fair

3----Good

4----Excellent

From these scores, equivalent S/N to speech samples can be obtained by VFR method. The relation between S/N and mean opinion scores (MOS) for the noise additional speech samples is shown in Fig. 5.8. The equivalent S/N ratio corresponding to synthesized speech is shown in Fig. 5.9.

Opinion scores for all samples are listed in Appendix 2. The

quality of synthesized speech at 1200 bps is equivalent to a speech sample with 9 dB, and that at 1500 bps is equivalent to 14 dB noise added to speech. In the case of synthesized speech at 1500 bps, mean opinion scores have large variance. This difference depends on the kinds of speech impairment between VFR and noise added to speech. In the case of noise added to speech, the acceptable S/N ratio is nearly the same among the subjects. On the other hand, opinion scores are divided for impairment in synthesized speech, for example, the lack of naturalness and pitch rapid transition. The relationship between spectral distortion and opinion score is shown in Fig. 5.10. Mean opinion scores decrease with the increase in spectral distortion. However, when the spectral distortion is the same, MOS has a large variance. This is mainly caused by miss-matching of spectral envelope parameters and excitation source parameters.

A new VFR scheme considering excitation source parameters must be developed.

5.6 Conclusions

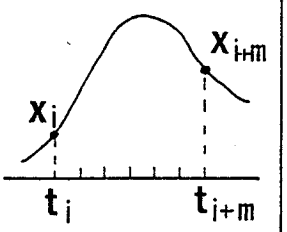
Speech data compression by VFR was presented, using LSR parameters, which have good interpolation properties.

(1) The variable frame rate extraction method by dynamic programming was introduced. The information bit rate was reduced to 65 % of that in the case of uniform sampling.

(2) A straightforward VFR scheme was proposed. This method can compress the mean spectral distortion in each segment under the threshold value. Using this method, 70 % spectral distortion reduction can be achieved, compared with uniform sampling.

(3) Synthesized speech samples were tested by subjective measurement,too. The synthesized speech quality at 1500 bps by VFR method was equivalent to a speech sample which had 14 dB added multiplicative noise.

Table 5.1 Several interpolation methods.

a	$x_{i+j} = x_i$	
b	$x_{i+j} = (x_i + x_{i+m})/2$	
c	$x_{i+j} = \frac{x_{i+\cdots+x_{i+m-1}}}{m}$	
d	$x_{i+j} = x_i + \frac{x_{i+m}-x_i}{m} j$	$j = 0, \bullet\bullet\bullet, m-1$

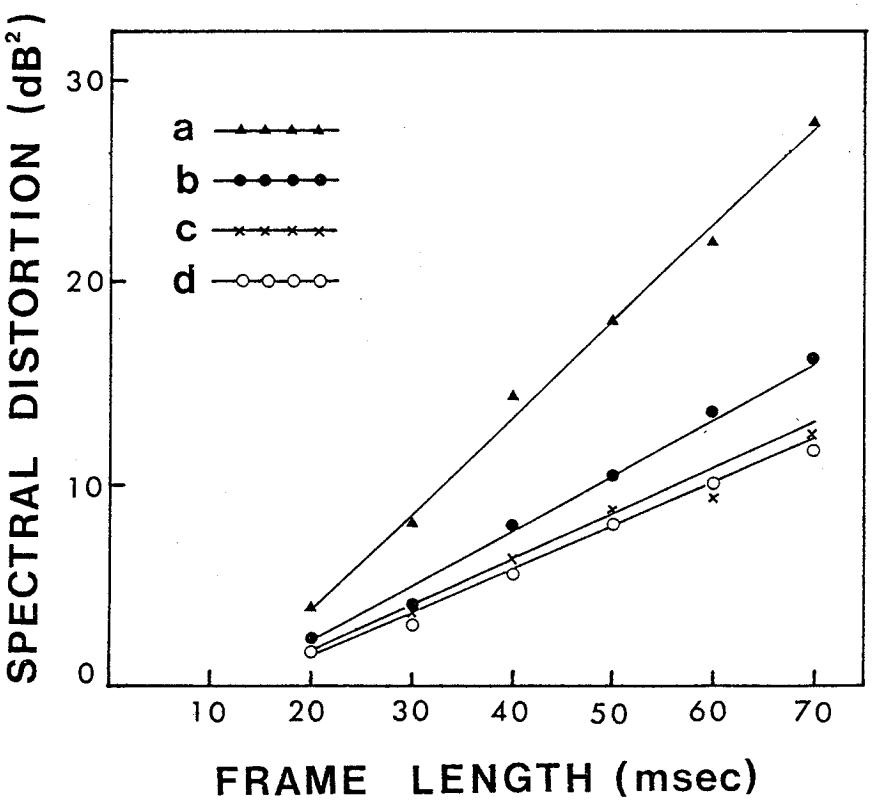


Fig. 5.1 Spectral distortion by interpolation versus frame length.

Table 5.2 Combinations of block length (T) and number of divisions (N).

MEAN FRAME LENGTH (msec)	T (msec)	N
20	200	10
30	300	10
40	400	10
50	500	10
60	480	8
70	490	7

Table 5.4 Combinations of sentences and speakers for a pair comparison listening test.

SPEAKER / SENTENCE	FEMALE		MALE	
	TS	EN	SM	SG
A	T_1	T_3	T_2	
B		T_2	T_1	T_3
C	T_2	T_1	T_3	

(A) ' bakuoNga giNsekaino koogeNni hirogaru '

(B) ' choseN naNbuni teikiatsuga atte, kyusokuni hattatsu shinagara '

(C) ' mukashi mukashi arutokoroni ojiisaNto obaasaNga imashita '

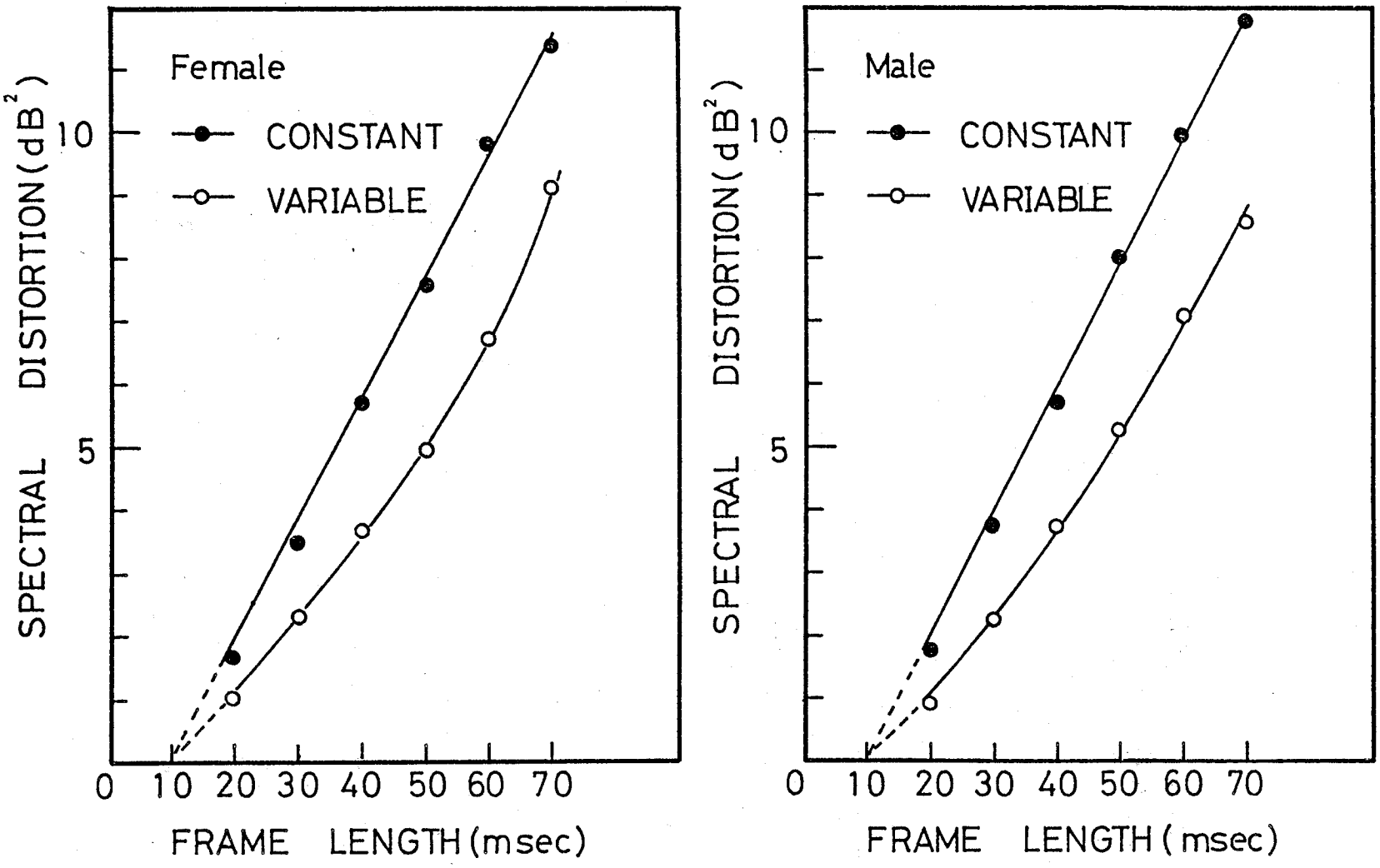


Fig. 5.2 Time spectral distortion difference between uniform and VFR sampling versus mean frame length.

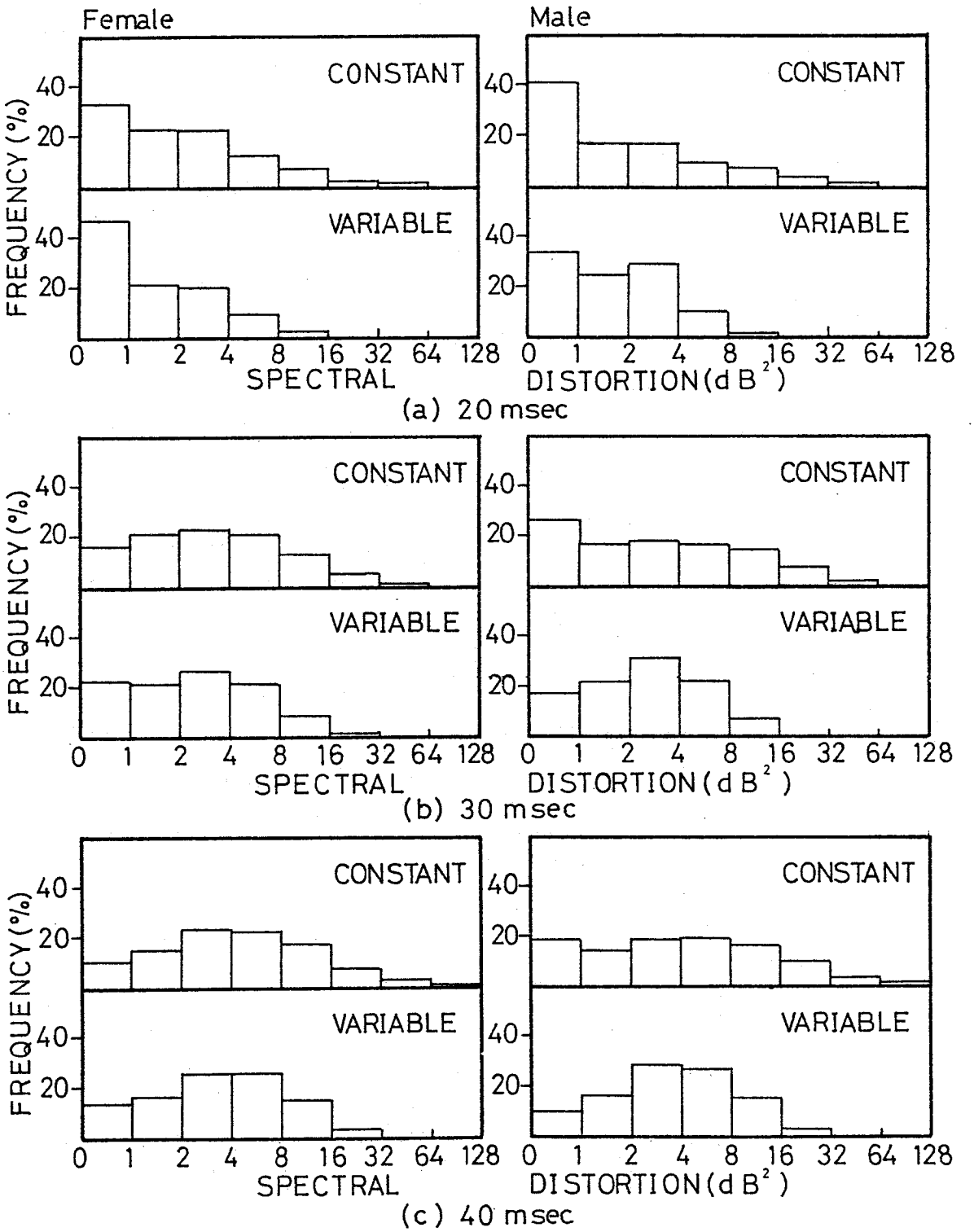


Fig. 5.3 Spectral distortion occurences at each frame in VFR scheme by dynamic programming.

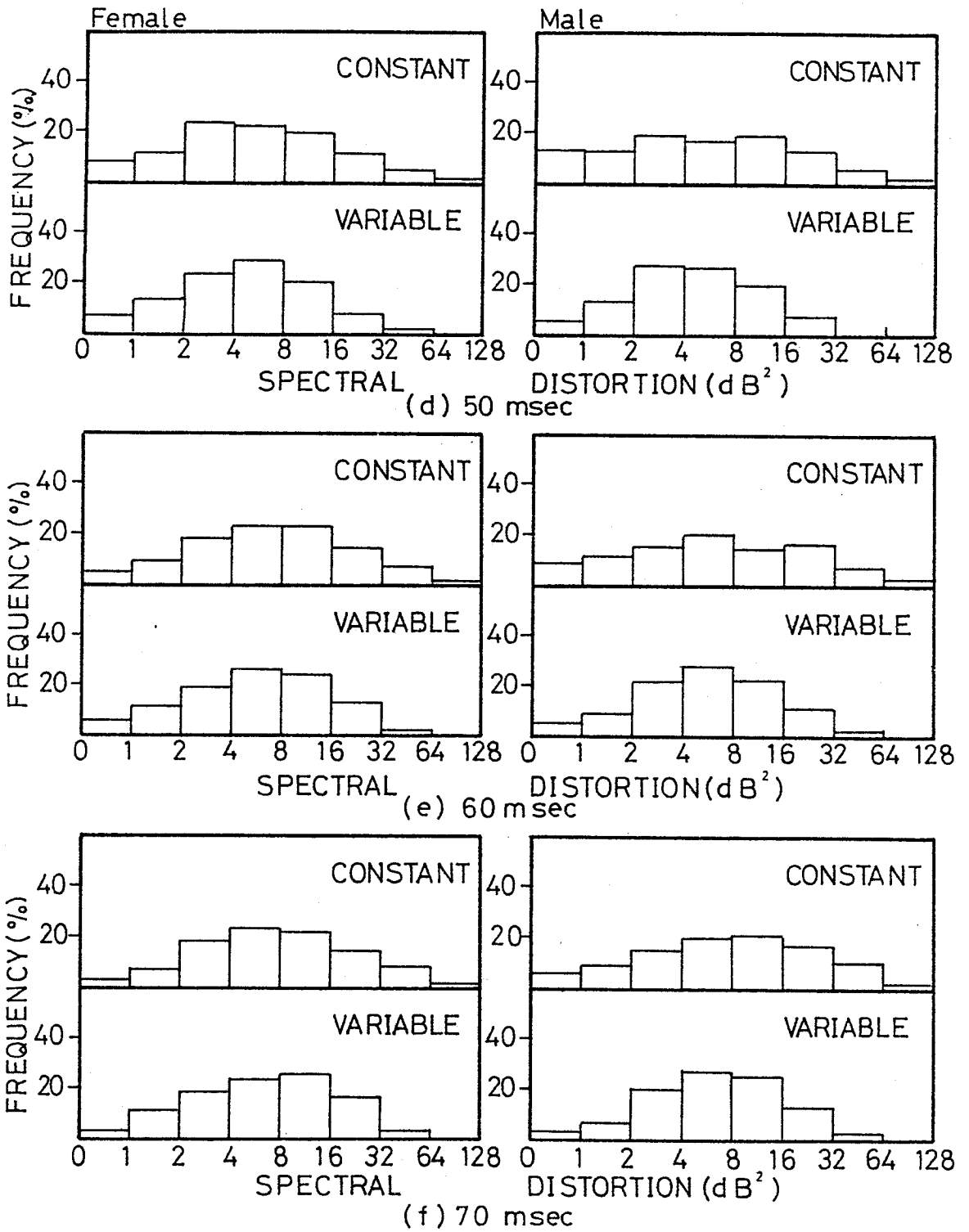


Fig. 5.3 (continuation)

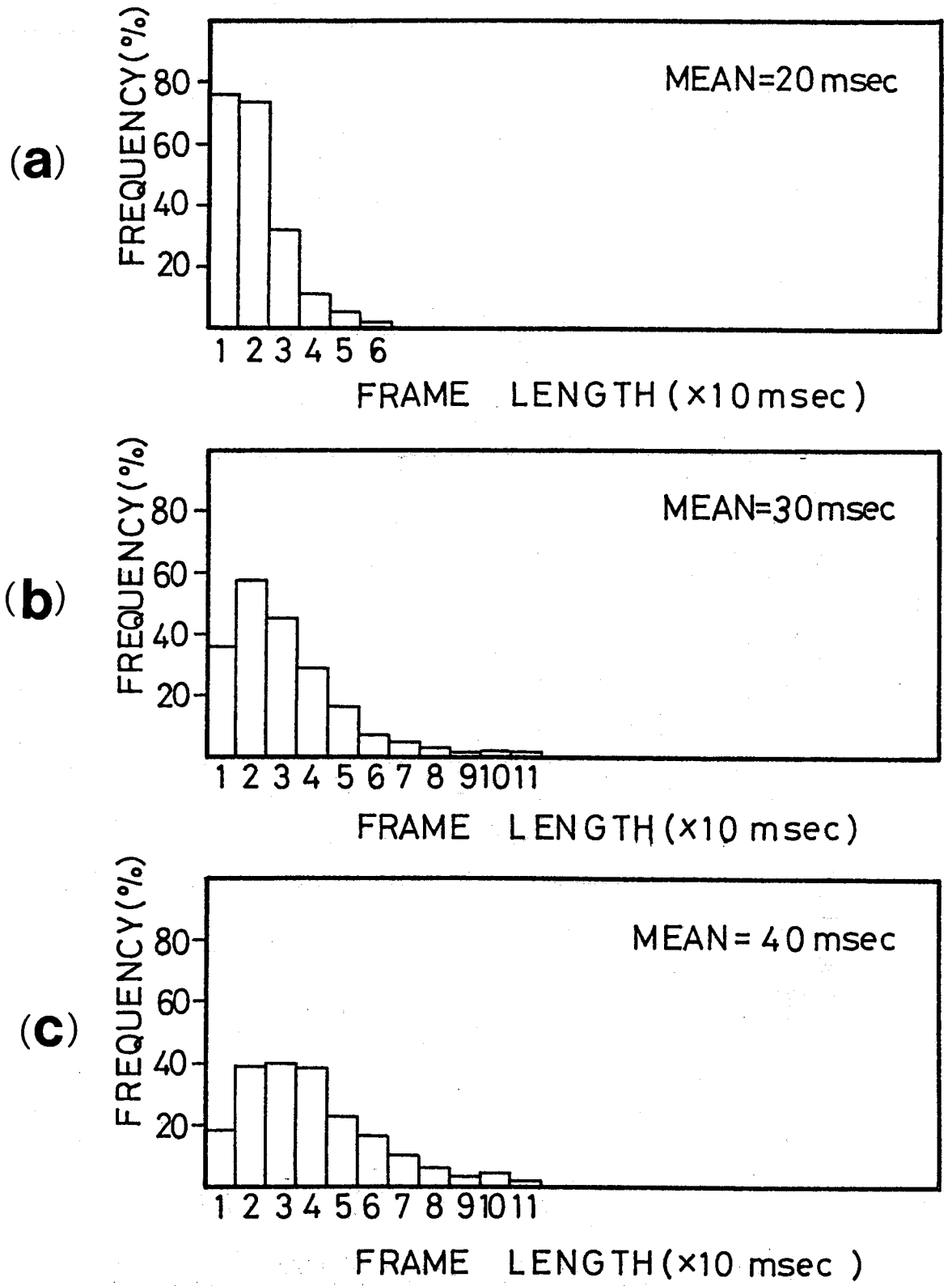
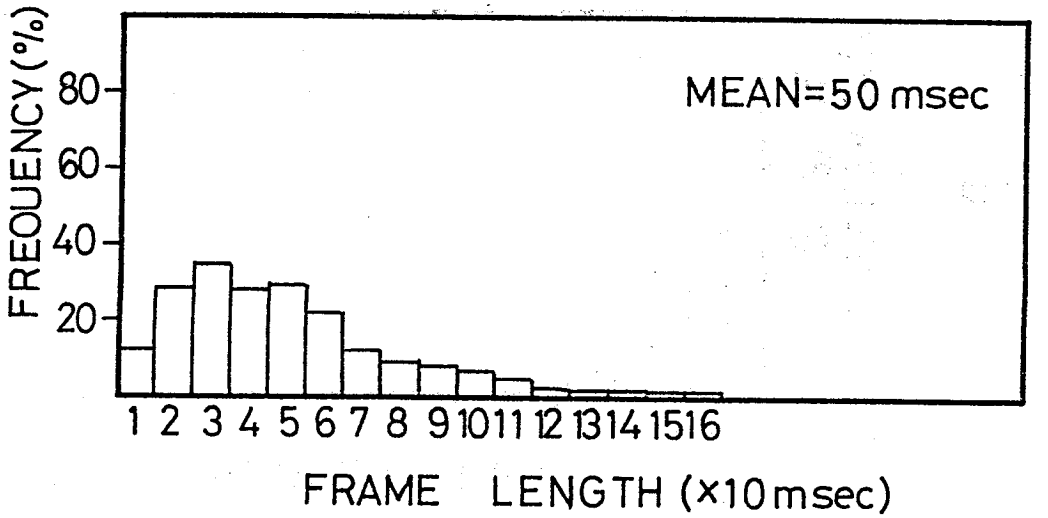
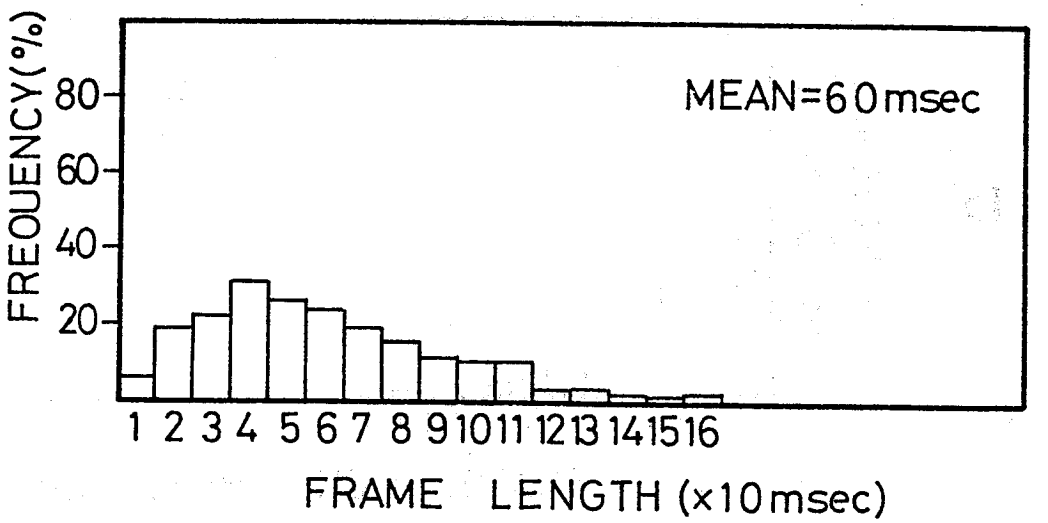


Fig. 5.4 Frame length distribution in VFR scheme by dynamic programming.

(d)



(e)



(f)

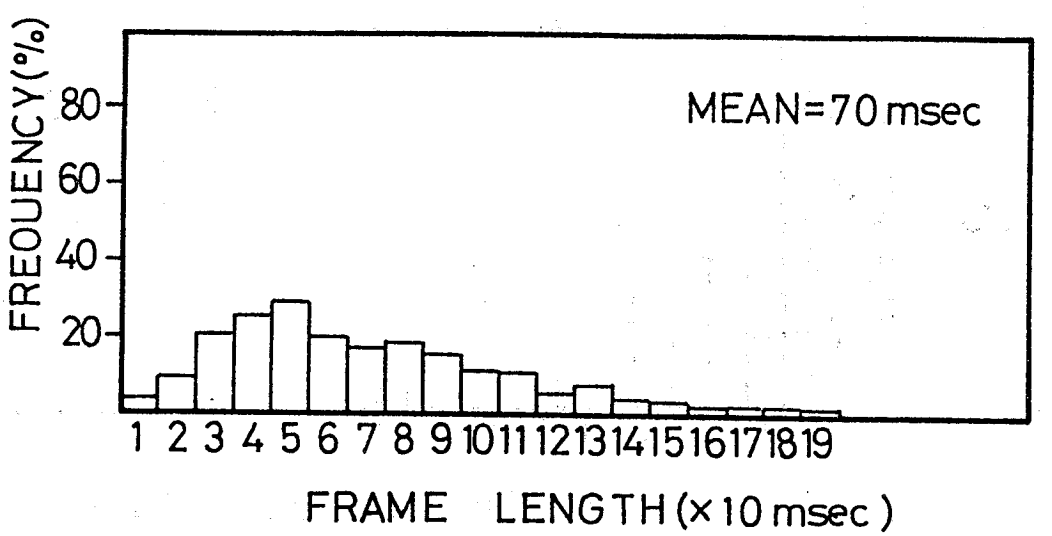


Fig. 5.4 (continuation)

Table 5.3 Spectral time distortion in relation to utterance speed rate (male speaker).

(dB^2)

SPEAKER		K I		N A		M A		I T		T O	
UTTERANCE SPEED(sec)		3.2		2.7		4.1		2.8		4.1	
FRAME PERIOD		CON.	VAR.	CON.	VAR.	CON.	VAR.	CON.	VAR.	CON.	VAR.
MEAN FRAME LENGTH (msec)	20	1.75	0.80	1.80	0.91	1.61	0.83	1.95	1.09	1.57	1.09
	30	3.80	2.21	3.65	2.58	3.48	1.69	3.82	2.64	3.62	2.41
	40	5.72	3.96	5.60	4.12	5.24	3.00	6.35	4.36	5.50	3.70
	50	7.87	5.33	7.56	5.78	6.99	4.58	10.2	6.03	7.49	4.97
	60	9.76	6.95	10.0	7.36	9.11	6.31	12.2	8.93	9.13	6.60
	70	12.1	8.54	10.0	8.89	11.0	8.17	14.1	10.4	11.3	7.59

Table 5.3 Spectral time distortion in relation to utterance speed rate (female speaker).

(dB^2)

SPEAKER		E N		U E		S G		K W		N A	
UTTERANCE SPEED(sec)		3.4		3.6		4.2		3.1		2.9	
FRAME PERIOD		CON.	VAR.	CON.	VAR.	CON.	VAR.	CON.	VAR.	CON.	VAR.
MEAN FRAME LENGTH (msec)	20	1.66	1.20	1.47	0.91	1.42	0.90	1.94	1.11	2.04	1.19
	30	3.61	2.78	2.96	2.10	3.00	1.78	3.89	2.47	4.18	2.50
	40	6.19	4.32	4.80	3.15	5.08	2.97	5.76	3.94	7.01	4.18
	50	8.37	5.90	6.31	4.49	6.53	4.10	7.92	5.11	8.88	5.50
	60	10.9	8.12	8.68	5.95	8.37	5.89	10.2	6.39	11.4	7.36
	70	13.0	9.80	10.7	6.60	9.45	7.15	11.6	8.24	12.2	8.73

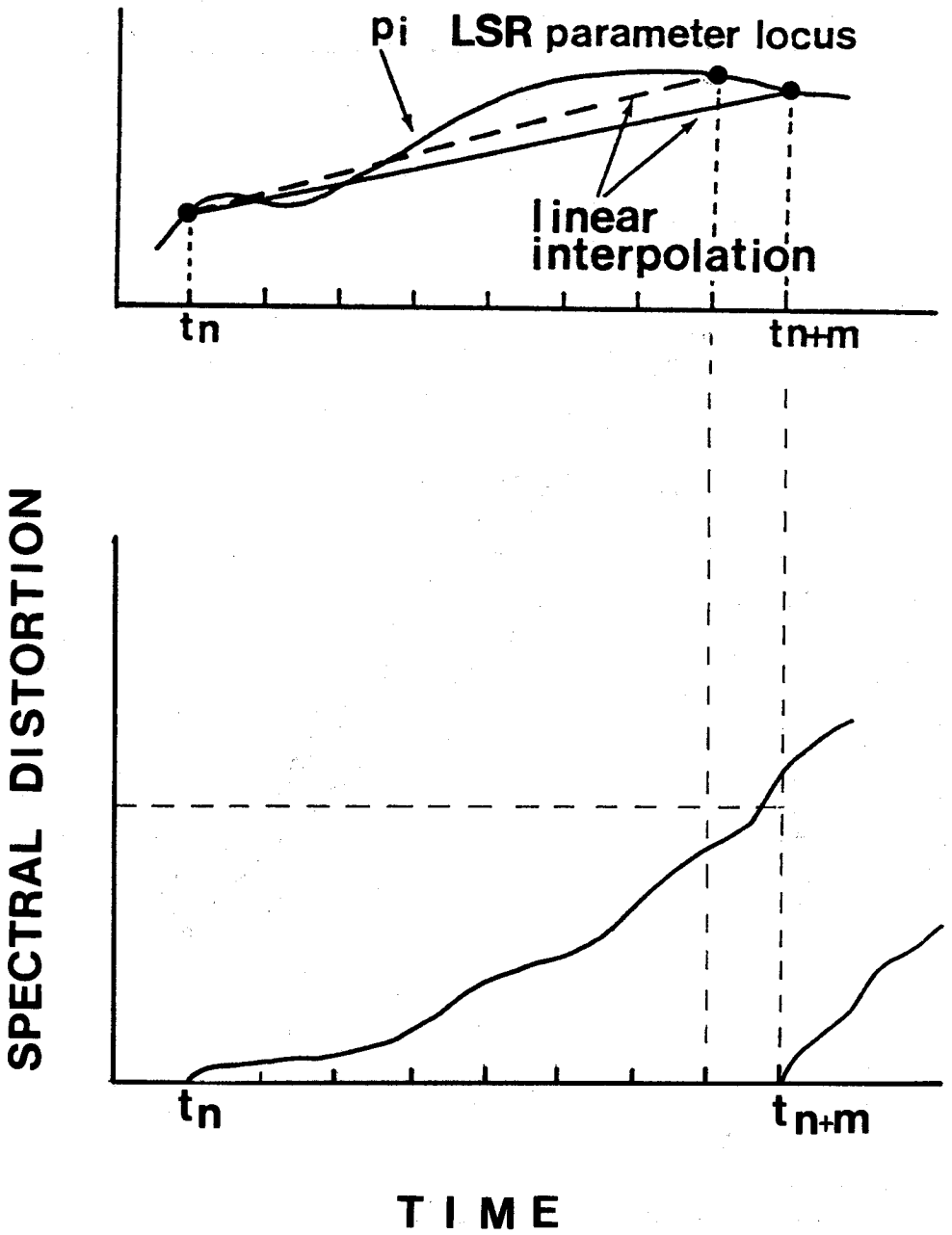


Fig. 5.5 Straightforward VFR scheme illustration.

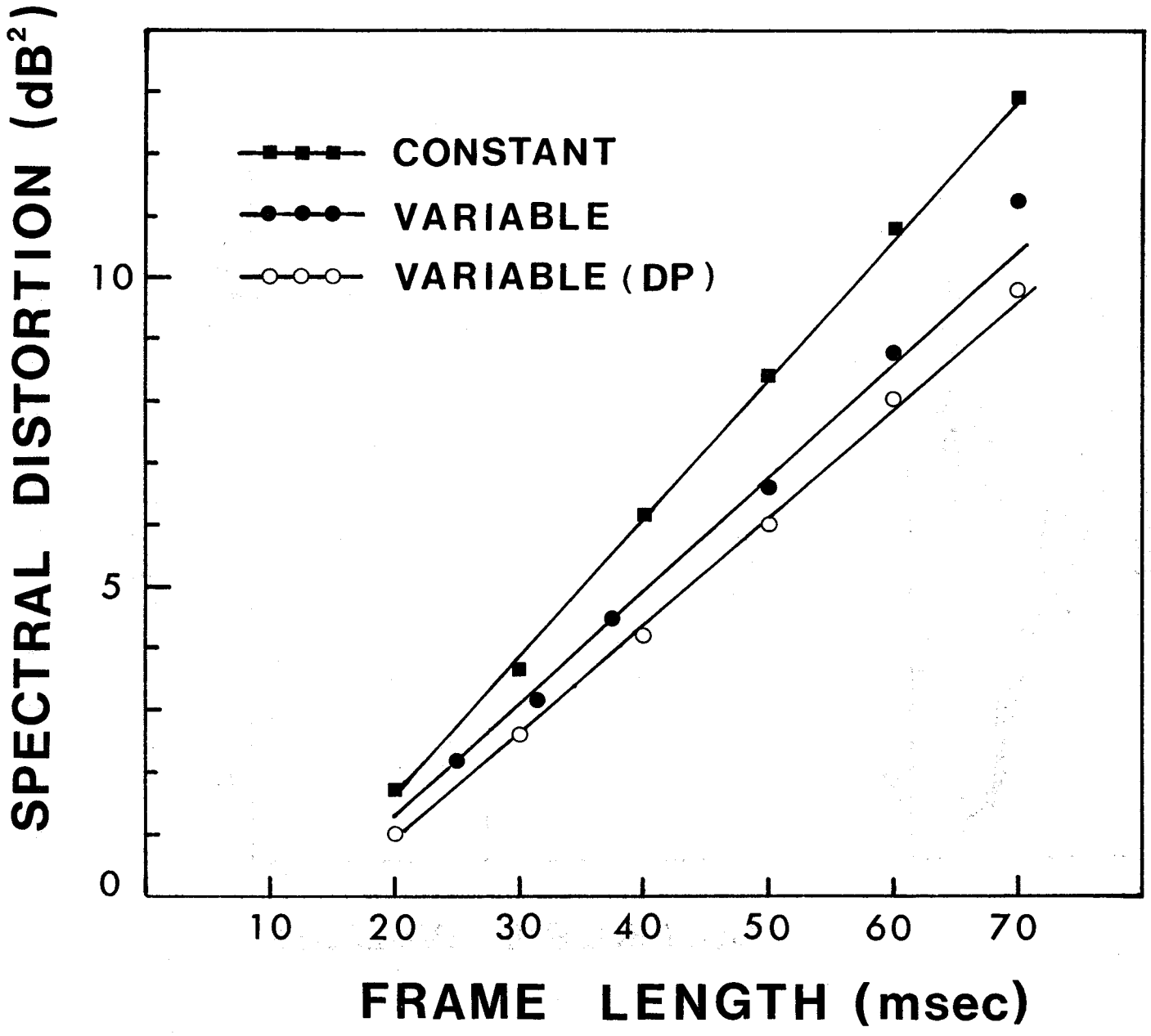


Fig. 5.6 Spectral distortion comparison between two VFR methods.

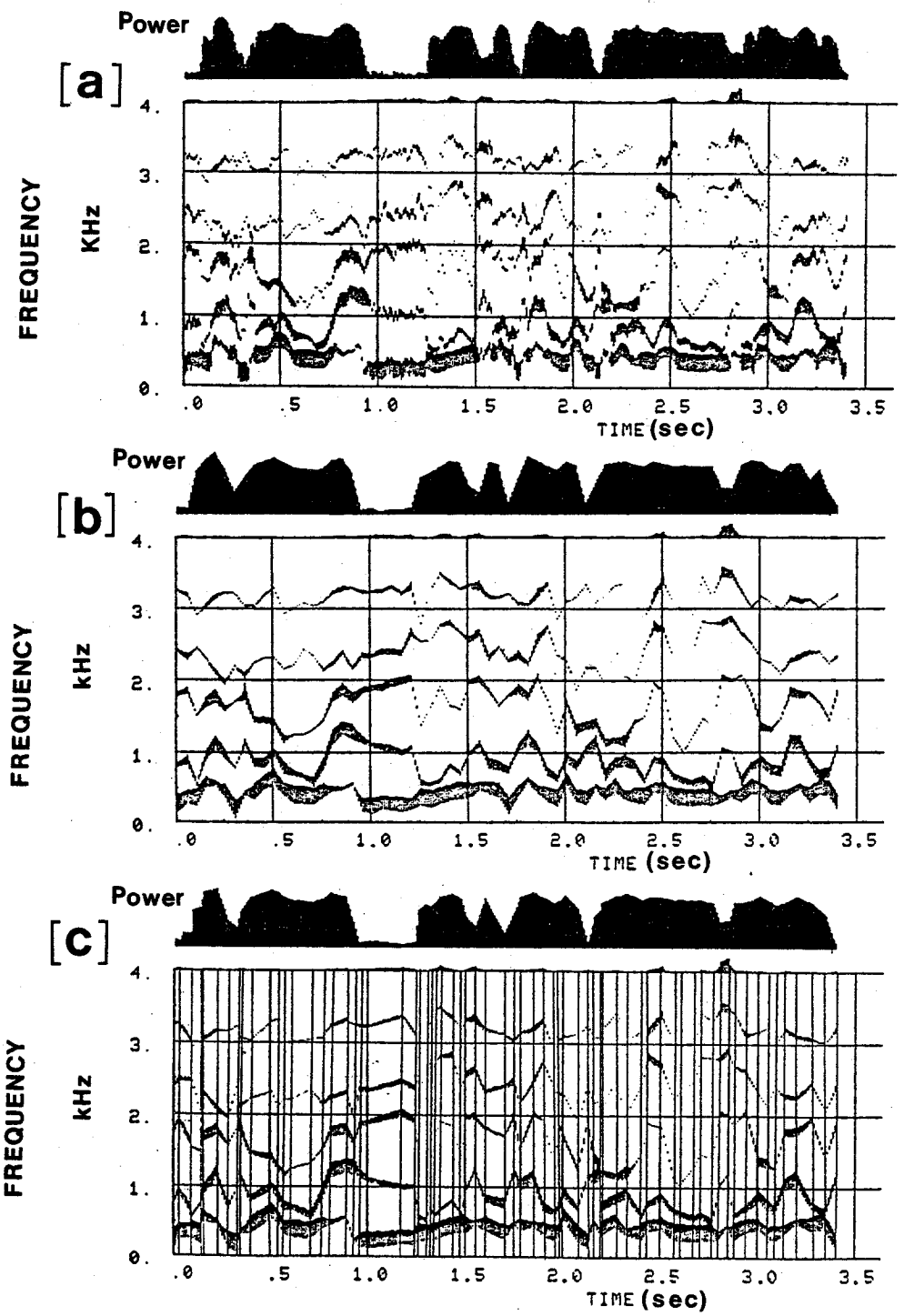


Fig. 5.7 LSR fluctuations by uniform sampling and VFR scheme.
(a) original pattern (b) uniform sampling
(c) VFR sampling

Table 5.5 Bit allocations in synthesized speech samples.

FRAME PERIOD	BIT RATE (bps)	MEAN FRAME LENGTH msec	BIT ALLOCATION			EXCITATION		LSR PARAMETERS										
								FREQUENCY					AMPLITUDE					
			EX.	SP.	TIME	POW.	PITCH	f_1	f_2	f_3	f_4	f_5	m_1	m_2	m_3	m_4	m_5	m_6
CONSTANT	1800	30	12	41	1	5	7	4	4	4	4	3	3	4	4	4	4	3
	1500	30	10	34	1	4	6	3	4	3	3	3	3	3	4	3	3	2
	1500	40	12	47	1	5	7	4	5	5	4	4	4	4	5	5	4	3
	1500	50	14	60	1	6	8	5	6	6	5	5	5	6	6	6	5	5
	1200	30	9	26	1	4	5	2	3	3	2	2	2	2	3	3	2	2
	1200	40	11	36	1	5	6	3	4	4	3	3	3	3	4	4	3	2
	1200	50	12	47	4	6	8	4	5	5	4	4	4	4	5	5	4	3
VARIABLE	1500	30	10	31	4	4	6	3	3	3	3	3	3	3	3	3	2	2
	1500	40	11	45	4	5	6	4	5	4	4	4	4	4	5	4	4	3
	1500	50	14	57	4	6	8	5	6	5	5	5	5	6	5	5	5	5
	1200	30	9	23	4	4	5	2	3	3	2	2	2	2	2	2	2	2
	1200	40	10	34	4	4	6	3	4	3	3	3	3	3	4	3	3	2
	1200	50	11	45	4	5	6	4	5	5	4	4	4	4	4	4	4	3

Table 5.6 Quantization ranges for LSR parameters.

	DISTRIBUTION MIN1% —— MAX1%		QUANTIZATION RANGE
f_1	0.08	0.52	0.44 kHz
f_2	0.40	1.20	0.80
f_3	0.96	2.12	1.16
f_4	1.88	2.88	1.00
f_5	2.68	3.56	0.88
f_6	4 (CONSTANT)		
m_1	0.047	0.990	13.2 dB
m_2	0.0079	0.758	19.8
m_3	0.00014	0.575	36.1
m_4	0.000036	0.288	39.0
m_5	0.000021	0.331	42.0
m_6	0.000004	0.200	47.0

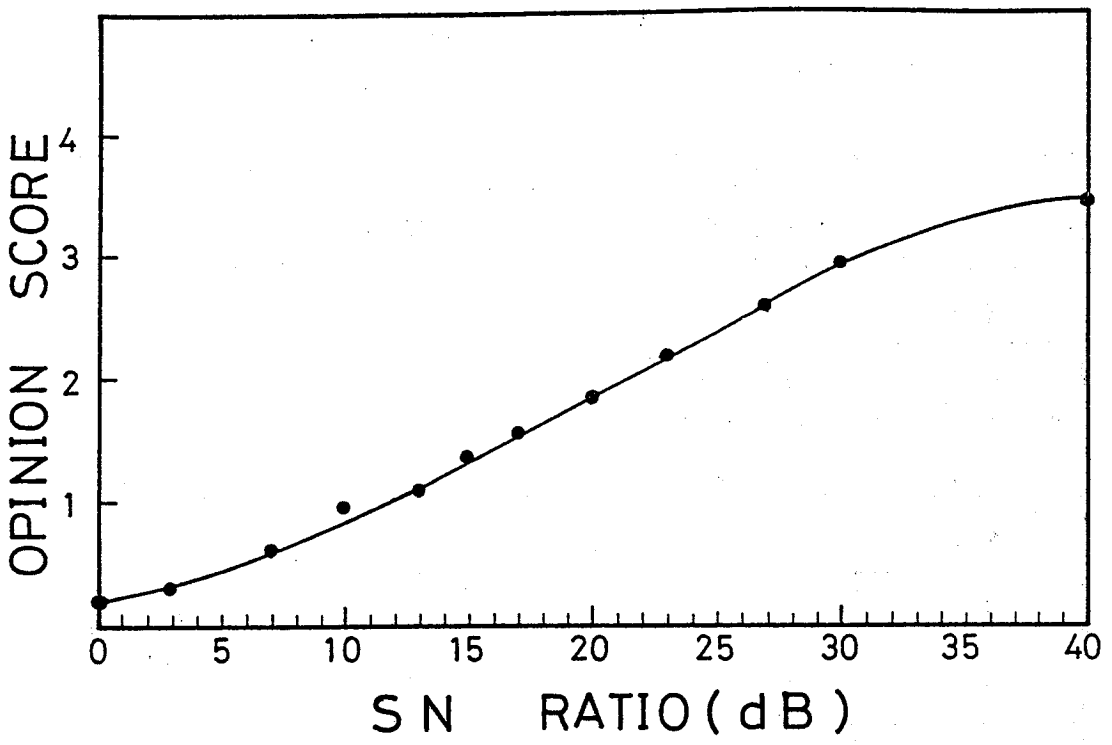


Fig. 5.8 Relationship between S/N ratio and mean opinion score(MOS).

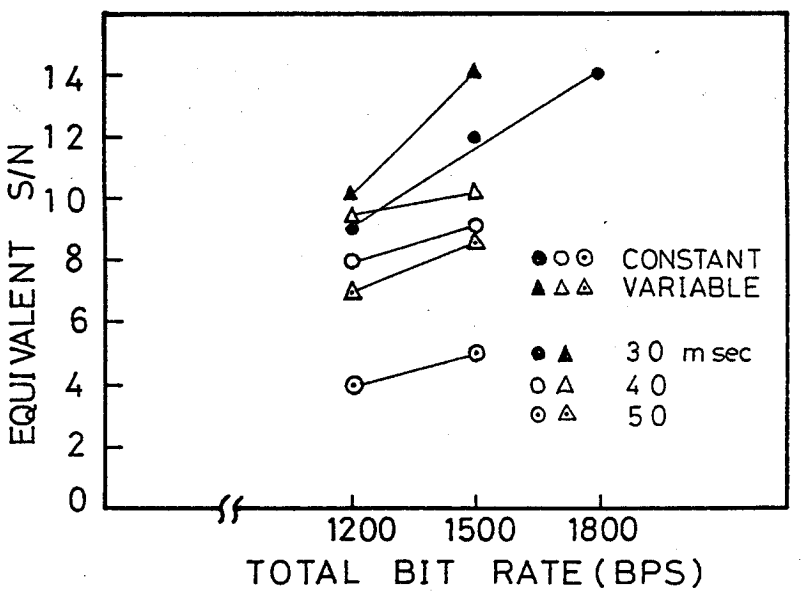


Fig. 5.9 Equivalent S/N ratio corresponding to synthesized speech.

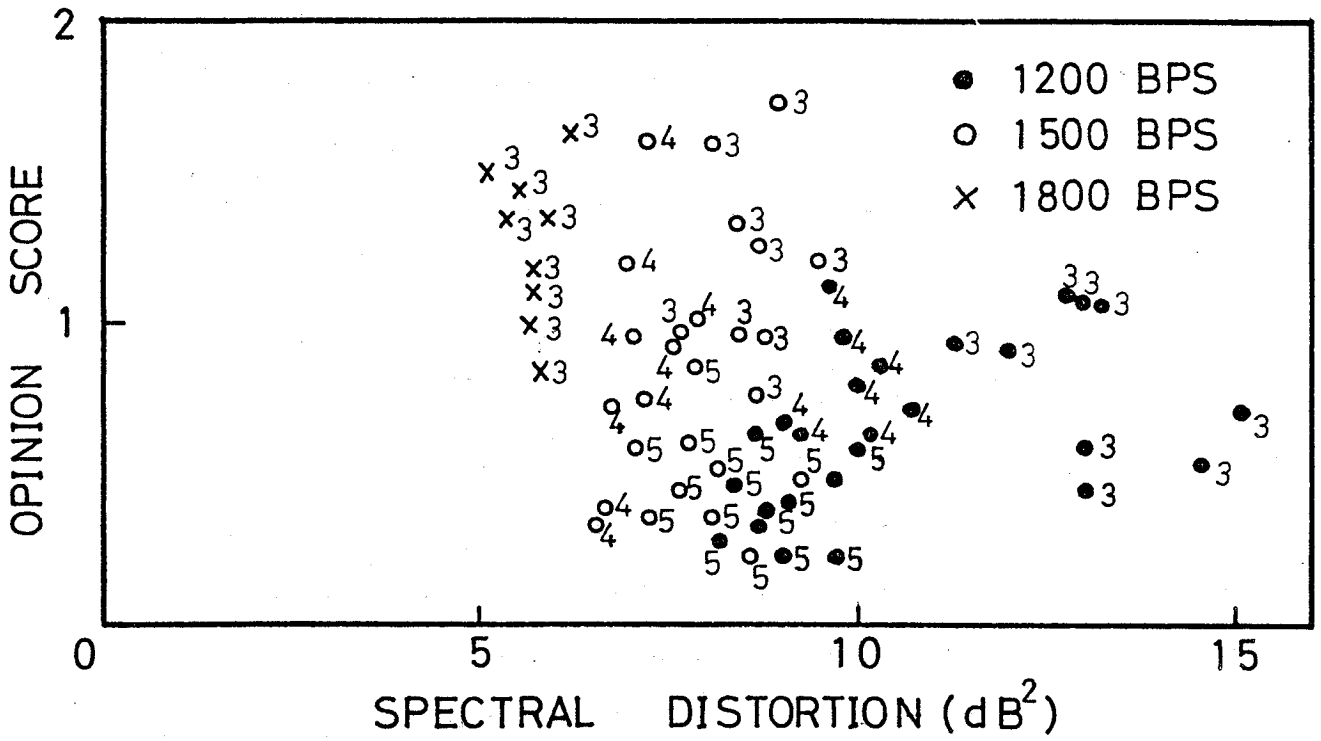


Fig. 5.10 Relationship between spectral distortion and mean opinion score for synthesized speech by VFR method.

CHAPTER 6 VECTOR QUANTIZATION USING LSP PARAMETERS

6.1 INTRODUCTION

In this chapter, a pattern matching vocoder using the LSP parameter is presented. In this system, standard spectral patterns are generated automatically by a simple clustering algorithm. Differences between scalar and vector quantization are discussed as a function of spectral distortion.

6.2 Pattern Matching Vocoder System

Speech analysis synthesis by PARCOR or LSP is a good method to compress the speech signal in narrow band transmission. Several quantization techniques have been proposed. However, these techniques are based on scalar quantization of feature parameters. Namely, each parameter is independently quantized, regardless of the relation with parameters in a frame. Thus, the limitation for data compression is around 2.4 kbps. Moreover, in order to compress the speech signal, it is necessary to eliminate the feature parameters redundancy in the time domain and space domain. One of these is VFR transmission, proposed in the previous chapter. The other is to use of the parameter distribution redundancy. Feature parameters are not generally distributed uniformly, and it is useful to represent spectrum patterns as a parameters set. Only a code, which corresponds to a parameters set, is transmitted instead of the parameters themselves. This method is called a "Pattern Matching Vocoder". The basic idea was proposed by C.P.Smith in 1957 [13],[61]. Figure 6.1 shows a typical pattern matching vocoder system.

In this system, typical spectral patterns, which correspond to several phonemes, are stored previously as discrete symbols using training speech samples. Each discrete symbol corresponds to one spectrum pattern represented by several feature parameters. Input speech is analyzed every 10 mseconds and LPC parameters are extracted. The spectral distance between a current frame and each stored spectrum pattern is calculated. Only the code which corresponds to the nearest spectrum pattern is transmitted. At the receiver, it is necessary to decode spectral parameters from a received code for speech synthesis. Excitation signals, for power and pitch and voicing, are transmitted as scalar quantization or another quantized technique (differential adaptive coding et al.) in most of these systems. It is possible to achieve drastic bit savings, in comparison with transmitting the parameters themselves. Recently, this method is called "Vector quantization", in comparison with scalar quantization. Several quantization techniques have been reported and compared with conventional scalar quantization [62],[63],[64].

The pattern matching vocoder has significant features over other vocoder systems. First, it is possible to carry speech information at under 1000 bps. Second, it furnishes high reliability to keep speech contents secrecy. The reason is, if there is no coincidence between coder and decoder, it is rather difficult to get intelligible synthesized speech quality. On the contrary, in the pattern matching vocoder, there are two big technical problems. First, synthesized speech quality is generally not as good as that for other methods. The reason is that, when the speech signal is represented by limited spectral patterns, spectral unsmoothed transitions occur. The rapid parameter change sometimes damages synthesized speech quality. The

second problem is how to select typical spectral patterns. It is very difficult to decide the number of discrete patterns required to represent the virtually unlimited variety of speech spectra. Therefore, adjacent spectral patterns are sometimes very different and it is necessary to interpolate parameters smoothly, in order to generate high quality synthesized speech. The third problem is spectral distance measure. It is desired that the distortion measure be meaningful subjectively. To solve these problems, the new pattern matching vocoder is presented. This system have several features over the preceeding methods.

For the first problem, LSP parameters are used in order to interpolate feature parameters smoothly. LSP parameter have better interpolation properties than other LPC parameters [65]. This feature can be used efficiently in a pattern matching vocoder system.

For the second problem, typical spectrum patterns are generated automatically without linguistic information. Much data can be dealt with by this method.

For the third problem, the logarithmic difference in power spectra between input speech and reference spectral pattern is used. This measure was used for evaluating synthesized speech quality in the PARCOR or LSP synthesis method. This measure is convenient for comparing the speech quality of conventional synthesized speech objectively. Moreover, this measure was proved to have good correspondence to subjective measurement [48],[66].

6.3 Selection for Discrete Patterns of Speech Spectra

There are mainly two methods in selecting discrete patterns for

speech spectra. One is to select discrete patterns using the phoneme information. The other method is to use clustering technique for dividing a space where speech spectra are distributed. In this thesis, the latter method is preferred to select discrete patterns for speech spectra. Speech samples can be dealt with regardless of phonemes and the kind of languages. In this automatic process, there are no ambiguous processes guided by human decision. The process for selecting discrete spectral patterns is presented as follows.

(1) Speech analysis

Training speech samples are analyzed every 10 mseconds. The 10-th order LSP parameters and excitation signals are extracted.

(2) Pre-processing

Data compression for clustering is executed by the following three methods.

The silence duration are eliminated using power threshold. If spectral distance between adjacent frames is less than the fixed threshold, reference frame data is only used for succeeding processes.

(3) Spectral patterns are divided into two classes according to voicing (voiced or unvoiced).

(4) Generation for discrete patterns of speech spectra

The speech spectra are generated automatically by a clustering technique using training utterances.

[Vector Quantization Algorithm]

$E_0 = (g_1, g_2, ----, g_M)$ is a training spectral vector set, which consists of a few thousand frames arbitrarily selected from training

utterances. From E_0, the K phoneme-like templates set $F=(f_1, f_2, \text{----}, f_K)$ is generated. $C(E_i)$ is defined as the mean vector of E_i and symbol $d(g_i, g_j)$ is defined as the spectral distance between g_i and g_j.

(1) Initialization : Set Θ (the distance threshold for vectors which are regarded as the same cluster) and K (the number of phoneme-like templates desired). Set k=1.

(2) Compute the spectral distance $d(g_i, g_j)$ for every combination of i and j.

(3) Count the number of g_i values which satisfies $d(g_i, g_j) < \Theta$ for all i. This vector set is represented as $B(g_i, \Theta)$ and the number is represented as N(i).

(4) Find the vector g_i which has the maximum N(i) number.

(5) Generate two kinds of speech spectral patterns. One is the central pattern itself. The other is the mean vector C determined by calculating the mean value of every autocorrelation coefficient from all the vectors belonging to $B(g_i, \Theta)$. These data are stored as f_i and $\overline{f_i}$, respectively, one of the phoneme-like templates.

(6) $E_{k+1} = E_k - B(g_i, \Theta)$

(7) If $E_{k+1} = 0$ (empty set) or k=K, stop; otherwise replace k by k+1 and go to step (3).

This algorithm is very simple, because it is only necessary to give K and Θ. In this method, the spectral distance between two vectors is calculated only one time. After step (3), only the number which satisfies $d(g_i, g_j) < \Theta$ is calculated, considering the vectors eliminated at step (6). These processes are illustrated in Fig. 6.2.

This clustering method proposed here is basically minimax

criterion. A spectral pattern in high density space can be extracted earlier than that in low density space. The spectral patterns in high density play a more important role than that in low density. This method can suppress the spectral distortion in the worst case at given threshold and it is not necessary to divide each cluster more precisely. Several other clustering techniques have been reported [62]. Most of these algorithms are based on decreasing mean spectral distortion. In these methods, the mean spectral pattern is influenced by isolated patterns, in order to decrease mean spectral distortion. The algorithm proposed here can avoid these isolated patterns influences in generating mean spectra.

Another reason is that, it is possible to apply the Huffman coding technique to reference spectra according to the occurences of such spectral patterns.

Let's assume L discrete spectral patterns are obtained finally. If ℓ spectral patterns are selected, these patterns can cover the training spectral patterns by

$$R(\Theta, \ell) = \sum_{i=1}^{\ell} N(i) \; / \; M \qquad (6.1)$$

This factor is defined as "Covering Rate" and is used for selecting the number of stored spectral patterns.

6.4 Speech Coding at Low Bit Rate Using Vector Quantization [67]

6.4.1 Speaker-Independent Experiment

Speech data base in the experiment, shown in Table 6.1, include

three kinds of short sentences, which are phoneme-balanced as much as possible. These sentences were uttered by a total of 25 speakers, twelve male and thirteen female speakers.

Three male and female speaker utterances for sentences (a) and (b) were used to generate reference spectral patterns. This data set was called Group A here. Total frame number was approximately 20000 frames, 15000 frames for voiced sound, the other was for unvoiced sound. The relation between the number of spectral patterns (1) and covering rate (R) defined in Eq. (6.1) is shown in Fig. 6.3. as a function of spectral distortion threshold. In Fig. 6.3, the solid line is for voiced sound and the dotted line is for unvoiced sound. Six pattern sets were selected according to covering rate, as shown in Table 6.2. Speech samples which were not used in generating standard spectral patterns were used to verify a new pattern matching vocoder. The speech data was analyzed and the distance between individual frames and vectors is calculated. The spectral distortion versus the number of spectral patterns (represented as bits) is shown in Fig. 6.4. For Group A, the spectral distortion is about eight dB in the case of 860 patterns. This distortion corresponds to that for two bits scalar quantization. On the contrary, the spectral distortion for Group B is greater than that for Group A. Furthermore, it doesn't decrease rapidly when the number of spectral patterns increases.

Several sonograms are shown in Fig. 6.5. The upper one shows original spectral patterns. The lower three patterns are sonograms for synthesized speech at 570, 630 and 700 bps, respectively. These speech spectra are synthesized under the following conditions. The frame period was 30 mseconds, six and five bits were allocated for pitch and power, respectively, in each frame. The speaker belonged to Group B.

Even when using 280 patterns, some feature of the original sound are reflected in synthesized speech.

For synthesized speech quality, there are several defect portions at 570 bps. Synthesized speech becomes easy to hear and gets more intelligibility at 700 bps.

In Fig. 6.4, the central patterns are used. Instead of central patterns, using mean spectral pattern, the spectral distortion decreases ten percent, as shown in Fig. 6.6. However, there were still differences between Groops A and B. This is because the characteristics for Group B are not reflected in reference spectral patterns. The re-generation method for selecting spectral patterns are introduced.

First, the same algorithm as that for data belonging in Group B is applied. Spectral patterns, which have a small number of N(i), are replaced by new generated patterns in Group B, which have a large N(i) number. Using the new patterns set, there are small differences between Group A and Group B, as shown in Fig. 6.7.

6.4.2 Speaker-Dependent Pattern Matching Vocoder [68]

In the previous section, the pattern matching system for the speaker-independent pattern matching method is discussed. In this section, the same algorithm is applied to speaker-dependent pattern matching. It is well known that there are many spectral variations for the same phoneme among different speakers. It is a very important and difficult problem to cover these variations with the limited number of spectra. In order to clarify the difference between speaker-dependent and independent pattern matching, vector quantization is applied for a single female speaker.

The parameter distribution ranges are narrow in the speaker-dependent case. Ranges for LSP parameters are shown in Fig. 6.8. The dotted line is for speaker-dependent case and the solid line is for speaker-independent case. For speaker independent case, fourteen male and eleven female speakers are used for the same short sentences. Ranges for this speaker-dependent case are reduced from 50 % to 15 % of those in speaker independent case. The same algorithm was applied for generating reference spectral patterns.

Using 25000 frames, six pattern sets, the same number as listed in Table 6.2, were selected. Another pattern set, which satisfies R=100 %, was also selected. In this case, there was a total of 1263 patterns (1023 voiced, 240 unvoiced).

The relation between the number of spectral patterns and spectral distortion for untrained speech data for the same speaker is shown in Fig. 6.9. Moreover, this result is compared with the speaker-independent pattern matching vocoder and optimum scalar quantization in LSP speech analysis synthesis method, as shown in Fig. 6.10. This result leads to the following conclusions.

(1) By vector quantization, ten bits saving per frame can be achieved in comparison with the optimum scalar quantization.

(2) In the case of a speaker-dependent pattern matching vocoder, the spectral distortion decreases to the 60 % of that for the speaker independent case.

Several sonograms are shown in from Fig. 6.11 to Fig. 6.13. Figure 6.11 shows original speech sound, Fig. 6.12 and Fig. 6.13 show speech sonagrams synthesized by 270 and 1263 patterns, respectively. The lower parts of these figures show LSP parameters fluctuation. Both edges of each bar indicate the LSP parameter value. When the bar is short, there

is strong resonance in the speech spectra, as explained Chapter 4.

6.5 Optimum Bit Allocation to Reference Spectral Pattern

It is evident that reference spectral patterns have different occurences from its generation algorithm. Then, it is effective to allocate variable bit lengths according to the occurences of each vector to decrease the average information rate. The Huffman coding technique is well known to achieve this requirement. In speaker independent coding, the Huffman coding technique was applied at 58 and 1263 patterns. Maximum and mean code lengths are shown in Table 6.3. The mean code lengths were 3.5 and 5.24, respectively, in the case of speech without pause duration. In the case of speech data including pauses, 4 bits are saved per frame on an average. From these experiments, the technique was found effective for speech data including pause duration. One pattern, which has one bit code length, is mainly applied to frames in pause duration.

6.6 Conclusions

A pattern matching vocoder system using LSP parameters was proposed. A simple clustering technique was proposed for generating the discrete spectra patterns. This clustering algorithm was developed by taking into account acceptable spectral distortion and the Huffman coding.

(1) Spectral distortion by quantization was examined as a function of transmission bit rate. It was clarified that about 10 bits saving can be achieved by vector quantization, compared to optimum scalar

quantization in LSP vocoder.

(2) The difference between speaker-dependent and independent vector quantization was examined using the same speech utterances. According to the experimental results, two or three bits saving can be achieved with limited speaker.

(3) Huffman coding technique for spectral patterns was applied. This coding technique was effective for speech data which contains pause durations.

(4) Synthesized speech samples by pattern matching proposed here had fairly good speech quality and intelligibility. This speech quality is due to the good interpolation property of LSP parameters.

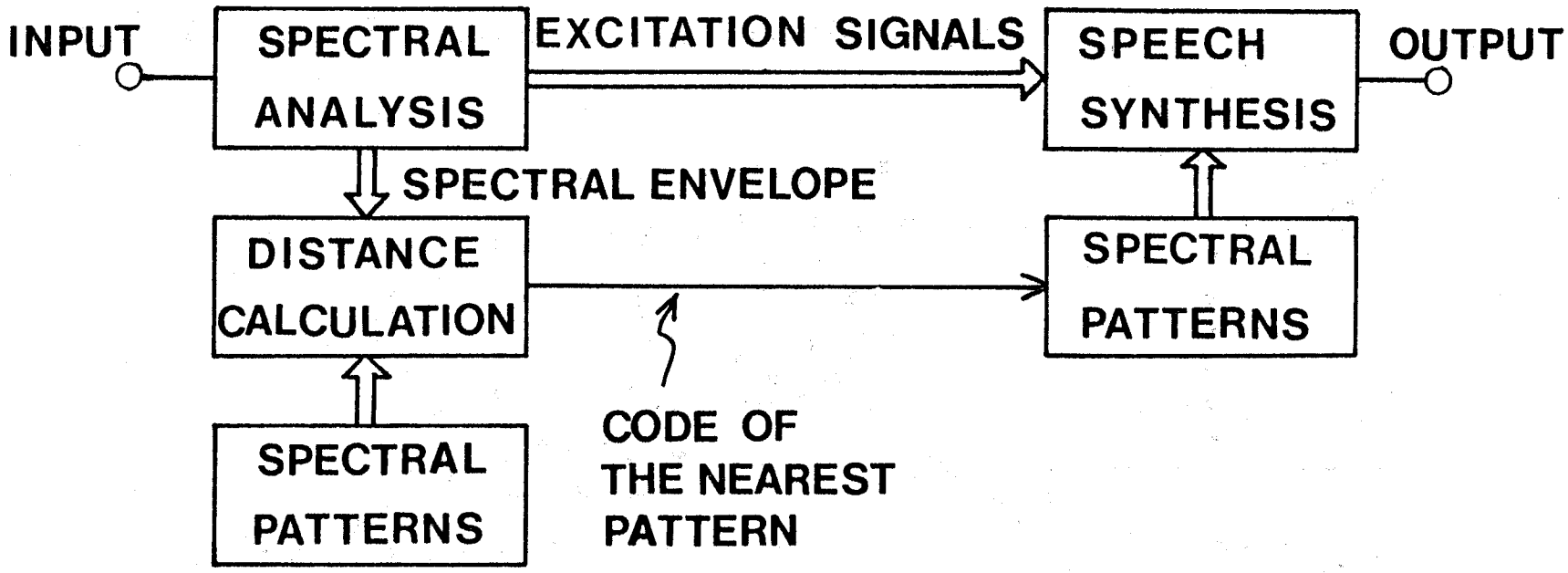


Fig. 6.1 Pattern matching vocoder system organization.

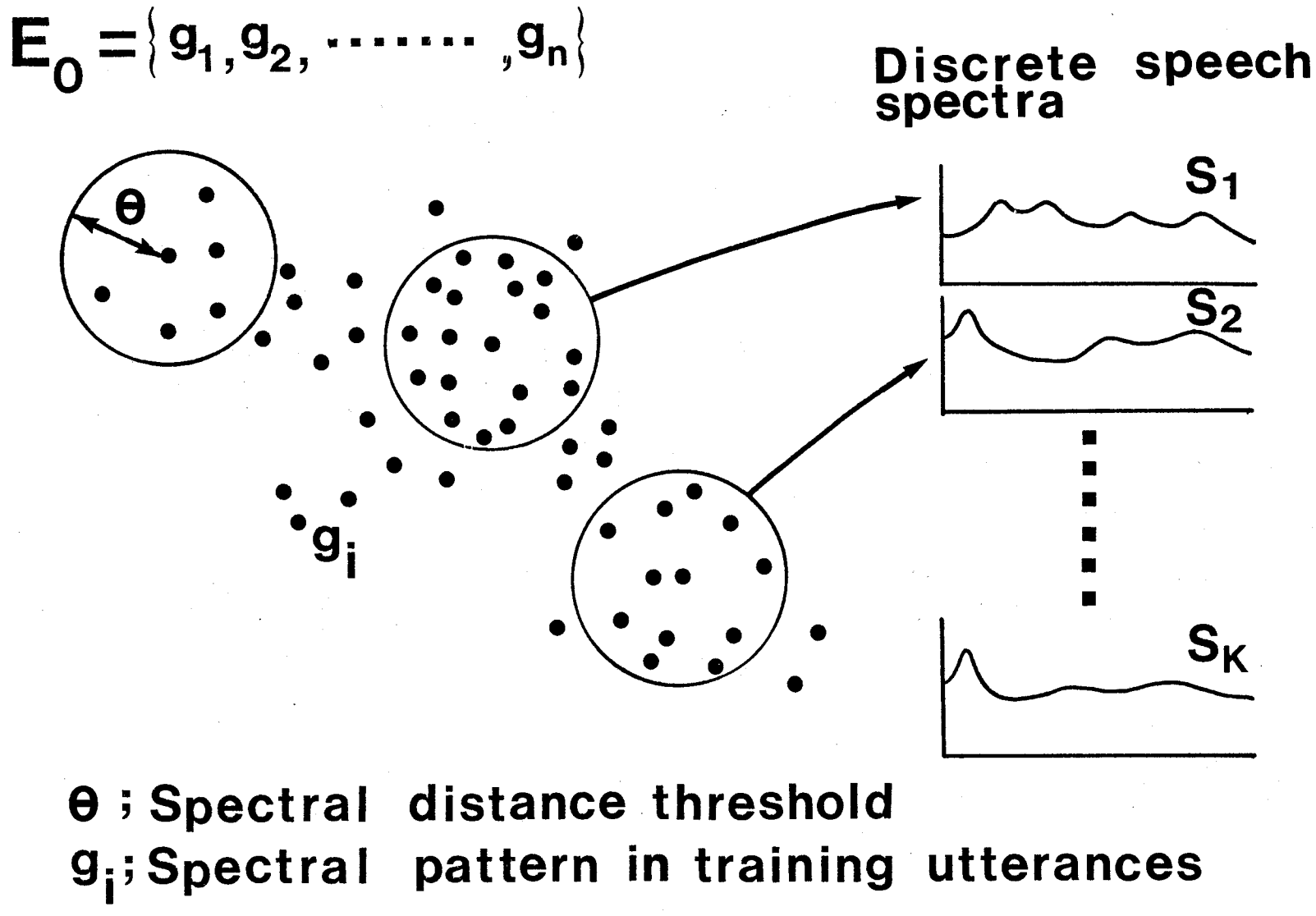


Fig. 6.2 Illustration of generating reference discrete spectra.

Table 6.1 Experimental conditions.

SPEECH DATA	
Contents	' bakuoNga giNsekaino koogeNni hirogaru '
Speakers	6 male speakers 6 female speakers
Contents	' chooseN naNbuni teikiatsuga atte kyusokuni hattatsu shinagara higashi naishi toohokutooni susuNde imasu '
Speakers	2 male speakers 3 female speakers
Contents	' Mukashi mukashi arutokoroni ojiisaNto obaasaNga imashita. OjiisaNwa yamae shibakarini obaasaNwa kawae seNtakuni dekakemashita. '
Speakers	6 male speakers 6 female speakers
Sampling frequency	8 kHz
Cut off frequency for low pass filter	4 kHz
AD conversion accuracy	14 bits
L S P ANALYSIS	
Frame period	10 mseconds
Window	30 mseconds Hamming window
Prediction order	10
Boundary condition at the glottis	$k_{p+1} = -1$ (complete opening)

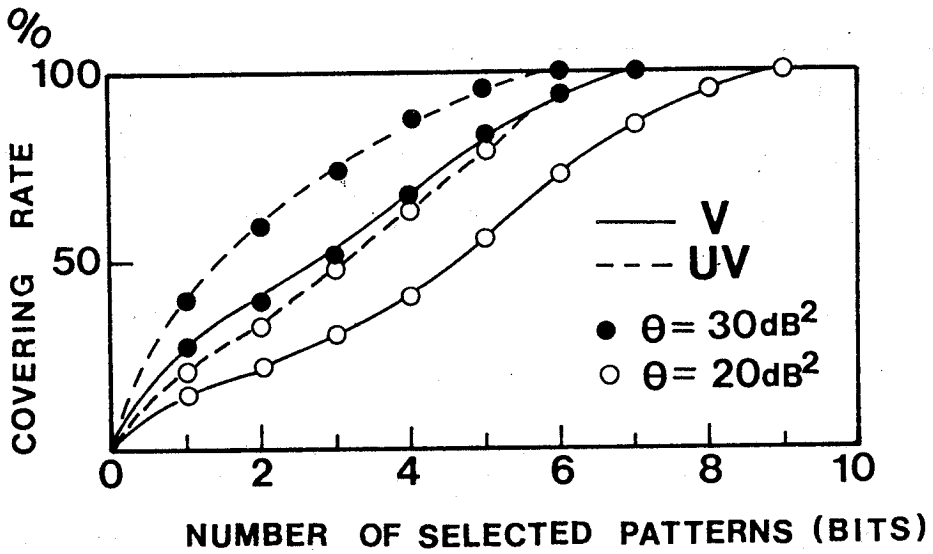


Fig. 6.3 Relationship between covering rate and the number of selected patterns.

Table 6.2 Selected reference pattern sets.

	NUMBER OF PATTERNS		TOTAL	BIT RATE	COVERING RATE (%)
	VOICED	UNVOICED			
1	49	9	58	5.9	75
2	147	27	174	7.4	80
3	237	43	280	8.1	85
4	318	59	377	8.6	90
5	466	84	550	9.1	95
6	714	146	860	9.8	100

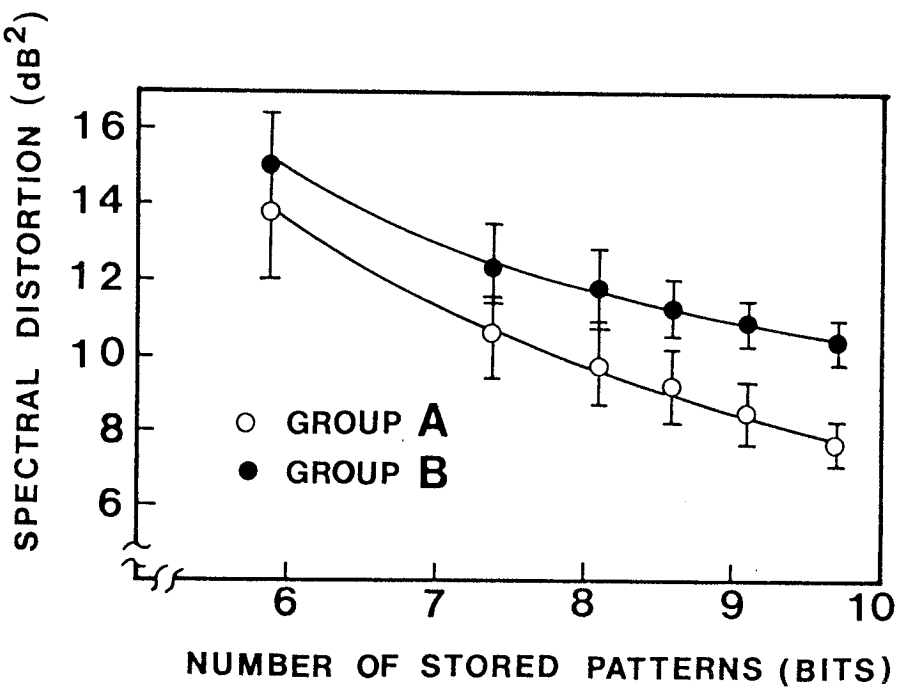


Fig. 6.4 Spectral patterns by vector quantization for Group A (trained) and Group B(untrained) speakers using central reference patterns.

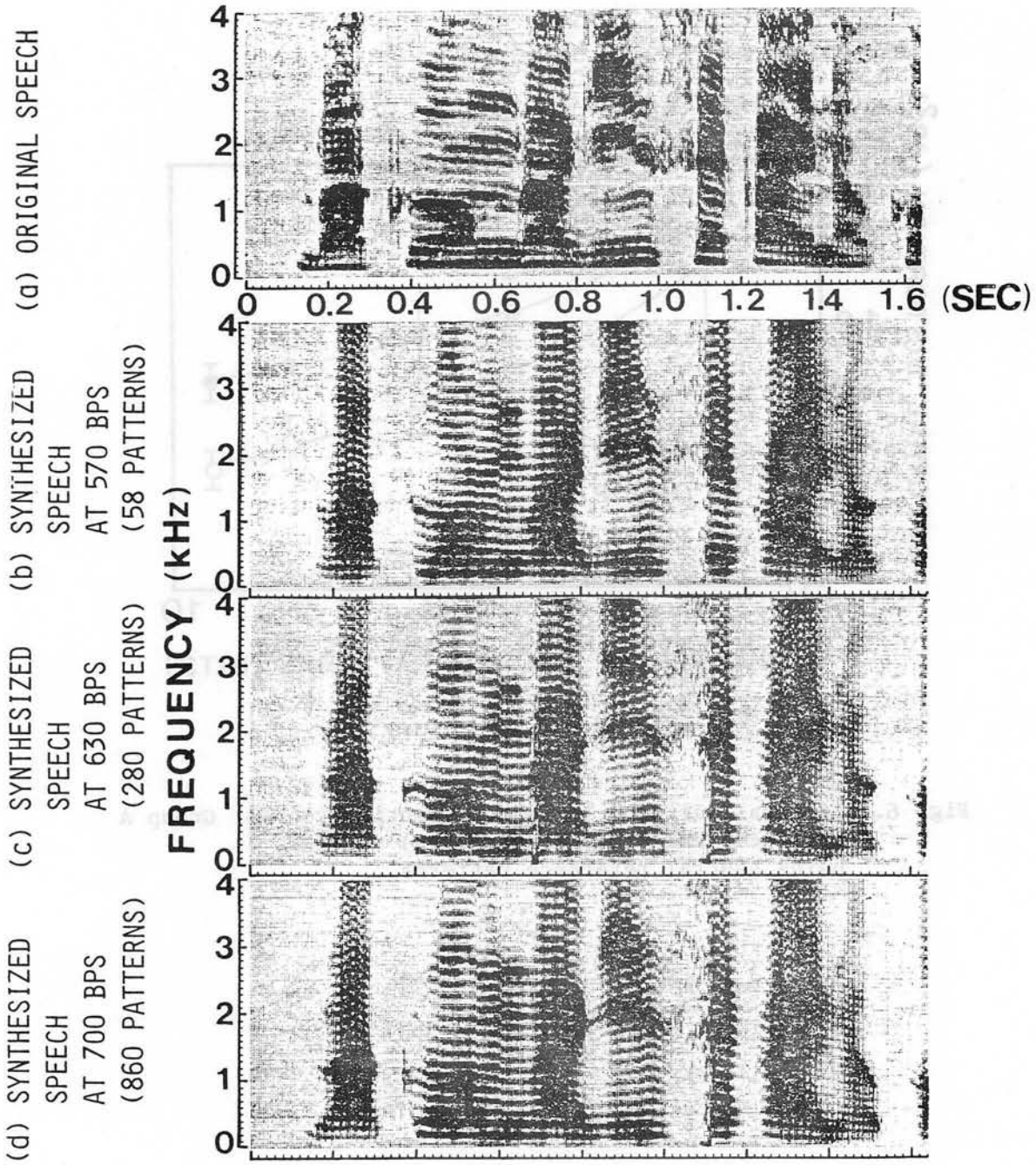


Fig. 6.5 Sonograms comparison between original and synthesized speech by vector quantization (male speaker, frame length -- 30 mseconds, pitch-- 6 bits, power-- 5 bits).

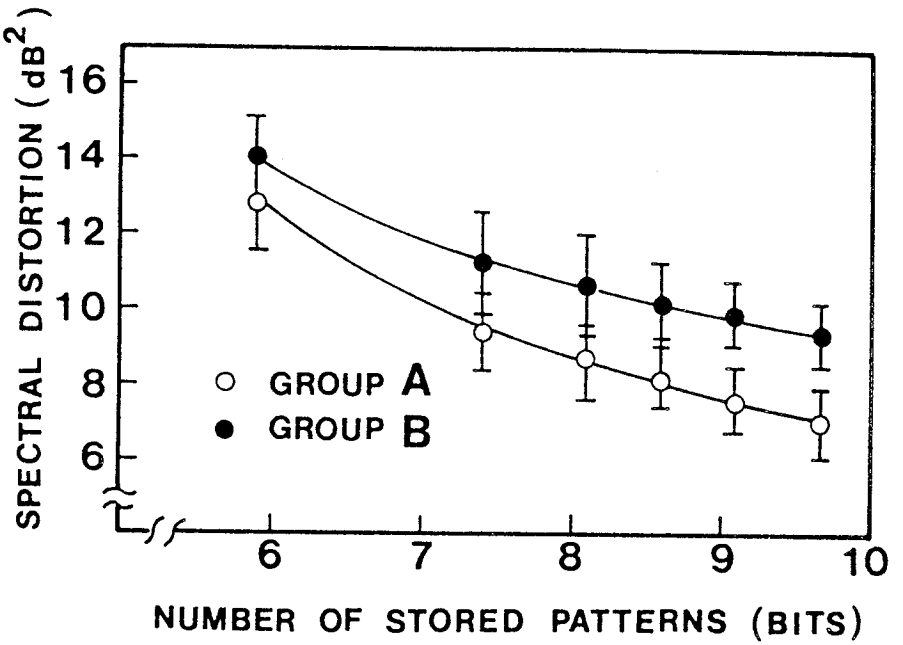


Fig. 6.6 Relationship between the number of stored patterns and spectral distortion, when using averaged patterns.

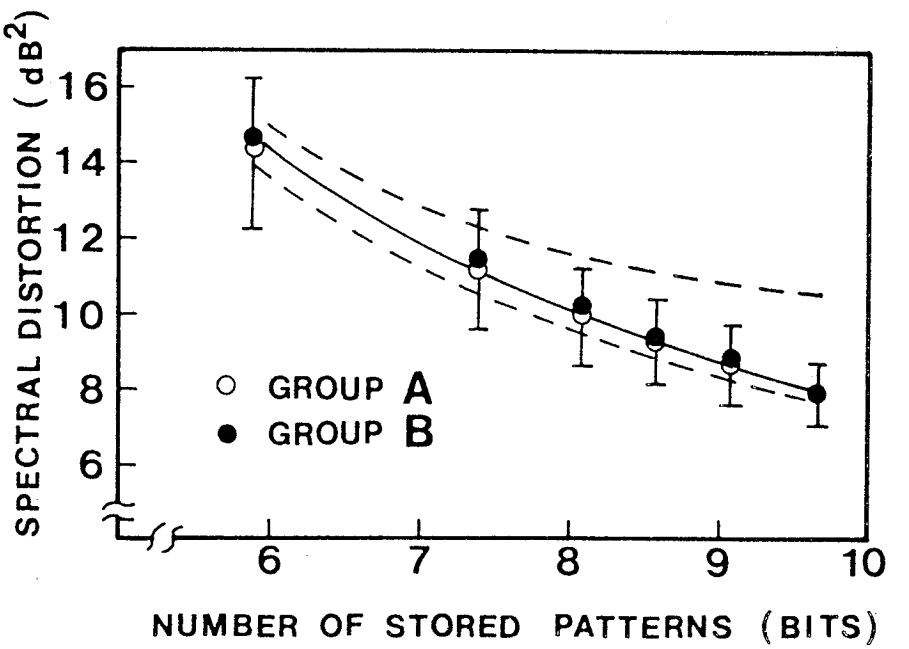


Fig. 6.7 Relationship between the number of stored patterns and spectral distortion, when using re-generated patterns.

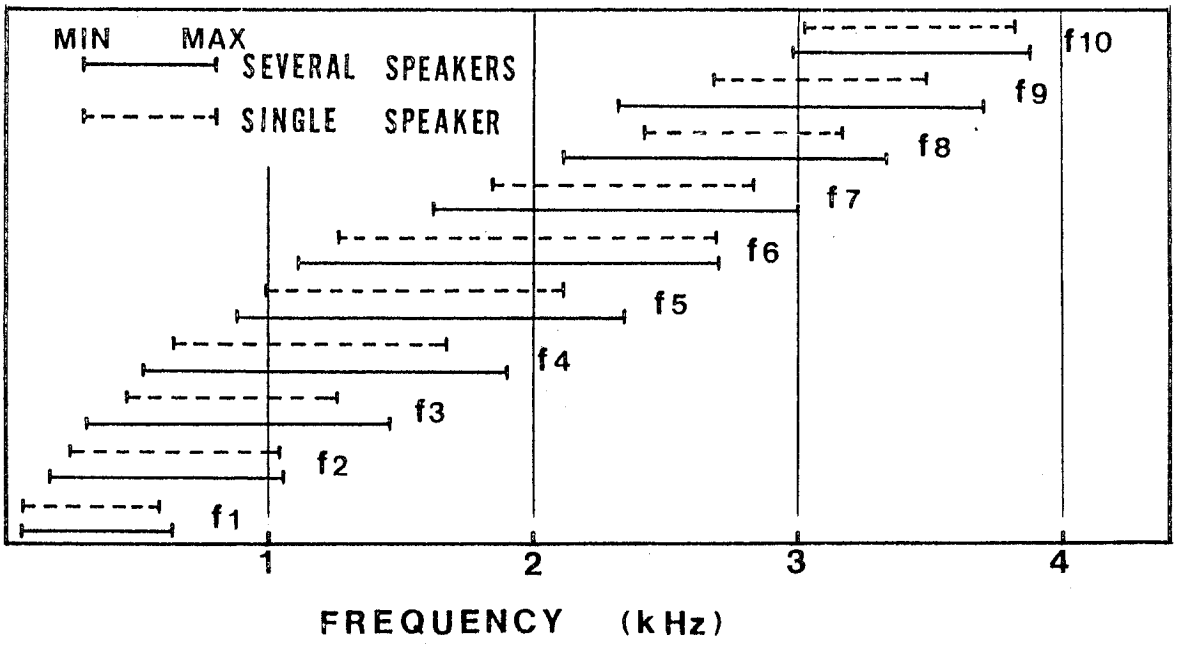


Fig. 6.8 Comparison between LSP parameter distribution ranges.

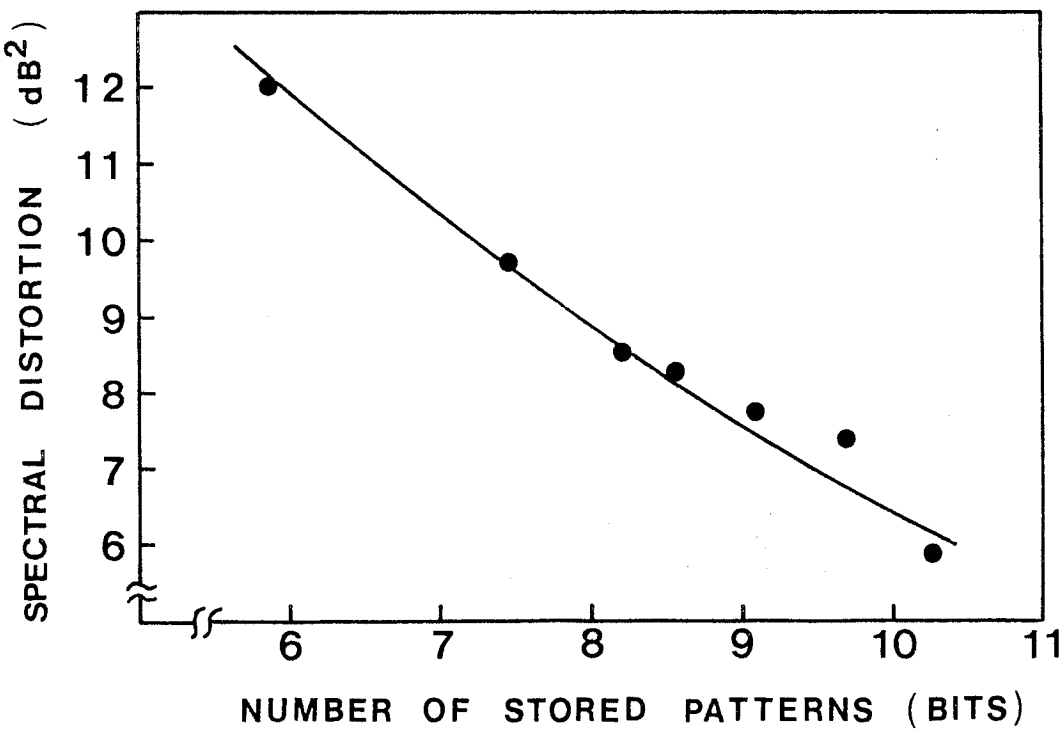


Fig. 6.9 Spectral distortion versus number of stored patterns for a single speaker.

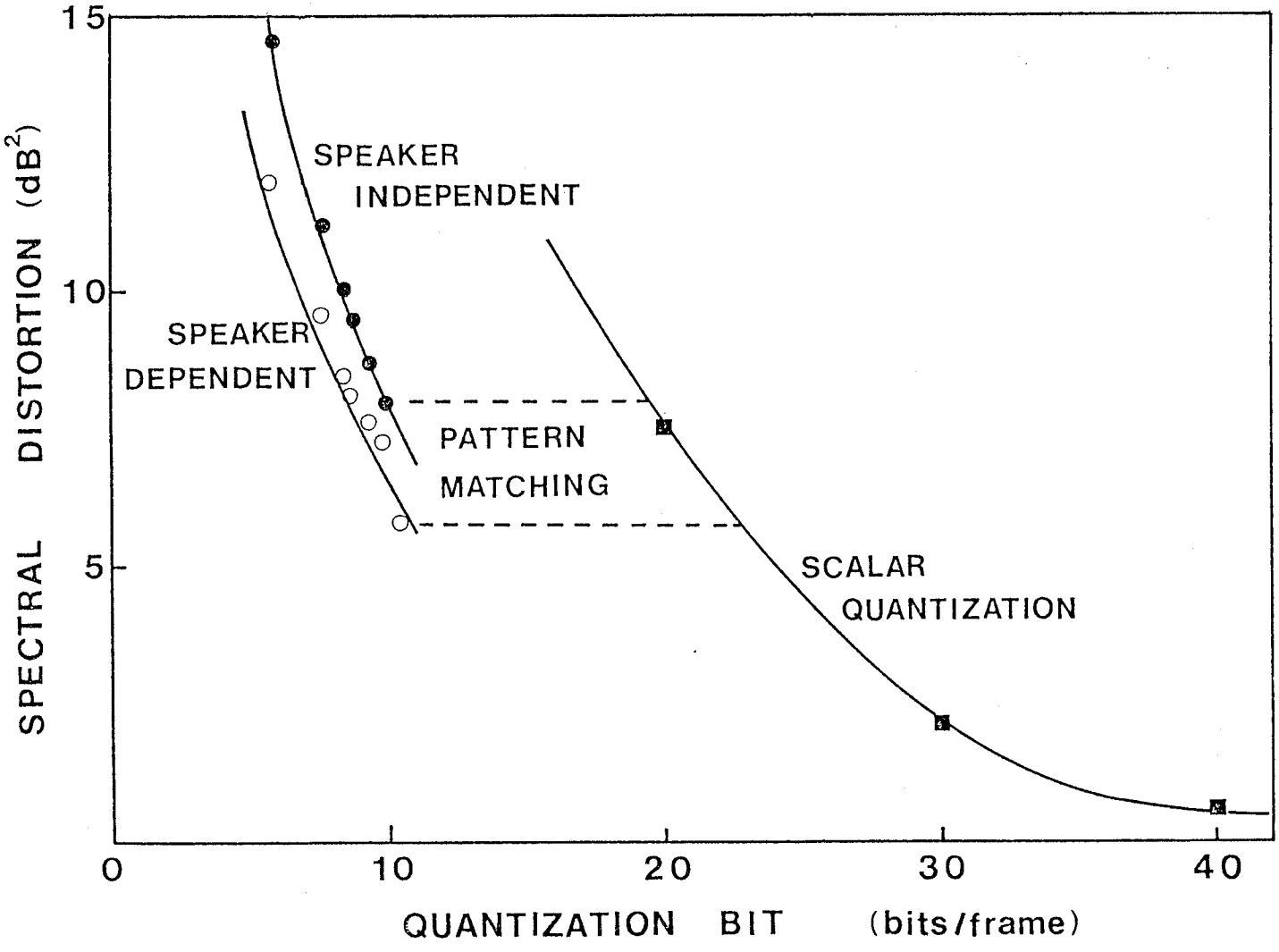


Fig. 6.10 Vector and scalar quantization comparison with respect to bit rate and spectral distortion.

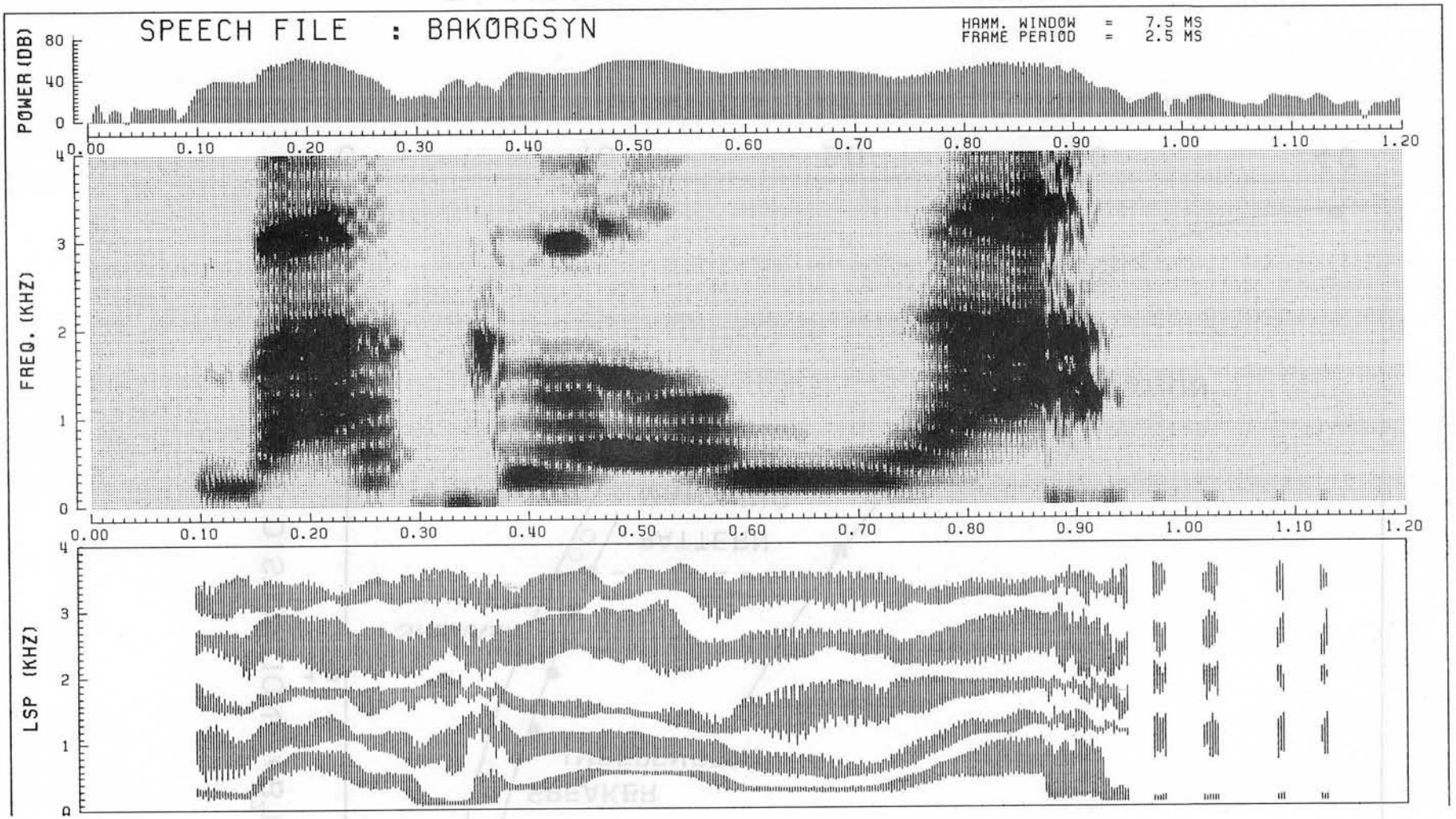


Fig. 6.11 Sonograms of original speech sound uttered by a female speaker.

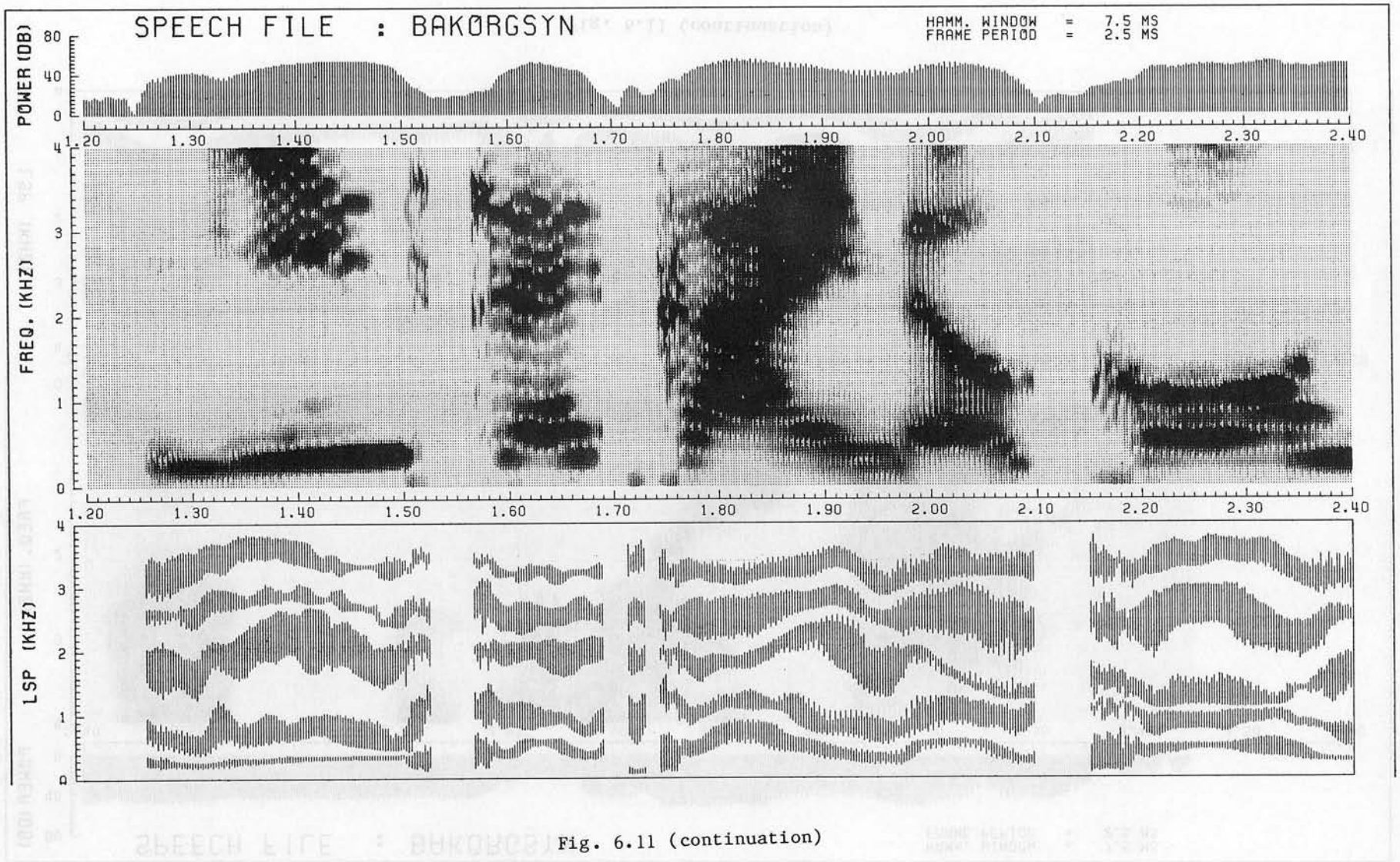


Fig. 6.11 (continuation)

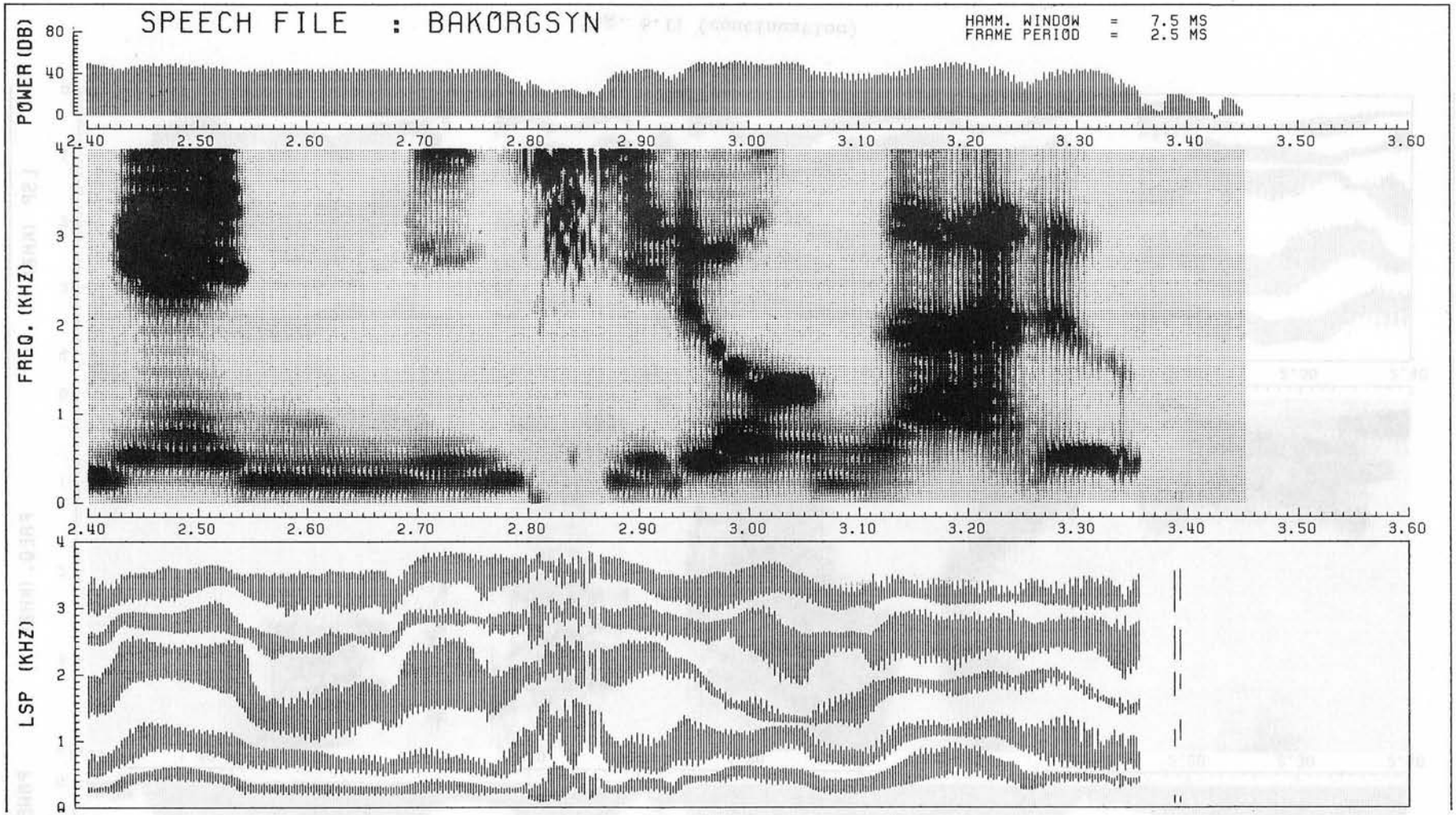


Fig. 6.11 (continuation)

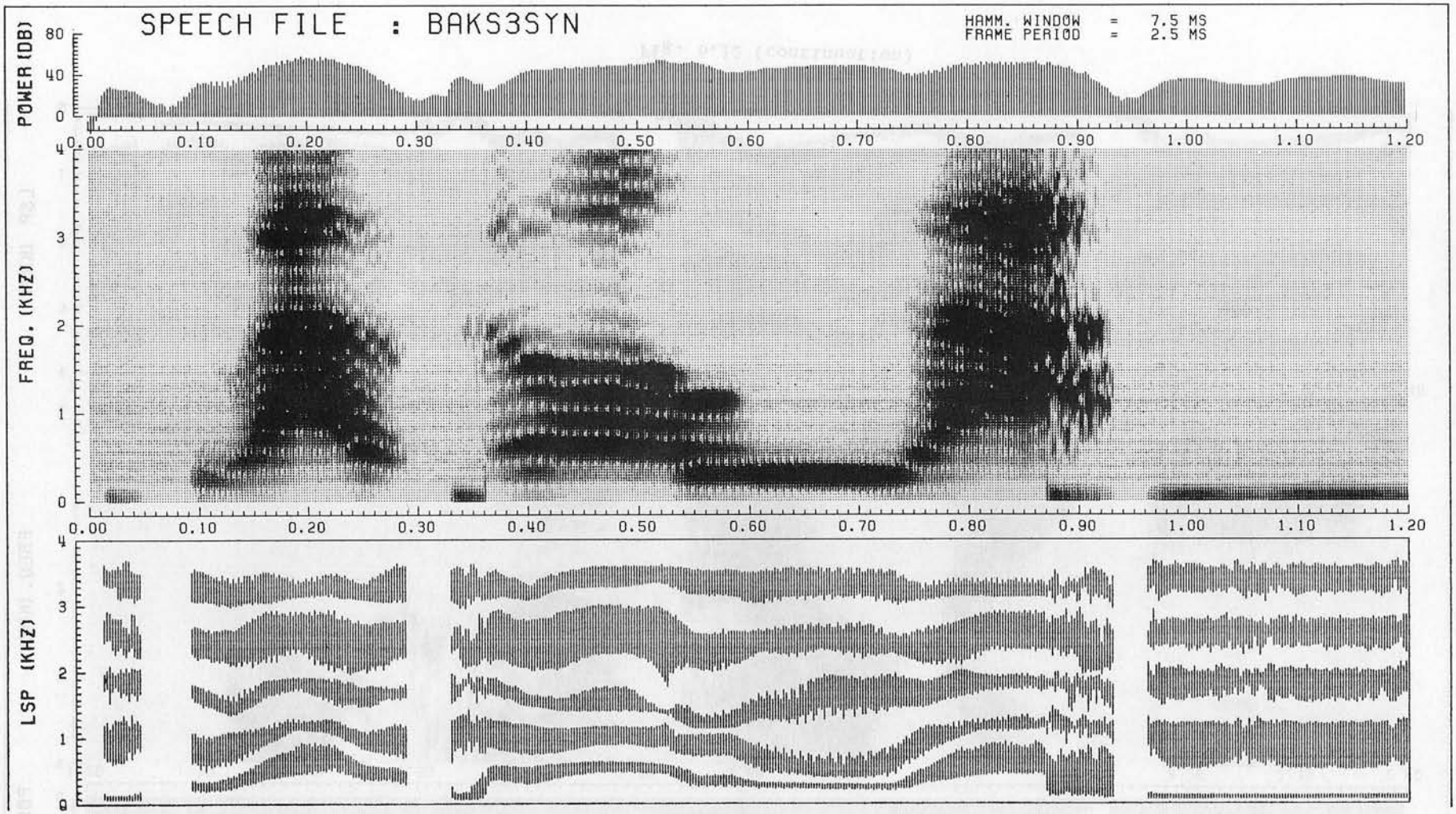


Fig. 6.12 Sonograms of synthesized speech sound by 280 discrete spectra

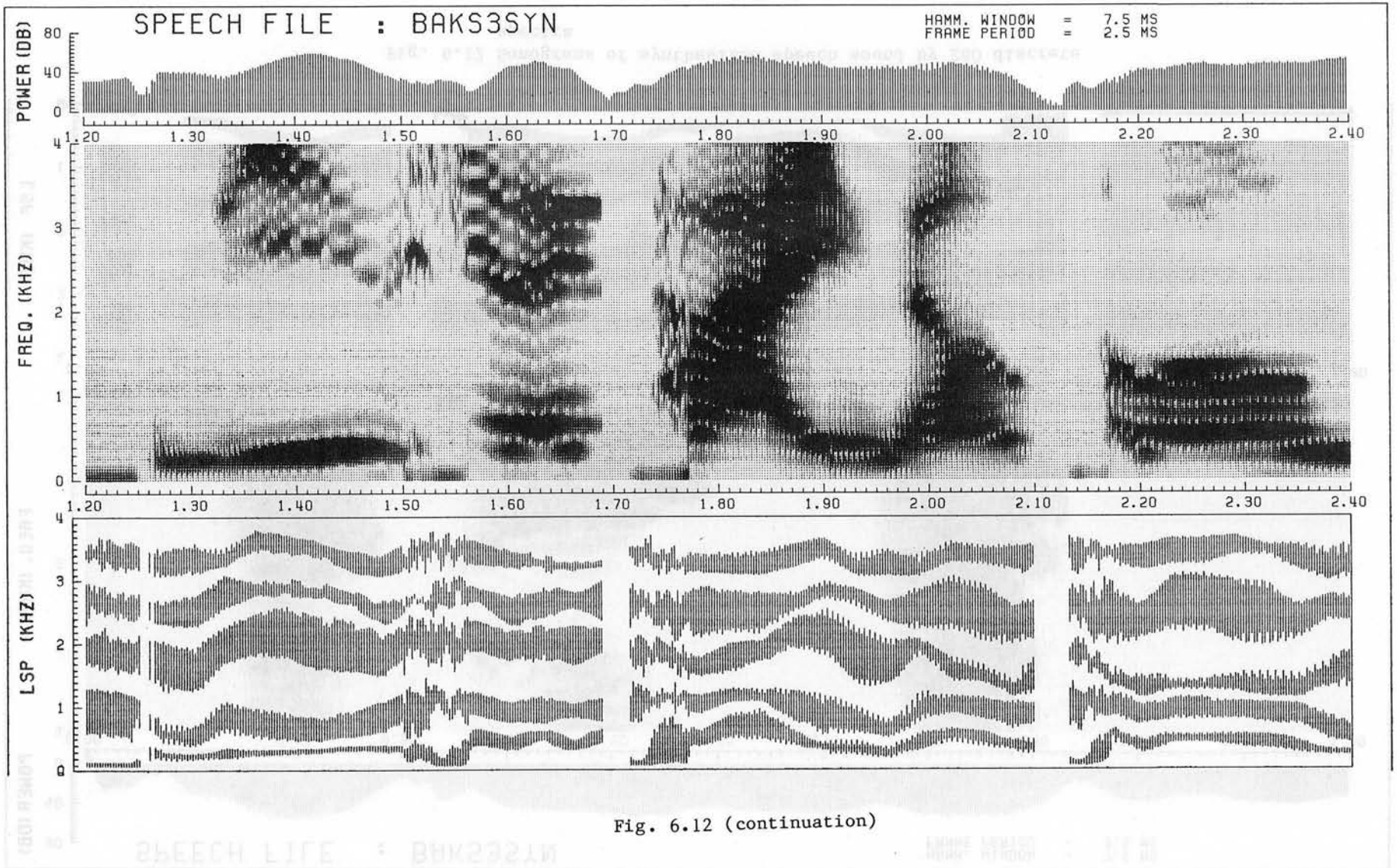


Fig. 6.12 (continuation)

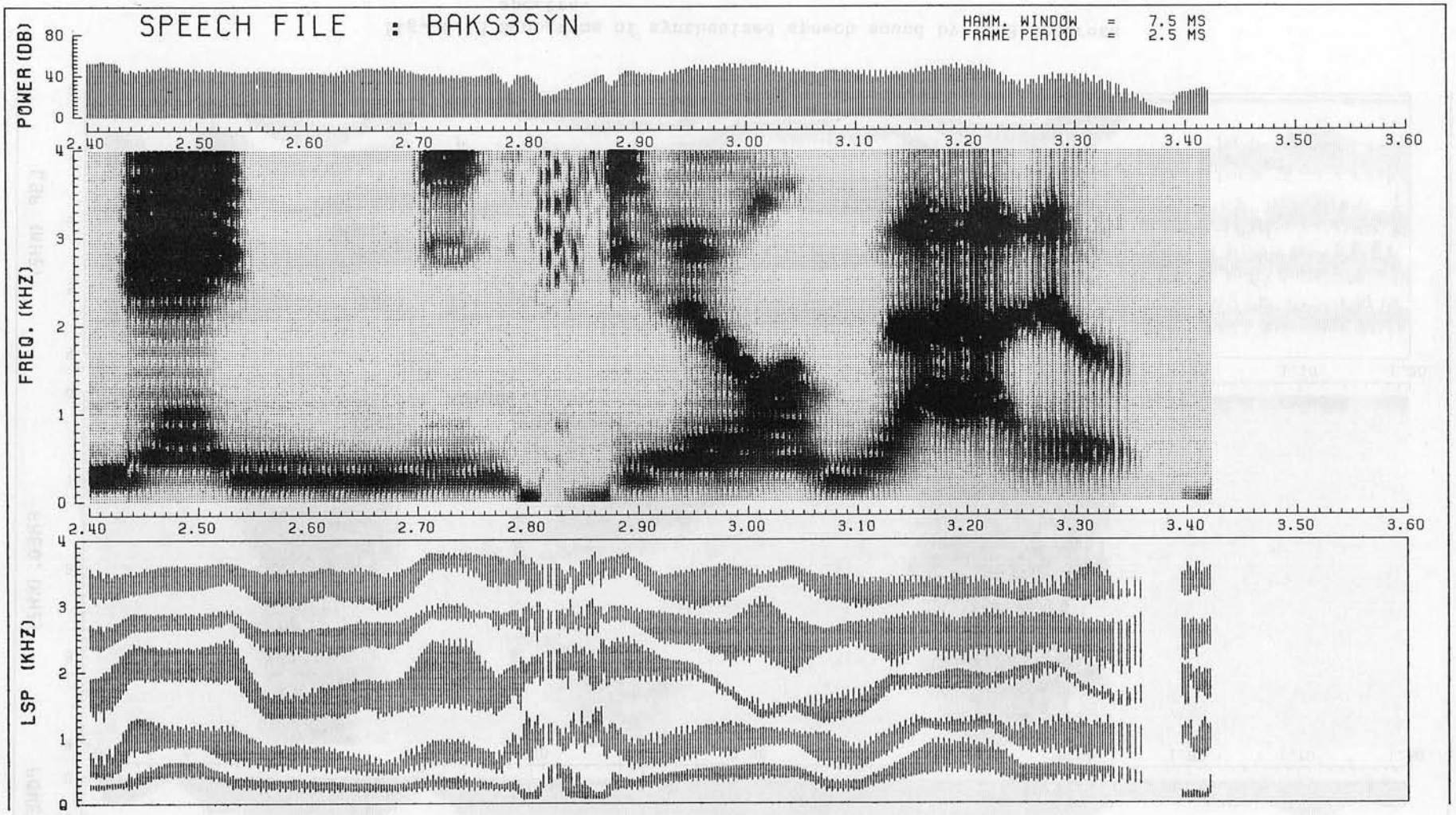


Fig. 6.12 (continuation)

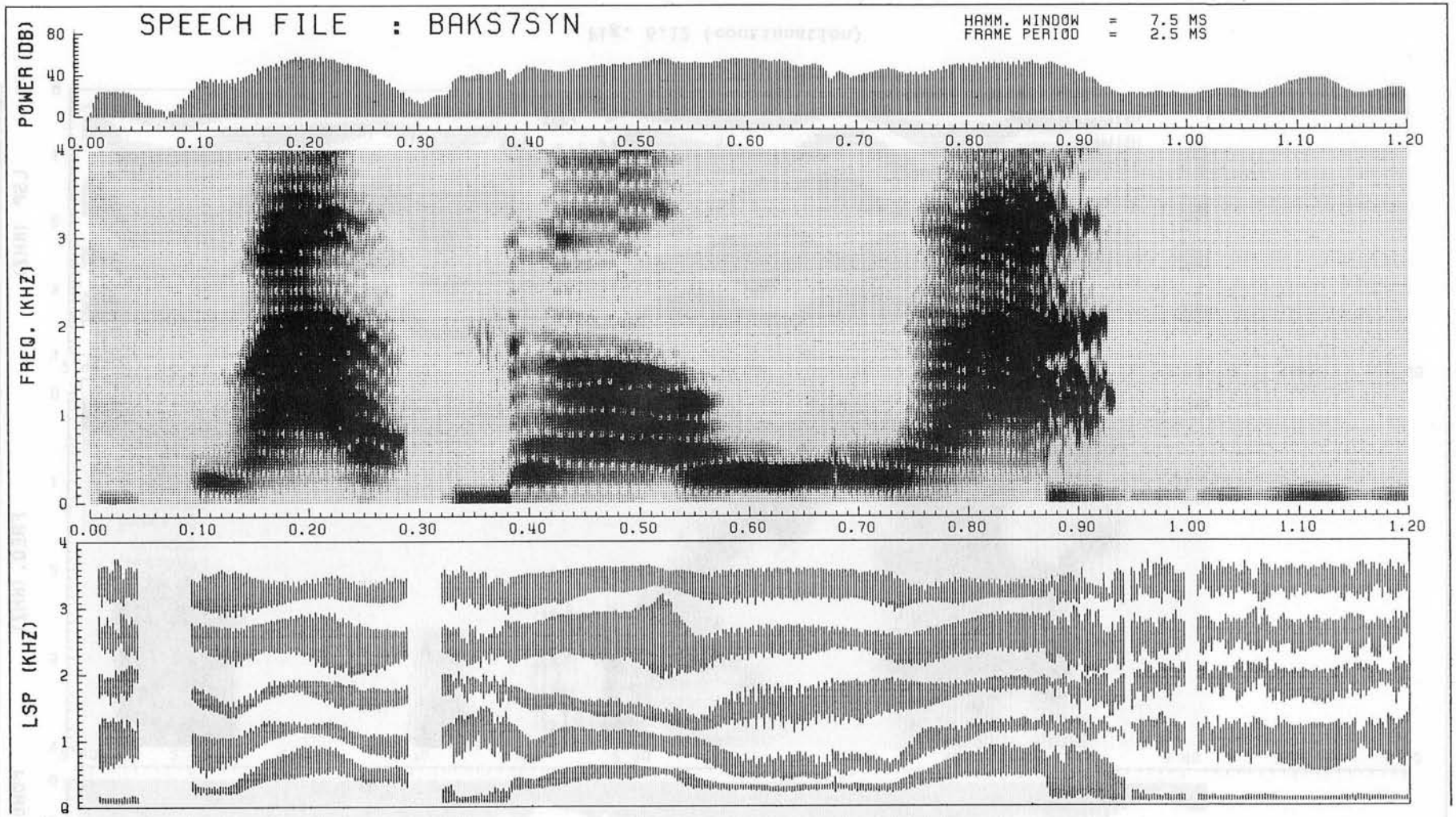


Fig. 6.13 Sonograms of synthesized speech sound by 1263 discrete spectra.

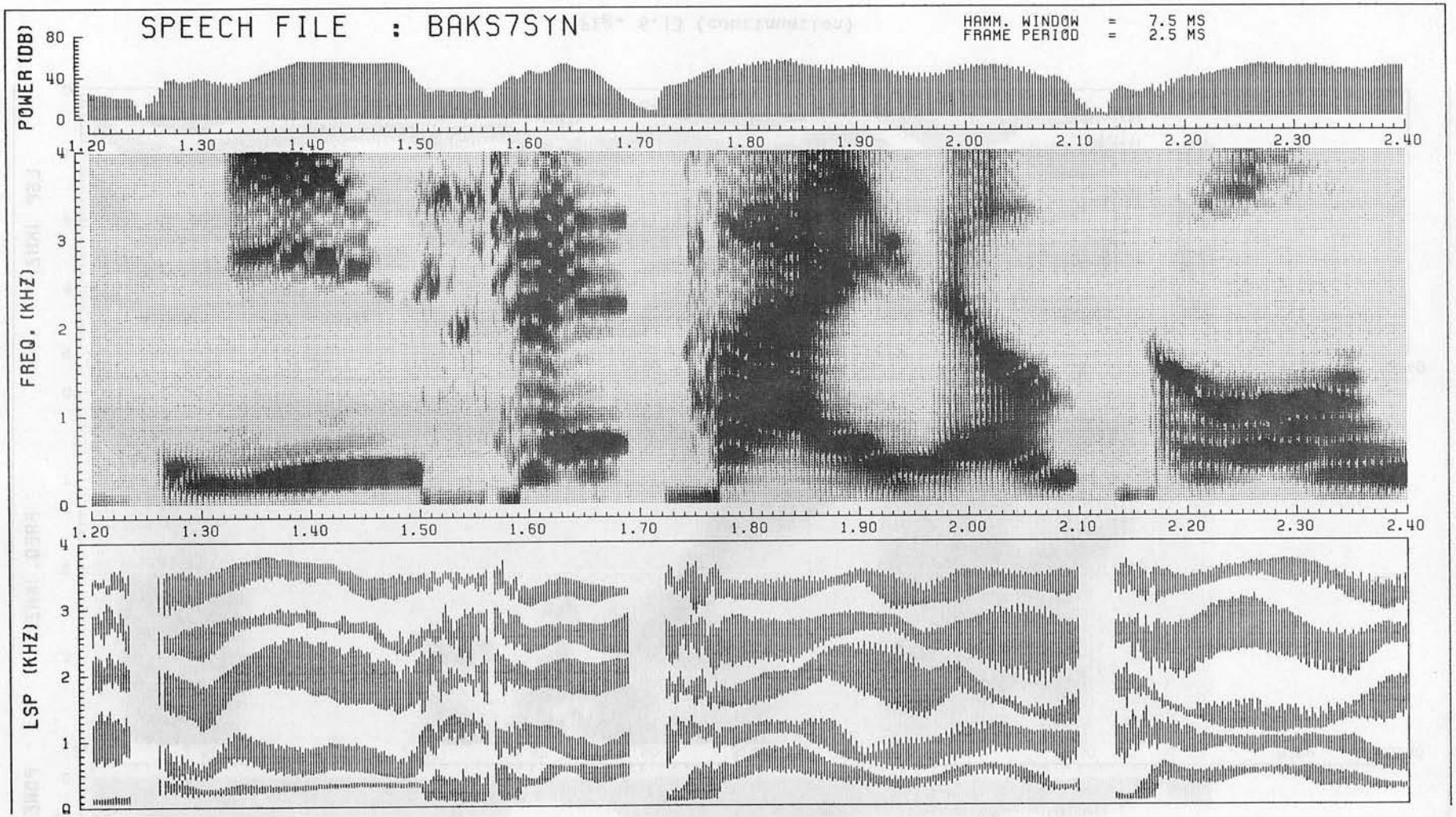


Fig. 6.13 (continuation)

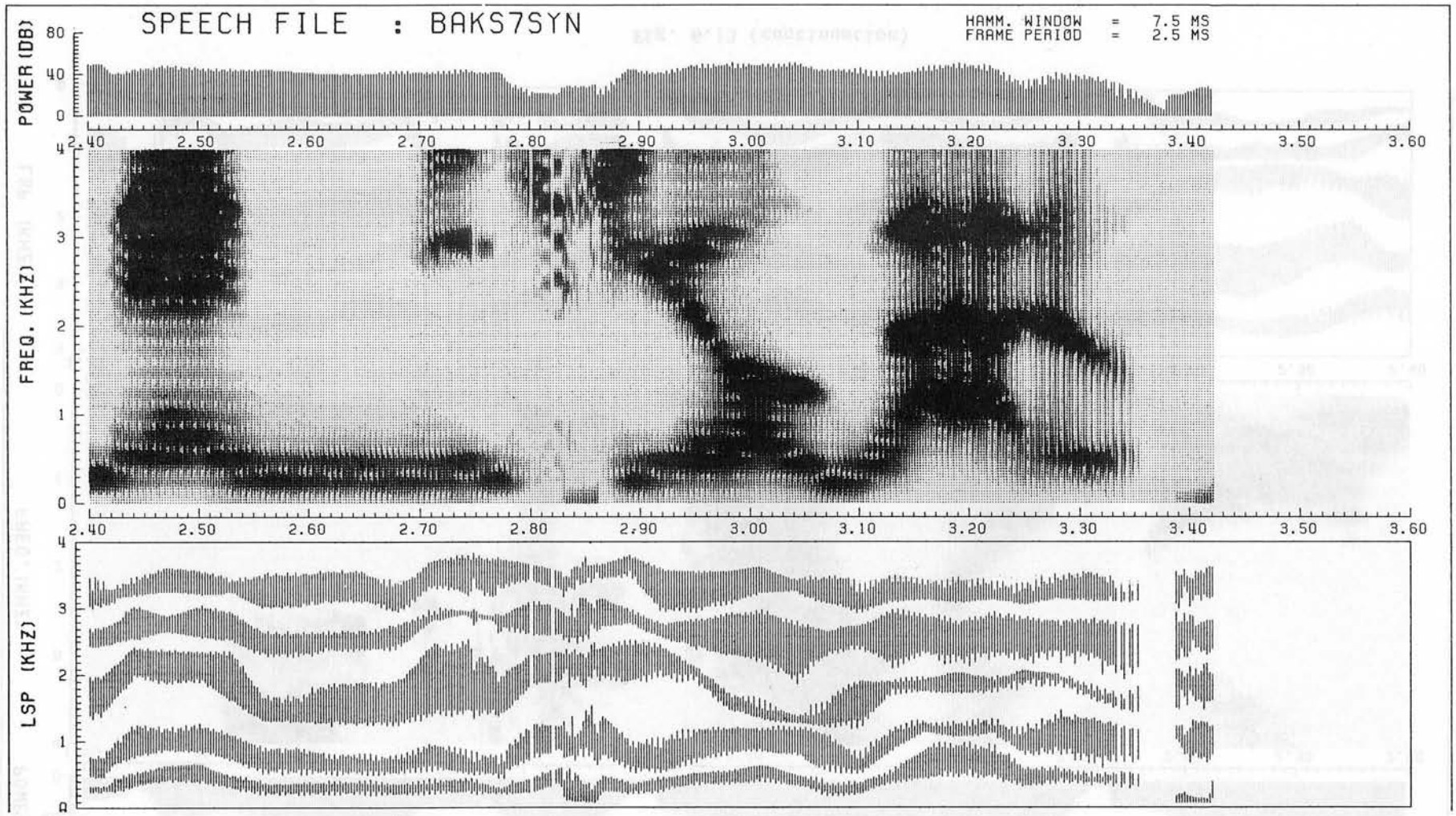


Fig. 6.13 (continuation)

Table 6.3 Code lengtgh in applying Huffman coding technique to reference spectral patterns.

NUMBER OF STANDARD PATTERNS	SPEECH DATA	CODE LENGTH	
		MAXIMUM	MEAN
58	INCLUDING PAUSE	11	3.50
	EXCLUDING PAUSE	10	5.54
1263	INCLUDING PAUSE	16	5.24
	EXCLUDING PAUSE	14	9.75

CHAPTER 7 ISOLATED WORD RECOGNITION USING STRINGS OF PHONEME-LIKE TEMPLATES (SPLIT)

7.1 Introduction

In this chapter, the vector quantization technique mentioned in Chapter 6 is applied to an isolated word recognition system. The basic idea of this system is to represent a speech spectral sequence by several discrete spectra symbols. Namely, in this system, word templates are represented as sequences of discrete phoneme-like templates (pseudo-phoneme), which are automatically selected from a training set of word utterances by a clustering technique. A new word recognition system and its feature are explained. This recognition system is especially effective in speaker-dependent large vocabulary word recognition or speaker-independent word recognition based on multiple word templates.

7.2 Isolated Word Recognition [69],[70],[71],[72]

In most isolated word recognition systems, pattern matching techniques based on dynamic time warping are used. Whole vocabulary words are assumed to be uttered in advance for training. Feature parameters, such as band pass filter outputs or LPC parameters, are extracted from those training utterances and stored frame by frame as word templates. In the recognition stage, a time sequence of feature parameters extracted from the input speech and word templates are directly compared. This method is called "direct-matching" in this thesis. This method configuration is shown in Fig. 7.1(a). In the

direct-matching method, since spectral information for every frame is precisely described using extracted parameters themselves, high recognition accuracy can be obtained. However, in the large vocabulary word recognition, the amount of spectral distance calculation for dynamic time warping and the memory size for the word templates become very large, because spectral calculation must be executed in every frame for every template. Another recognition system was also reported to avoid these problems [73],[74]. In this system, phoneme spectral patterns and word templates, which contain the duration time for each phoneme are used. However, it is sometimes difficult to determine phoneme spectral patterns and word templates.

Apart from the word recognition field, narrow band transmission has been studied based on vector quantization techniques mentioned in Chapter 6. Through the examination, the speech has been ensured to be intelligible at an extremely low bit rate, under 1000 bps. After these experiments, it was conceived to represent word templates in speech recognition as sequences of discrete spectral patterns, which are vector quantized roughly. This system configuration is shown in Fig. 7.1(b), compared with a conventional system shown in Fig. 7.1(a). A new word recognition system, using Strings of Phoneme-LIke Templates, is named the SPLIT method. SPLIT is also a term derived from SPLITting the feature space where many spectra are distributed.

7.3 New Isolated Word Recognition System, SPLIT [75],[76]

A SPLIT system is shown in Fig. 7.2 in detail. In the first stage, phoneme-like templates and word templates are generated, using previously determined training speech samples.

7.3.1 Generating Phoneme-like Templates

The phoneme-like templates are generated automatically. The same clustering algorithm, as that introduced in Chapter 6, is applied to generate phoneme-like templates. Phoneme-like templates do not exactly correspond to real phonemes. They are only used as symbols representing discrete speech spectra here.

After this work, another vector quantization technique [77] was applied to generate phoneme-like templates in the SPLIT method. Differences between two clustering methods were examined as functions of spectral distortion and recognition accuracy [78].

7.3.2 Generating Word Templates

Word templates are represented as sequences of phoneme-like templates. Namely, each training word utterance is divided into a 16 mseconds duration succession and spectral distance between each frame and each phoneme-like template is computed. The symbol for phoneme-like template, which minimize the spectral distance, is stored in every frame. Word templates are stored as sequences of phoneme-like templates not using the exact spectral parameters.

7.3.3 Word Recognition

The phoneme-like templates and word templates are stored at (c) and (d), respectively, in Fig. 7.2. Word recognition is carried out as follows.

The input word utterance is analyzed every 16 mseconds and autocorrelation coefficients and LPC cepstrum coefficients are extracted at (e) in Fig. 7.2. The spectral distance between individual input word frames and individual phoneme-like templates is stored as an element of a distance matrix in (f) in Fig. 7.2. In the experiments, two kinds of spectral distance measures were used. One is cepstrum distance measure (CEP) and the other is Weighted Likelihood Ratio (WLR) [79]. These spectral distance measures are defined as :

$$CEP = 2 \sum_{i=1}^{P} (C_i^{(f)} - C_i^{(g)})^2 \qquad (7.1)$$

$$WLR = 2 \sum_{i=1}^{P} (C_i^{(f)} - C_i^{(g)}) (V_i^{(f)} - V_i^{(g)}) \qquad (7.2)$$

where $C_i^{(f)}$ and $C_i^{(g)}$ are LPC cepstrum parameters and $V_i^{(f)}$ and $V_i^{(g)}$ are autocorrelation coefficients. Suffixes (f) and (g) correspond to input word utterances and phoneme-like templates, respectively.

7.3.4 Dynamic Time Warping [80],[81]

In isolated word recognition system using dynamic time warping, the recognition accuracy is influenced by the efficiency of the dynamic time warping algorithm. In the present experiments, efficient slope-constrained dynamic time warping is used. (Appendix 3)

7.3.5 Several Features of the SPLIT Method

The SPLIT method has the following significant features over the direct-matching method.

(1) Memory saving for word templates

In the SPLIT system, word templates are represented by sequences of phoneme-like templates. Thus, drastic memory saving can be achieved in comparison with the direct-matching system. The saving ratio is calculated approximately as a function of the number of word templates.

Assume the vocabulary size is L words. The i-th word has M_i frames. Each frame is N dimensional feature vector and its accuracy is N_a bits. Denote the number of phoneme-like templates as N_S. Using these notations, the memory amount for word templates in the direct-matching method (R_d) and SPLIT method (R_S) are given as

$$R_d = \left(\sum_{i=1}^{L} M_i \right) N N_a \tag{7.3}$$

$$R_S = N_S N N_a + \left(\sum_{i=1}^{L} M_i \right) n_b \tag{7.4}$$

where $n_b = \log_2 N_S$

The reduction ratio for the SPLIT method to the direct-matching method is

$$R_{MEM}(L) = \frac{N_S N N_a + \left(\sum_{i=1}^{L} M_i \right) n_b}{\left(\sum_{i=1}^{L} M_i \right) N N_a} \tag{7.5}$$

(2) Distance calculation saving for a dynamic time warping

In the SPLIT system, spectral calculation amount depends only on the number of phoneme-like templates. The savings ratio is calculated as follows.

The frame number for input speech is M_l. The window length in

dynamic time warping is assumed to be N_W. Calculation amounts are represented, respectively, by :

$$C_d = M_\ell \, N_W \, L \tag{7.6}$$

$$C_S = M_\ell \, N_S \tag{7.7}$$

where C_d is the distance calculation amount in the direct-matching method and C_S is the calculation amount in the SPLIT method.

The reduction ratio is

$$R_{CAL}(L) = \frac{M_\ell \, N_S}{M_\ell \, N_W \, L} = \frac{N_S}{N_W L} \tag{7.8}$$

The reduction ratio for spectral distance calculation and memory amount for word templates, compared with direct-matching method, is shown in Fig. 7.3 as a function of the number of words to be recognized. In this figure, parameter values are assumed to be :

M_i = 50 frames (for all i,for simplicity)

N = 16 parameters

N_a = 16 bits

N_S = 256 templates

N_W = 15 frames

n_b = 8 bits

7.4 Large Vocabulary Size Speaker-Dependent Word Recognition

7.4.1 Using 256 Phoneme-like Templates

In order to verify the SPLIT system ability in large vocabulary word recognition, 641 city name utterances (Appendix 4) by four male speakers were used. Every speaker uttered each word twice at two week intervals. Experimental conditions are as follow.

Input speech is band limited to 4 kHz, sampled at 8 kHz and then converted into digital signals by a 12 bit AD converter. After passing the 32 msecond Hamming window, thirteen auto-correlation coefficients are derived every 16 mseconds. After these processes, the 10th order LPC analysis is executed. The first utterance set is used in generating phoneme-like templates and word templates. The second utterance set was used for an evaluation at the first experiment. At the second experiment, such conditions were inverted. In the experiments, 2048 frames were used for generating phoneme-like templates and 256 phoneme-like template were generated for every speaker. In these experiments, the following points were tested.

(1) Comparison of recognition accuracies between SPLIT method and the direct-matching method.

(2) Comparison of the spectral measure efficiency between cepstrum distance and WLR distance.

Experimental results are concentrated in Fig. 7.4. This figure shows the relationship between ranking and the accumulated recognition accuracy. Based on these results, the following conclusions were reached.

(1) 96.3 % recognition accuracy can be obtained in 641 word recognition on an average for four speakers by the SPLIT method, using

WLR spectral distance measure. Degradation from the direct-matching method was only 0.4 %. These results mean that spectral information can be roughly quantized in each frame without largely decreasing recognition accuracy.

(2) 98.3 % recognition accuracy can be obtained within the top five choices.

(3) Recognition accuracy using WLR is much higher than that for cepstrum distance in both systems. This means that the WLR spectral distance measure is effective in large vocabulary word recognition.

7.4.2 Classification of Unrecognized Words

Words which were not recognized correctly in the experiments are classified into the following categories.

(A) Endpoint detection error	(ex. SAKU → KUSATSU)
(B) Vowel confusion	(ex. UOZU ⟶ OOZU)
(C) Head consonant confusion	(ex. SAGA ⟶ KAGA)
(D) Consonant confusion, other than (3)	(ex. OOGAKI→OOMACHI)
(E) Others segregation categories	(ex. MINOO ⟶ MINO)

Classification results for four speakers are shown in Fig. 7.5. From this figure, the following conclusion can be obtained.

(1) Endpoint detection error is about 1 % of all errors.

(2) Vowel confusion were very few.

(3) Consonant confusions decrease, when using WLR spectral measure.

7.4.3 For Fewer Than 256 Phoneme-like Templates

The relation between the number of phoneme-like templates and the spectral distortion in generating word templates was examined and the results are shown in Fig. 7.6. The points indicated by the arrow show the maximum recognition accuracy. In this figure, strong correlation is observed between spectral distortion in generating word templates and recognition accuracy.

Next, the recognition accuracy was investigated versus the number of phoneme-like templates. Five kinds of phoneme-like template sets were generated, each of which consist of 16, 32, 64, 128 or 256 templates, by changing the threshold value θ for clustering for one of four speakers. θ values were set from 0.05 to 0.30 at every 0.05 intervals. Using these six phoneme-like template sets, recognition experiments were executed for one of the four speakers. The relationship between recognition accuracy and the number of phoneme-like templates is shown in Fig. 7.7. This result shows that high recognition accuracy can be obtained when there are fewer than 64 phoneme-like templates, if θ is set appropriately. Furthermore, even when there are 16 phoneme-like templates, 94.0 % recognition accuracy was obtained. Surprisingly, this number is much less than Japanese phonemes number. This result means the spectral pattern sequence has as much information which characterizes a word as spectral accuracy in each frame.

Based on results obtained from this experiment, phoneme-like templates were generated, so that total spectral distortion in the word templates, represented by the phoneme-like template sequences, became minimum. Using these optimum template sets, recognition experiments were carried out by the utterances of the same four speakers. The

relation between the averaged recognition accuracy and the number of phoneme-like templates is shown in Fig.7.8 and Table 7.1.

This figure shows that the recognition accuracy does not decrease rapidly due to the decreasing in the number of phoneme-like templates, and 92.9 % recognition accuracy was obtained, even when there were 16 phoneme-like templates.

When there were 32 phoneme-like templates, degradation from the direct-matching method was 2.2 %, where the amount of distance calculation and memory size for word templates are 0.3 % and 2% of the direct-matching method, respectively. The small number of phoneme-like templates can be used in several application fields.

7.4.4 Relationship Between Phoneme-like Templates and Real Phonemes

The relation between 32 phoneme-like templates and real phonemes was investigated. Each frame in several words was labeled phonemes through a human process. Correspondence between phoneme-like template and phoneme in each frame was investigated. The selected 32 phoneme-like templates are mapped on a two dimensional plane by multiple dimensional analysis, as shown in Fig. 7.9. Five or six phoneme-like templates belong to the same vowel cluster and several templates correspond to consonants, such as / s /,/ m,n /,/ p,t,k /. Spectral envelopes of these 32 phoneme-like templates are shown in Fig. 7.10.

7.5 Speaker-Independent Isolated Word Recognition Based On Multiple Word Templates

This section describes a speaker-independent word recognition system based on multiple word templates using the SPLIT method. One powerful speaker-independent word recognition method is based on multiple templates which cover the variance among speakers. However, this system has several defects. One of these is the increase in computation amount in proportion to the number of word templates.

In the SPLIT system, as the spectral distance calculation amount is independent of the number of word templates, the best use can be made in speaker-independent word recognition based on multiple word templates. When applying the SPLIT system to speaker-independent word recognition based on multiple templates, there are several important study items.

(1) How to make the multiple word templates [82], [83].

(2) How to make the phoneme-like templates which are used commonly for all speakers.

Nearly the same algorithm as that used in generating phoneme-like templates is applied for selecting multiple templates. The difference points are in replacing the spectral distance between two vectors by the spectral distance between two words using dynamic time warping and selecting word templates instead of computing average templates. The word templates are selected on the basis of the word utterances distribution. Namely, if W_i has many words in its near neighbourhood, that word W_i is picked up as one of the word templates. For the second problem, phoneme-like templates were generated using utterances by three male speakers and one female speaker.

In the recognition experiments, 8184 words (31 words/speaker) uttered by 264 speakers through telephone lines were used. (Appendix 5) To verify the effect of choosing multiple templates, the following

experiment was executed.

(1) Using 264 utterances, multiple templates were generated for each word by the clustering technique. Average number of word templates for each word was set to 18. Utterances by all the speakers, excepting the speaker whose utterance was used as one of the multiple templates, were used for recognition test.

(2) The recognition experiment without clustering was also carried out. 264 speakers were arbitrarily divided into groups of about 31 speakers. A speaker was selected from 31 speakers and word utterances by that speaker were recognized using utterances from 30 other speakers as multiple templates. Rotating this 31 times, the average recognition accuracy was calculated for each group.

Experimental results are shown in Table 7.2. Results show that multiple templates, which were generated by clustering, work quite well.

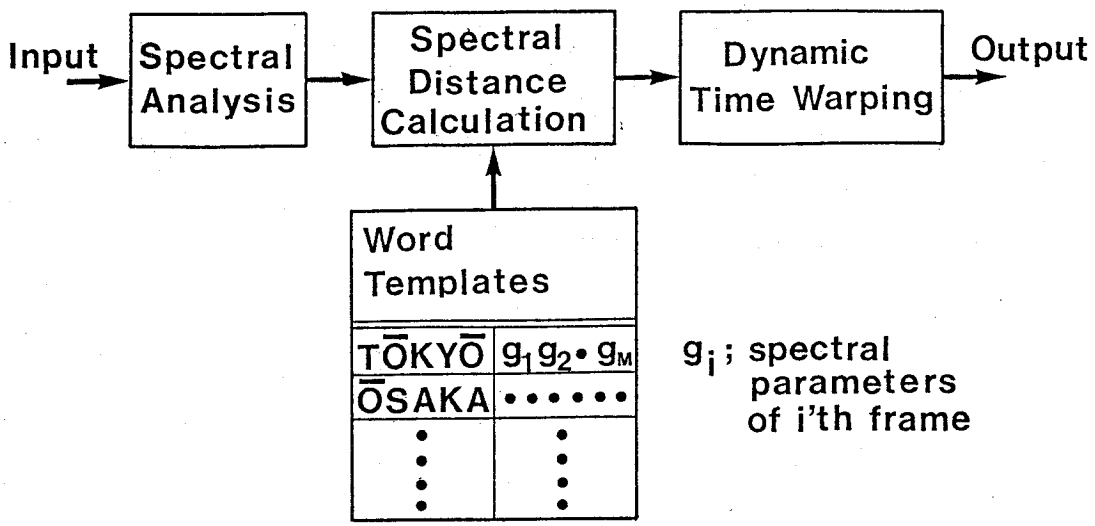
7.6 Conclusions

This chapter proposed a new word recognition system, named SPLIT, and described some feature of this system. The efficiency of this method in speaker-dependent large vocabulary word recognition and speaker-independent isolated word recognition, based on multiple-templates, was clarified through several experiments. These experimental results mean that spectral information in isolated word recognition based on dynamic time warping can be roughly quantized in each frame without a large decrease in the recognition accuracy.

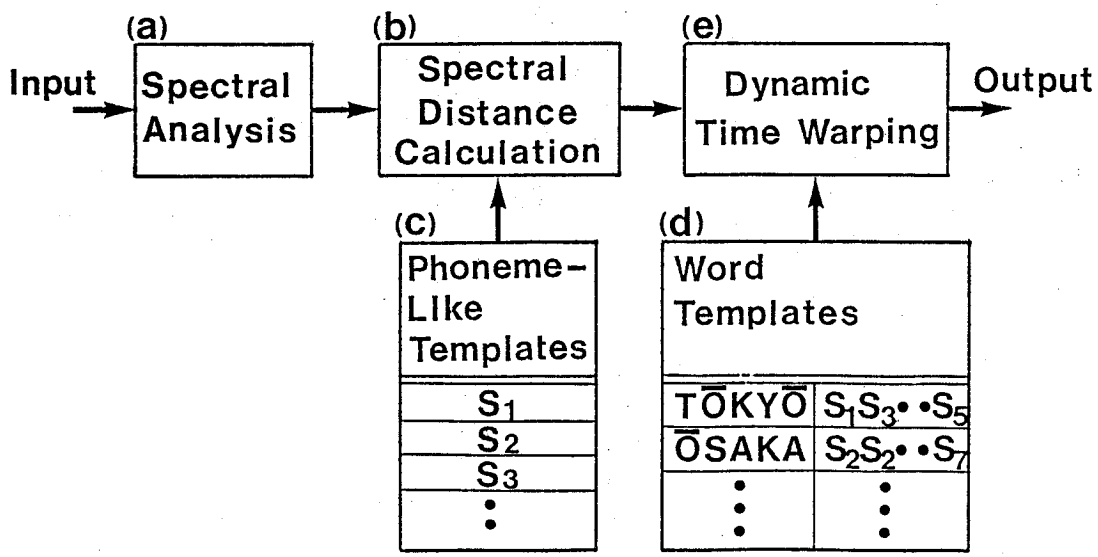
(1) 96.3 % recognition accuracy was obtained in 641 city names word recognition for four male speakers, when there were 256 phoneme-like templates.

(2) 94.5 % recognition accuracy was obtained when there were 32 phoneme-like templates. In this case, the amounts of spectral distance calculation and memory for word templates are 0.3 % and 2 % of the same categories obtained by the direct-matching method, respectively.

(3) A simple clustering technique was proposed to generate word templates in speaker-independent word recognition.



(a) Typical word recognition scheme



(b) Word recognition scheme based on vector quantization

Fig. 7.1 Isolated word recognition system.

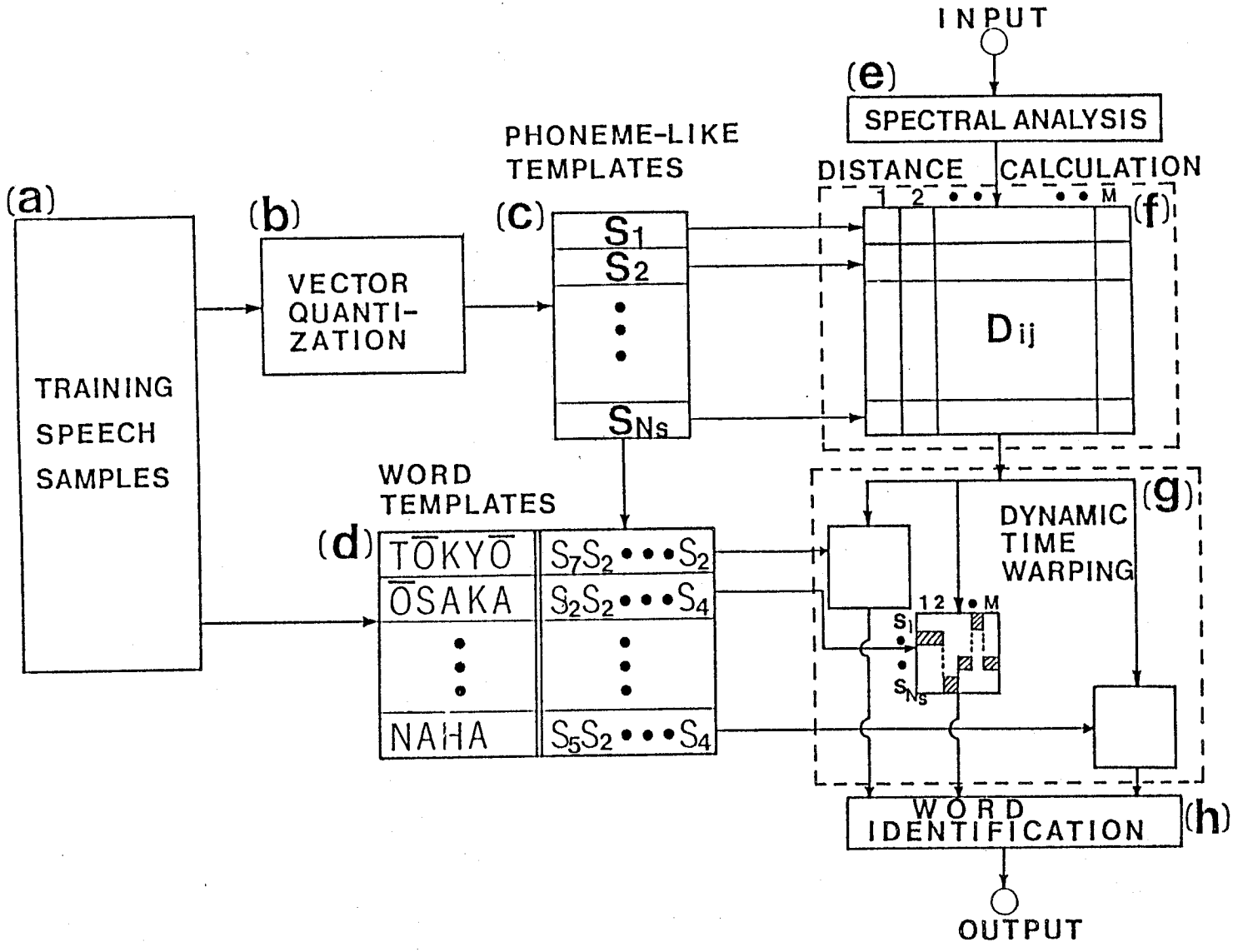


Fig. 7.2 Isolated word recognition system using phoneme-like templates (SPLIT).

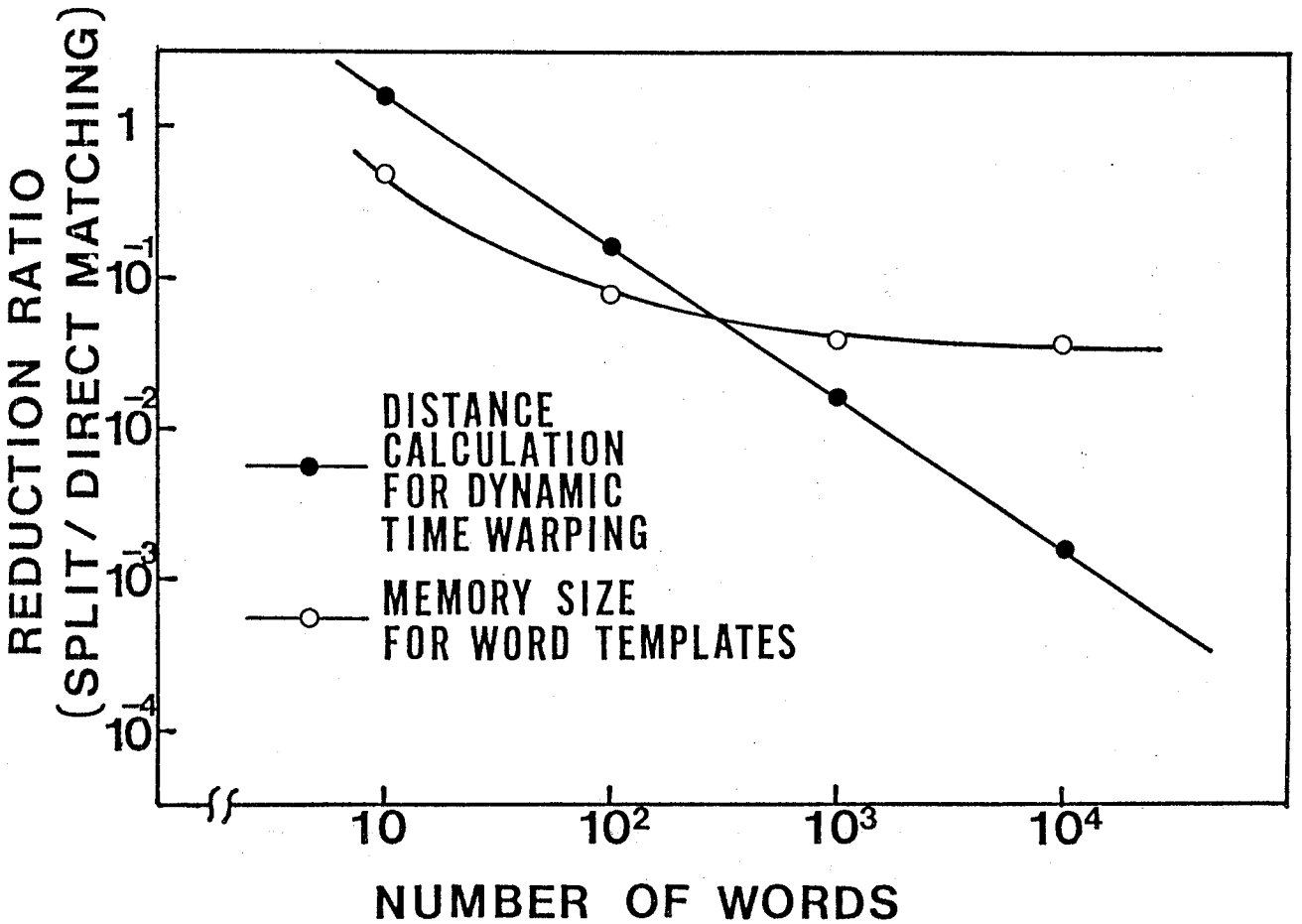


Fig. 7.3 Relationship between the number of words and SPLIT method to direct-matching method reduction ratio, from the viewpoint of memory size and the amount of distance calculation for dynamic time warping.

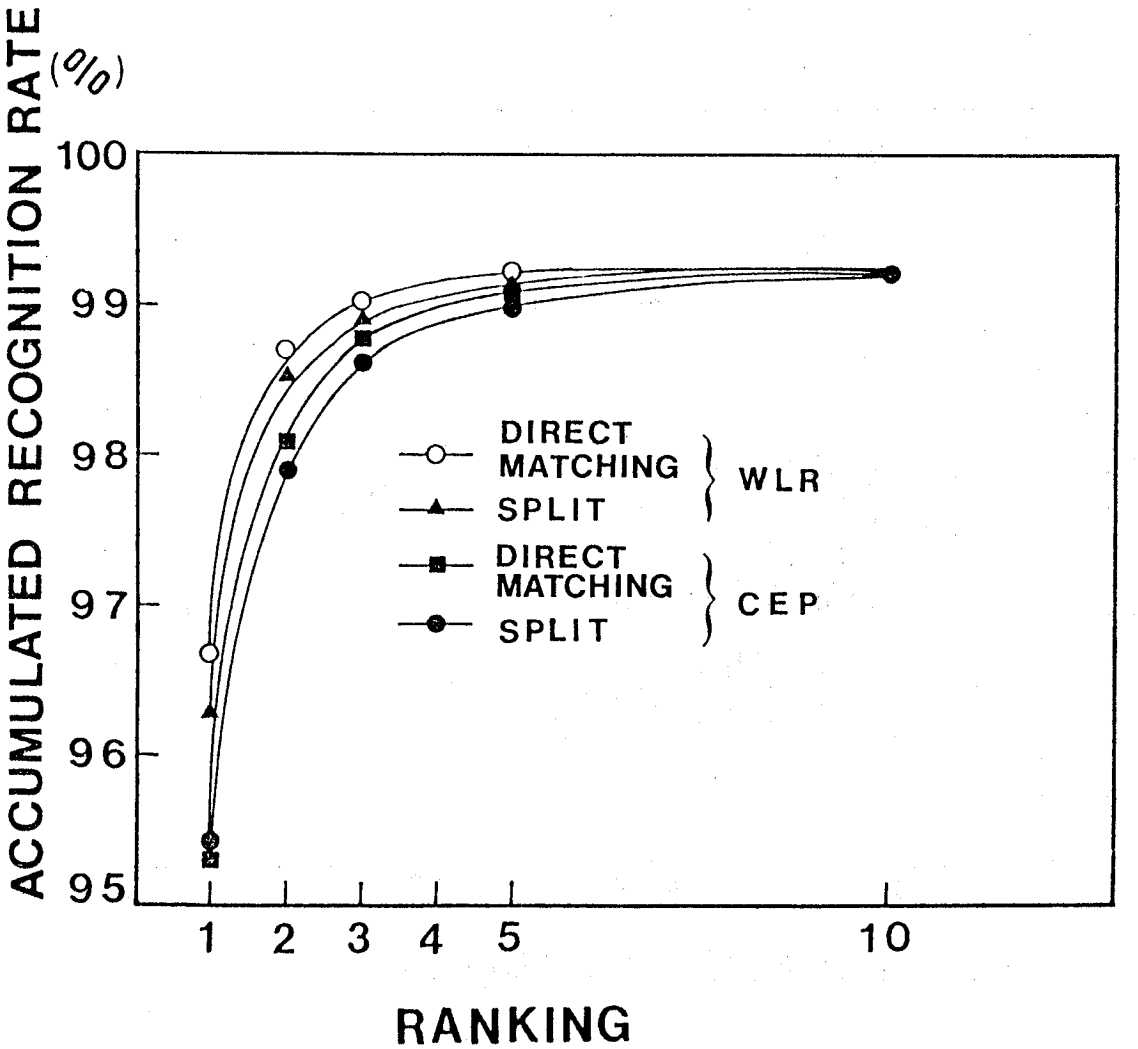


Fig. 7.4 Experimental results in speaker-dependent 641 city names word recognition.

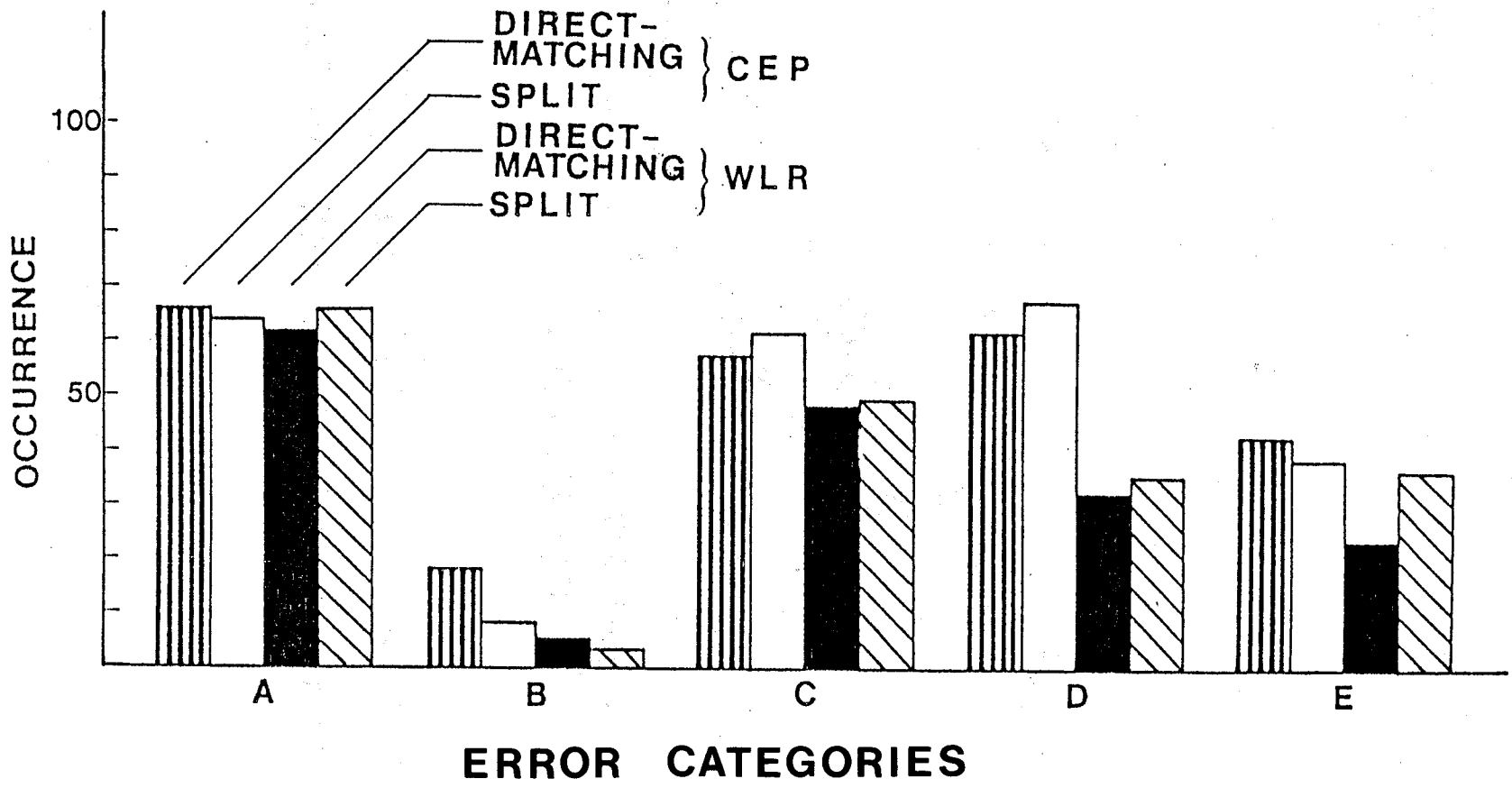


Fig. 7.5 Recognition error classification into five categories.

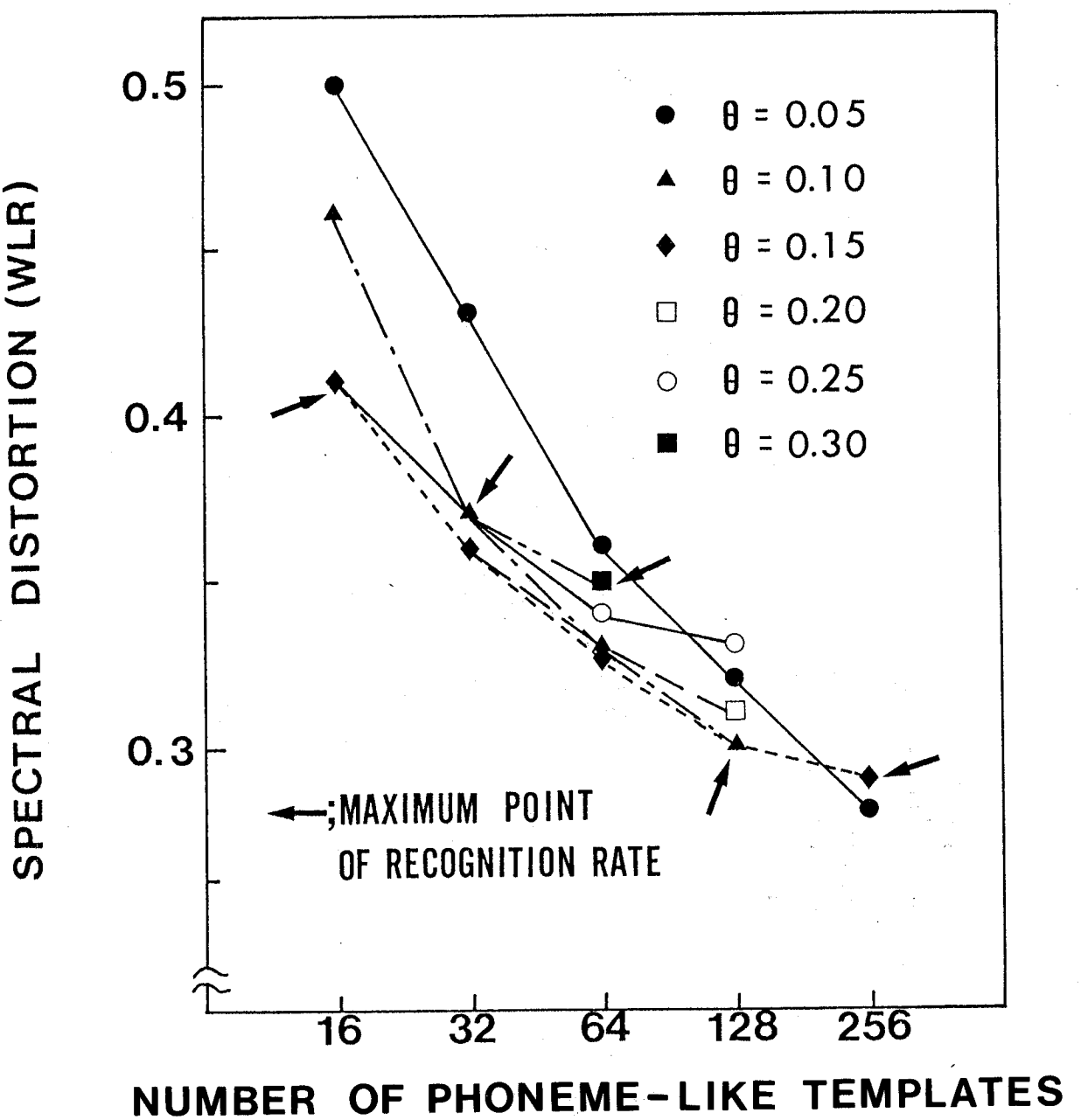


Fig. 7.6 Specral distortion in generating word templates.

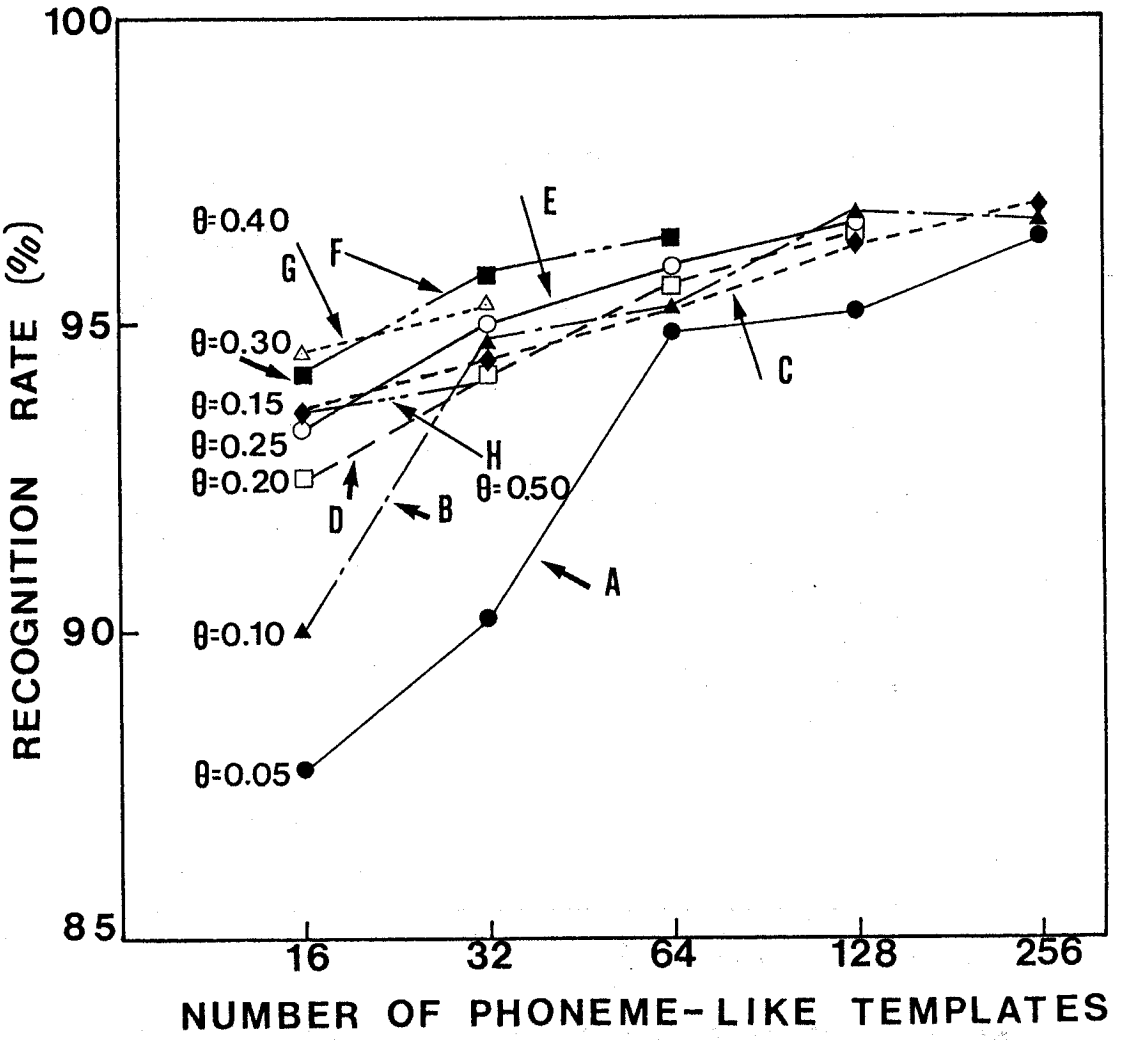


Fig. 7.7 Recognition accuracy versus number of phoneme-like templates.

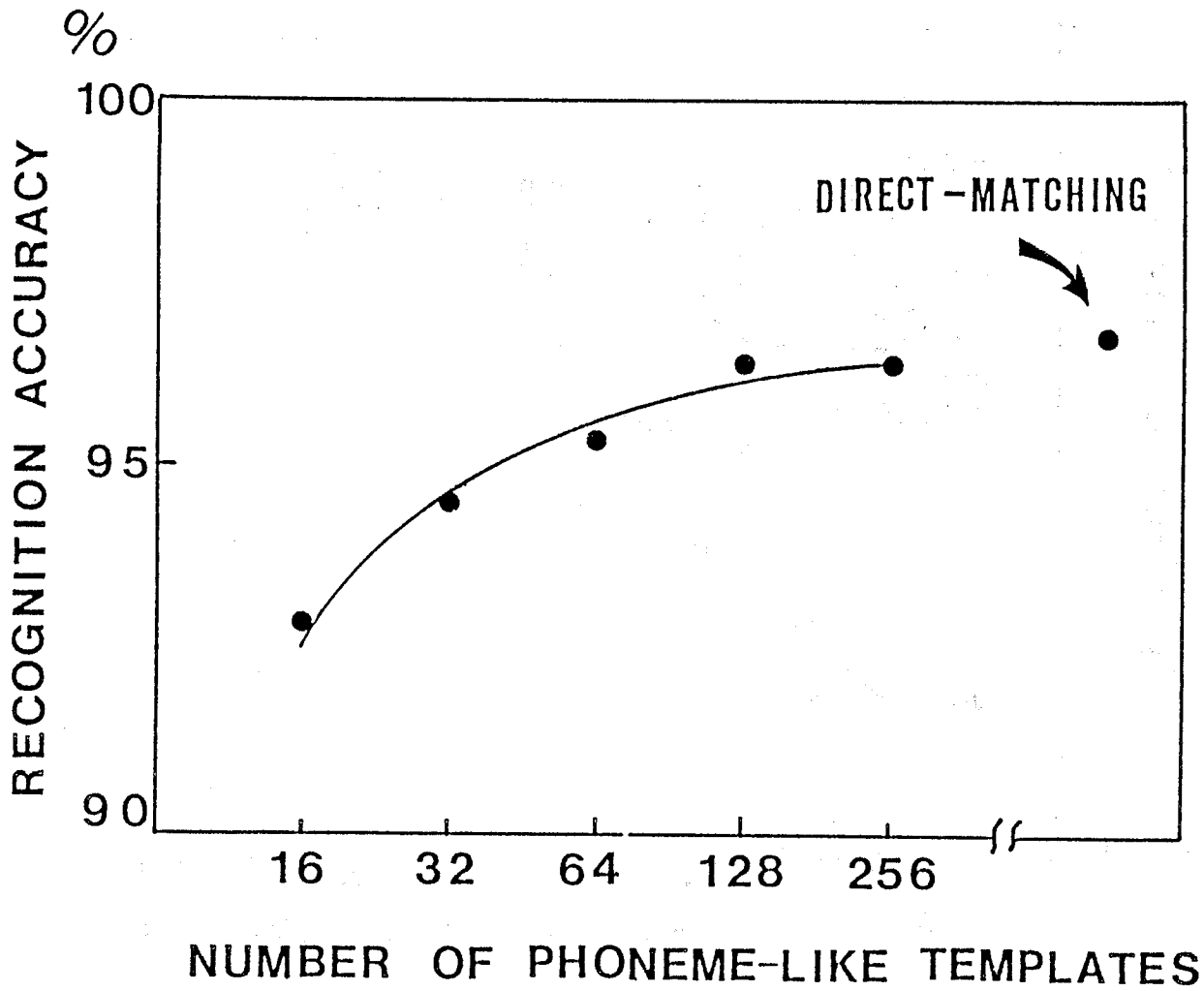


Fig. 7.8 Averaged recognition accuracy for four speakers versus the number of phoneme-like templates.

Table 7.1 Recognition accuracy for each speaker in relation to the number of phoneme-like templates.

N \ SPEAKER / R.R.		NS	KA	KI	KS	AVE.
16	R.R.	92.5	94.5	93.0	91.6	92.9
	θ	0.20	0.25	0.25	0.20	
32	R.R.	94.2	96.4	94.2	9.30	94.5
	θ	0.20	0.20	0.15	0.15	
64	R.R.	95.2	96.3	95.5	94.2	95.3
	θ	0.15	0.15	0.10	0.10	
128	R.R.	96.7	96.1	96.7	95.6	96.3
	θ	0.10	0.10	0.10	0.10	
256	R.R.	96.6	96.6	97.0	95.2	96.4
	θ	0.10	0.10	0.10	0.10	
DIRECT MATCHING		97.0	97.0	97.7	96.4	97.0

Table 7.2 Recognition accuracy in speaker-independent word recognition using multiple word templates.

Condition	With Clustering (18 templates/word)	Without Clustering (30 templates/word)
Recognition Accuracy	98.0 %	97.2 %

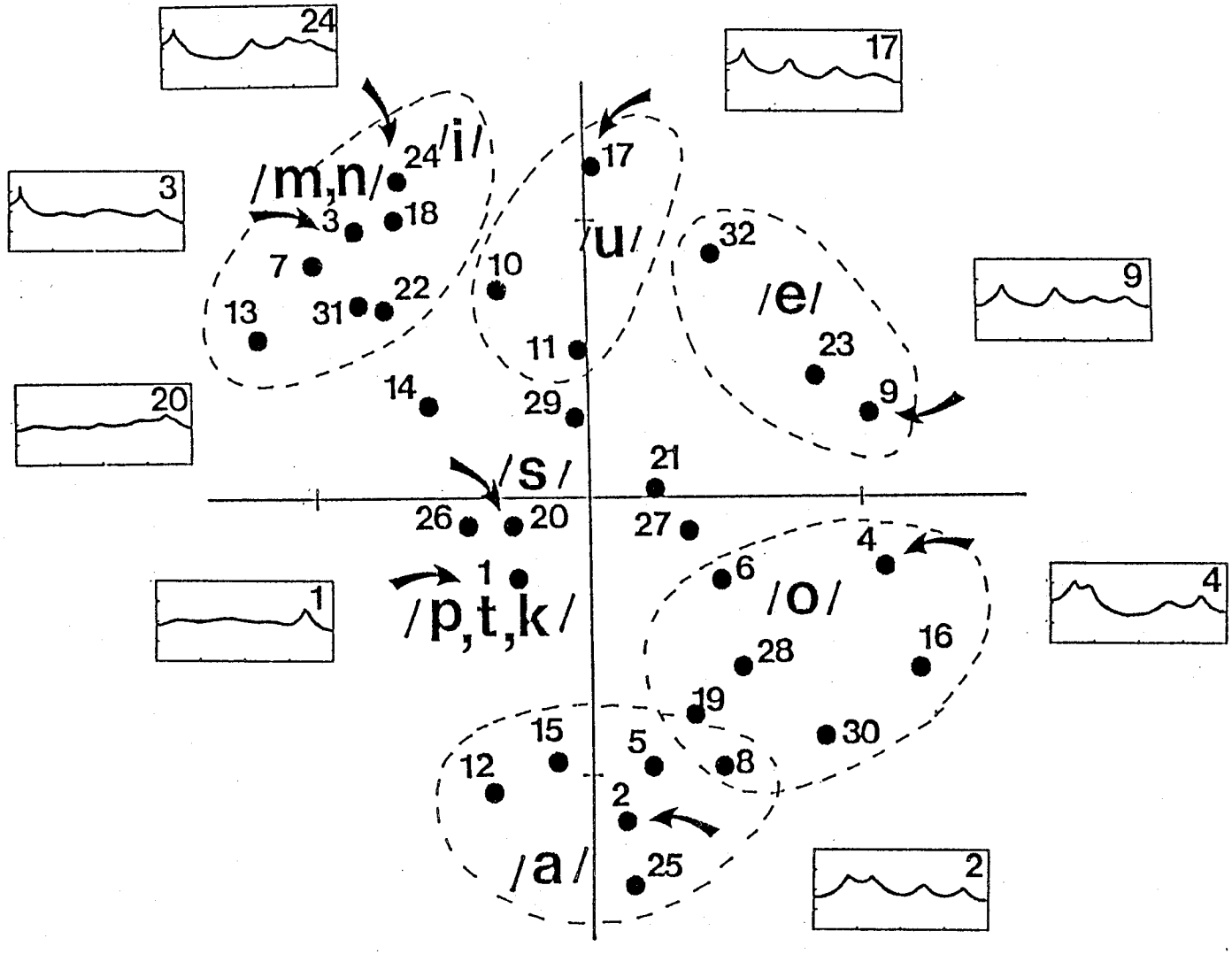


Fig. 7.9 32 phoneme-like templates mapped onto two dimensional plane by multiple dimensional analysis.

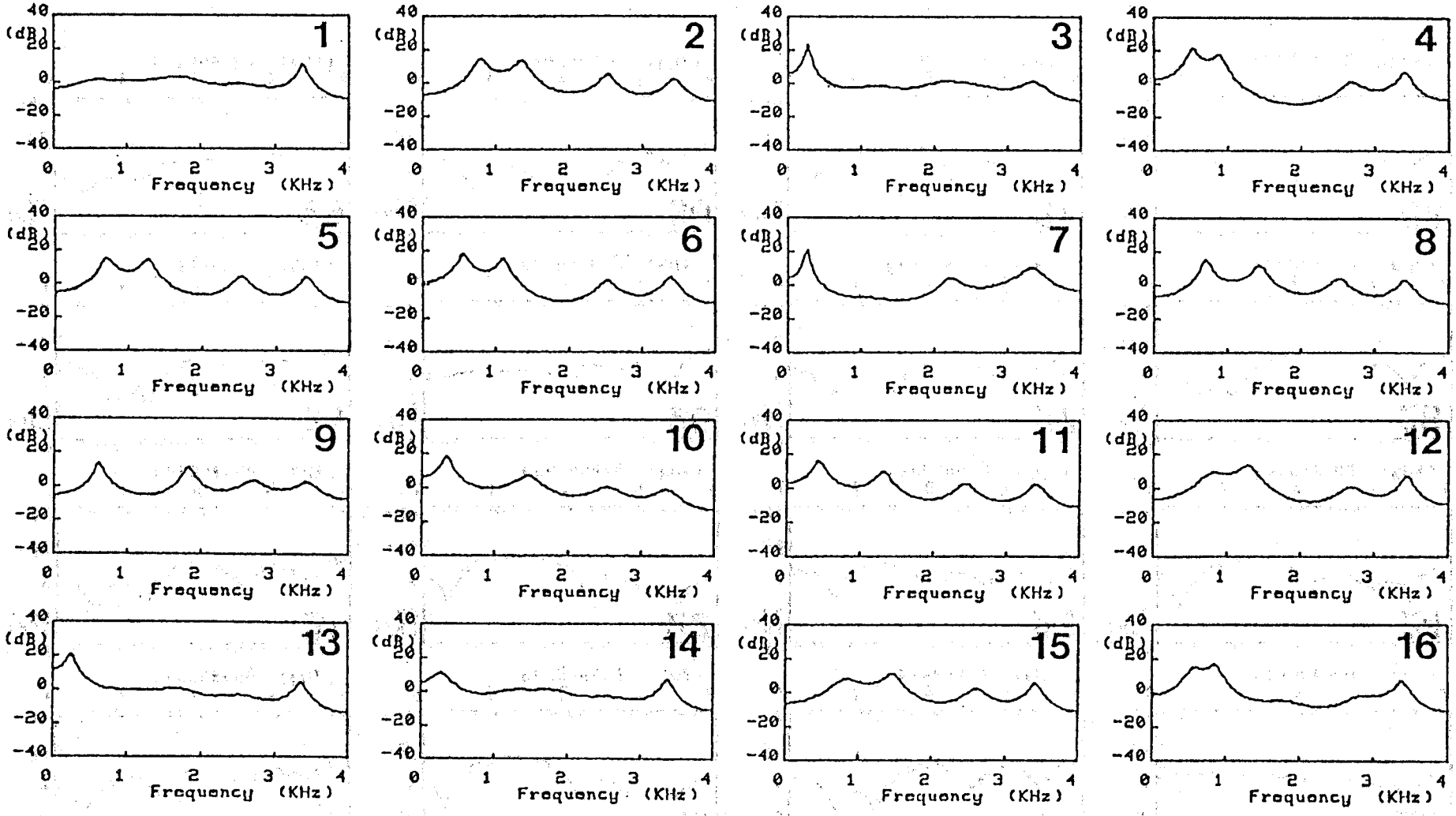


Fig. 7.10 Spectral envelopes for 32 phoneme-like templates.

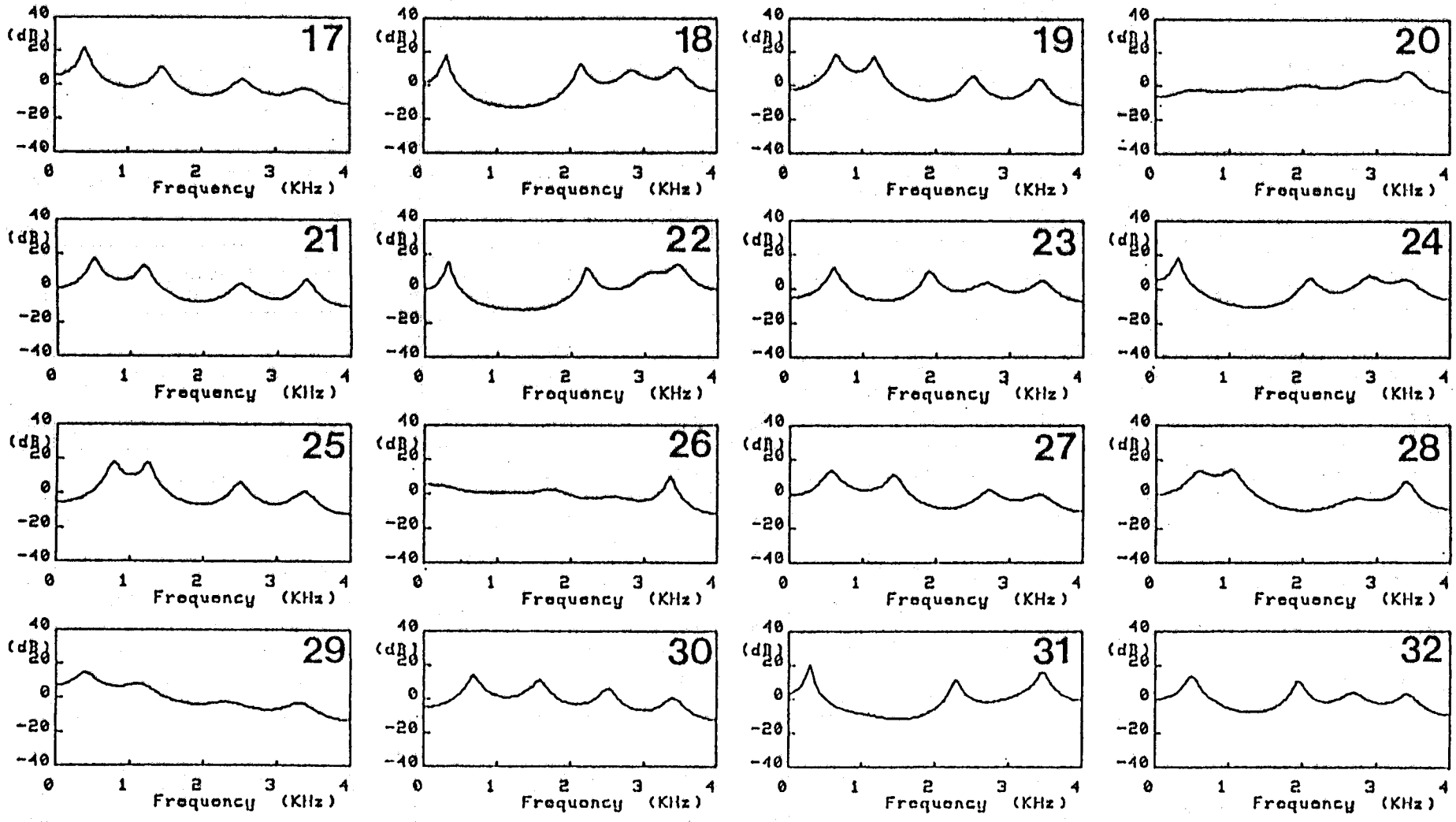


Fig. 7.10 (continuation)

CHAPTER 8 EPILOGUE

8.1 Conclusions of This Thesis

This research has been executed in order to represent a speech signal efficiently at an extremely low bit rate. The conclusions are summed up as follows.

(1) LSR parameters properties were examined as a function of spectral distortion. LSR parameters are one of the representations which are equivalent to LPC in an all pole digital filter model in speech generation. It was clarified as the result of several experiments that LSR parameters have better interpolation properties compared to PARCOR coefficients. An optimum bit allocation to LSR parameters was proposed by spectral sensitivity analysis. Spectral sensitivities for LSR parameters were proved to be rather more uniform than those for PARCOR coefficients.

(2) Another representation of LPC, line spectrum pair (LSP), was presented. Their physical meanings were explained analytically. The quantization and interpolation properties of LSP parameters were examined experimentally. The optimum quantization method was proposed through spectral distortion analysis, in the same manner as for LSR parameters. The relationship between information bit rates and spectral distortion by quantization was discussed, comparing PARCOR coefficients and LSR parameters. The synthesized speech quality was tested by objective and subjective measurements. Consequently, the PARCOR system 60 % information rate is sufficient to obtain equivalent speech quality by LSP parameters. The LSP speech analysis synthesis method is the one of the most efficient vocoder methods used nowadays.

(3) The parameter quantization technique with variable frame rate (VFR) was proposed using the speech signal non-uniformity in the time domain. Optimum frame selection, based on dynamic programming, was executed as a function of time spectral distortion. This method gave a minimum distortion when time intervals and the number of divisions were given. A straight forward variable frame rate scheme was also proposed and verified by subjectively. Synthesized speech quality at 1500 bps with VFR was proved to be equivalent to 14 S/N speech quality. However, other factors, including excitation signals, must be considered to obtain higher quality speech.

(4) Vector quantization, using LSP parameters, was presented to reduce the redundancy in parameter scalar quantization. Vector quantization has an advantage of 10 bits saving per frame over the conventional scalar quantization. Synthesized speech quality at 800 bps is proved to be equivalent to 2000 bps by scalar quantization.

(5) The vector quantization technique was applied to an isolated word recognition based on dynamic time warping. Word templates were represented as discrete spectral patterns sequences. According to this processing, dramatic reduction can be achieved in spectral calculations and memory requirements without degrading the performance, compared to a conventional pattern matching method, based on dynamic time warping. The relationship between the number of phoneme-like templates and recognition accuracy was clarified. It was proved that word templates can be quantized with 7 or 8 bits in each frame without degrading recognition accuracy

The relationship between the information rate and synthesized speech quality was clarified as a function of spectral distortion. During the process of this research, a speech synthesized technique,

based on LPC, was implemented on a single LSI chip and speech signal processing techniques have been paid much more attention. Some application systems, based on this research, are going to be realized in several fields.

To represent speech signals at an extremely low bit rate is a very interesting problem. It is related with speech property itself and it is closely connected with speech synthesis and recognition or other speech processings. The most suitable method exists in each information rate.

Several techniques proposed in this thesis are very popular nowadays. The author hopes that the basic idea and approaches in this thesis will be the foundation of speech processing research efforts in the near future.

REFERENCES

(1) J.L.Flanagan : "Speech Analysis Synthesis and Perception", Second Edition, 1972, Spriger-Verlag.

(2) J.D.Markel, A.H.Gray, Jr. : "Linear Prediction of Speech", 1976, Springer-Verlag.

(3) Y.Katsuki, A.H.Gray, Jr. : "Hearing and Speech", IECE Japan.

(4) F.deJager : "Delta Modulation, A Method of PCM Transmission Using the 1-Unit Code", Philips Res. Rept, pp.442-466, 1952.

(5) B.S.Atal, M.R.Schroeder : "Predictive Coding of Speech Signals", The 6-th International Congress on Acoustics, C-5-4, 1968.

(6) N.S.Jayant : "Adaptive Delta Modulation with A One-bit Memory", Bell Syst. Tech. J., Mar.,1970.

(7) P.Cumiskey, N.S.Jayant, J.L.Flanagan : "Adaptive Quantization in Differential PCM Coding of Speech", Bell Syst. Tech. J., Sept.,1973.

(8) R.E.Crochiere : "On the Design of Subband Coders for Low-Bit-Rate Speech Communication", Bell Syst. Tech. J., May-June, 1977.

(9) R.Zelinski, P.Noll : "Adaptive Transform Coding of Speech Signals", IEEE Trans. on ASSP, Vol.25, No.4, Aug.,1977.

(10) M.Honda and F.Itakura : "A Study of a Predictive Coding Scheme With Adaptive Bit Allocation", Proceeding of ASJ, June 1979.

(11) H.Dudley : "The Vocoder", Bell Labs Rec., 18, p.122, 1939.

(12) W.A.Munson, H.C.Montgomery : "A Speech Analyzer and Synthesizer", JASA, Vol.22, p.678(A), 1950.

(13) C.P.Smith : "Voice-Communication Method Using Pattern Matching for Data Compression", JASA., Vol.35, p.805(A), 1963.

(14) M.R.Schroeder : "Correlation Techniques for Speech Bandwidth Compression", J. Audio Eng. Soc., Vol.10, p.163,1962.

(15) M.R.Schroeder, E.F.David, Jr. : "A Vocoder for Transmitting 10 kc/s Speech Over a 3.5 kc/s Channel", Acoustica, Vol.10, p.35, 1960.

(16) J.L.Flanagan, R.M.Golden : "Phase Vocoder", Bell system Tech. J., Vol.45, p.1943, 1966.

(17) F.Itakura S.Saito : "An Analysis-Synthesis Telephoney Based On Maximum Likelihood Method", Proc. Int. Congr. Acoust. C-5-5, 1968.

(18) A.V.Oppenheim : "Speech Analysis-Synthesis System Based On Homomorphic Filtering" : "JASA, Vol.45, p.459, 1969.

(19) F.Itakura and S.Saito : "Speech Analysis-Synthesis Based On the Partial Autocorrelation Coefficient", Proceeding of JSA, pp.199-200, 1969.

(20) B.S.Atal, S.L.Hanauer : "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", JASA Vol.50, p.637, 1971.

(21) D.T.Magill, C.K.Un : "Residual excited Linear Predictive Vocoder", 87th Meeting of ASA, 1974.

(22) M.R.Schroeder : "Vocoders : Analysis and Synthesis of Speech", Proc. IEEE Vol.54, p.720, 1966.

(23) H.Dudley : "Phonetic Pattern Recognition Vocoder for Narrow-band Speech Transmission", JASA, Vol.30, pp.733-742, Aug.,1958.

(24) G.S.Kang, D.C.Coulter : "600 BPS Voice Digitizer", Proc. of IEEE Int. Conf. on ASSP, pp.91-94, 1976.

(25) A.W.F.Huggins, R.Viswanathan, J.D.Markel. : "Speech Quality Testing of Some Variable-frame-rate (VFR) Linear-predictive (LPC) Vocoders", JASA, Vol.12, No.2, pp.430-434, Aug., 1977.

(26) R.Reddy : "Use of Segmentation and Labelling in Analysis and Synthesis Speech ", Proc. of IEEE Int. Conf. on ASSP, pp.28-31, 1977.

(27) C.K.Un : "A Low-rate Digital Formant Vocoder", IEEE trans. on Comm., Vol.26, No.3, pp.344-355, Mar., 1978.

(28) R.Viswanathan, J.D.Markel, R.Wicke : "The Application of a Fundamental Model of Speech to Variable LPC Systems", Proc. of IEEE Int. Conf. on ASSP, pp.219-222, 1977.

(29) A.Buzo, R.M.Gray, J.D.Markel, A.H.Gray, Jr. : "A two Step Speech Compression System With Vector Quantization", Proc. of Int. Conf. of IEEE on ASSP, pp.52-55, 1979.

(30) D.Y.Wong, B.H.Juang, A.H.Gray, Jr. "Recent Developments in Vector Quantization for Speech Processing", Proc. of Int. Conf. of IEEE on ASSP, pp.1-4, 1981.

(31) S.Roucus : "A Segment Vocoder at 150 B/S", Proc. of Int. Conf. on ASSP, pp.61-64, 1983.

(32) D.Y.Wong, : "Very Low Data Rate Speech Compression With LPC Vocoder and Matrix Quantization", Proc. of Int. Conf. of IEEE on ASSP, pp.65-68, 1983.

(33) S.Saito and K.Nakata : "Fundamentals of Speech Signal Processing", Acoustic Society of Japan, 1981, Ohom Printing Inc.

(34) N.Kitawaki, K.Itoh, F.Itakura : "PARCOR Speech analysis-Synthesis System", Review of the E.C.L. of N.T.T., Japan, Vol. 26, Nos.11-12, Nov.-Dec., 1978.

(35) N.Kitawaki, F.Itakura, S.Saito : "Optimum Coding of Transmission Parameters in PARCOR Analysis Synthesis System", Iece Trans., J61-a, Vol.2, pp.119-128, Feb., 1978.

(36) F.Itakura : "Optimal Nonlinear transformation of LPCs to Improve Quantization Properties", JASA, Vol.56, S16, 1974.

(37) A.H.Gray,Jr.,R.M.Gray, J.D.Markel : "Comparison of Optimal Quantizations of Speech Reflection Coefficients", IEEE Trans. on ASSP, Vol.25, P.9, Feb., 1977.

(38) N.Kitawaki, F.Itakura : "Efficient Coding of Speech by Nonlinear

Quantization and Nonuniformity of PARCOR coefficients", IECE Trans., J61-A, No.6, pp.543-550, June 1978.

(39) Y.Tohkura : "Improvement of Voice Quality in PARCOR Bandwidth Compression System", IECE Trans., Vol.61-A, No.3, Mar. 1978.

(40) Y.Tohkura, F.Itakura, S.Hashimoto : "Spectral Smoothing Technique in PARCOR Speech Analysis-synthesis", IEEE Trans. on ASSP, Vol.26, Dec. 1978.

(41) F.Itakura : "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals", JASA, Vol.57, S35(A), p.s35, 1975.

(42) F.Itakura : "Line Spectrum Representation of Linear Prediction Coefficients of Speech Signal", Trans. of the Committee on Speech Research, ASJ, S75-34, Dec. 1975.

(43) F.Itakura, S.Saito : "Digital Filtering Techniques for Speech Analysis and Synthesis", 7th Int. Cong. Acoust. Budapest, 25C1, 1971.

(44) S.Sagayama : "Symmetry Between LPC and CSM Formulated by Theory of Orthogonal Polynomials", Trans. of the Committee on Speech Research, ASJ, S81-02, Apr. 1981.

(45) S.Sagayama, F.Itakura : "Composite Sinusoidal Modeling Applied to Spectral Analysis of Speech", IECE Trans., Vol.J64-A, No.2, pp.105-112, Feb. 1981.

(46) H.Fujisaki : "Current Problems in Speech Recognition", JASA, Vol.28, No.1, pp.33-41, 1972.

(47) N.Sugamura, F.Itakura : "Line Spectrum Representation of Linear Predictor Coefficients of Speech Signal and Its Statistical Properties", IECE Trans. Vol.J64-A, No.4, pp.323-340, Apr. 1981.

(48) N.Kitawaki, K.Itoh, K.Kakehi : "Speech Quality Measurement Methods for Synthesized Speech", Review of E.C.L.,N.T.T., Japan, Vol.29, Nos.9-11, Sep.-Dec. 1981.

(49) Y.Tohkura, F.Itakura : "Spectral Sensitivity of PARCOR and Difference Limen of Spectral Distortion", Proc. of IECE meeting in Spring, 1231, 1977.

(50) Y.Tohkura, F.Itakura : "Effect of Spectral Smoothing in PARCOR System", Tech. Group on Electrical Acoustics, Inst. Electronics Commun. Eng. Japan, EA76-22, July 1976.

(51) J.L.Flanagan : "A Difference Limen for formant Frequency", JASA, Vol.27, No.3, pp.613-617, May 1955.

(52) J.L.Flanagan : "Difference Limen for Formant Amplitude", JSHD, Vol.22, No.2, pp.205-212, Jan. 1957.

(53) N.Sugamura, F.Itakura : "Speech Data Compression by LSP Speech Analysis and Synthesis Technique", IECE Trans., Vol.J64-A, No.8, pp.599-605, Aug. 1981.

(54) F.Itakura, N.Sugamura : "LSP Speech Synthesizer , Its Principle and Implementation", Trans. of the Committee on Speech Research, ASJ, S79-46, Nov. 1979.

(55) S.Sagayama : "Stability Condition of LSP Speech Synthesis Digital Filter", Proc. of ASJ, pp.153-154, Mar. 1982.

(56) N.Sugamura, S.Sagayama, F.Itakura : "A study of Speech Quality of Synthesized Speech by LSP", Trans. of the Committee on Speech Research, ASJ, H80-29, July 1980.

(57) E.Blackman, R.Viswanathan, J.Makhoul : "Variable-to-fixed Rate Conversion of Narrowband LPC Speech", Proc. of Int. Conf. of IEEE on ASSP, pp.409-412, 1977.

(58) R.E.Crochiere, M.R.Sambur : "A Variable Band Coding Scheme for Speech Encoding At 4.8 KB/s", Proc. of Int. Conf. of IEEE on ASSP, pp.444-447 1977.

(59) W.J.Hess, J.Heiler : "Towards a Variable Frame Rate Speech

Transmission System With frame Selection by Time-domain Segmentation", Proc. of Int. Conf. of IEEE on ASSP, 1978.

(60) H.R.Silbiger : "IEEE Recommended Practice for Speech Quality Measurements", IEEE No.297, June 1969.

(61) C.P.Smith :,"Perception of Vocoder Speech Processed by Pattern Matching", JASA, Vol.46, No.6, pp.1562-1571, July 1969.

(62) A.Buzo, R.M.Gray, A.H.Gray, Jr., J.D.Markel : "Speech Coding Based Upon Vector quantization", IEEE trans. on ASSP, Vol.28, No.5, pp.562-574, Oct. 1980.

(63) B.H.Juang, D.Y.Wong, A.H.Gray, Jr. : "Distortion Performance of Vector Quantization for LPC Voice Coding", IEEE Trans. on ASSP, Vol.30, No.2, pp.294-304, Apr. 1982.

(64) D.Y.Wong, B.Juang, A.H.Gray, Jr. : "An 800 bit/s Vector Quantization LPC Vocoder", IEEE Trans. on ASSP, Vol.30, No.5, Oct. 1982.

(65) Y.Tohkura : "Experimental Comparison of Parameter Interpolation in LPC Vocoders", ASA*50 Speech Comm. Preprint Experiment, p.373, 1979.

(66) Y.Tohkura, F.Itakura : "Spectral Sensitivity Analysis of PARCOR Parameters for Speech Data Compression", IEEE Trans. on ASSP, Vol.27, No.3, P.273, 1979.

(67) N.Sugamura, F.Itakura : "Speech Data Compression by Spectral Pattern Matching", IECE Trans., Vol.J65-A, No.8, pp.834-841, Aug. 1982.

(68) N.Sugamura, F.Itakura : "Speech Data Compression by Spectral Pattern Matching In the Case of A Single Speaker", Trans. of the Committee on Speech Research, ASJ, S80-45, Oct. 1980.

(69) J.L.Flanagan : "Computers that Talk and Listen; Man-Machine Communication by Voice.", Proc. of IEEE, Vol.64, No.4, 1976.

(70) D.R.Reddy : "Speech recognition by Machine; A Review", Proc. of IEEE, Vol.64, No.4, 1976.

(71) T.B.Martin : "Practical Applications of Voice Input to Machines", Proc. of IEEE, Vol.64, No.4, pp.487-500, 1976.

(72) Y.Niimi : "Speech Recognition", 1979, Kyoritsu Printing Inc.

(73) M.Kohda, S.Saito : "Speech Recognition by Incomplete Learning samples", IEEE Conf. on Speech Commun., Processing, Rep. H-10, 1972.

(74) S.Furui : "A Training Procedure for Isolated Word recognition System", IEEE Trans. on ASSP, Vol.28, No.2, pp.129-136, Apr. 1980.

(75) N.Sugamura, K.Hakoda, S.Furui. : "Large Vocabulary Word Recognition by Pseudo-Phoneme Templates", Trans. of IECE, p.5-332, 1981.

(76) N.Sugamura, S.Furui : "A Large Vocabulary Word Recognition System Using Pseudo-Phoneme Templates", IECE Trans., Vol.J65A-D, No.8, pp.1041-1048, Aug. 1982.

(77) Y.Linde, A.Buzo, R.M.Gray : "An Algorithm for Vector Quantizer Design", IEEE Trans. on Com., Vol.28, pp.84-95, Jan. 1980.

(78) M.Sugiyama, K.Shikano : "Centroid for LPC Spectral Measure", Trans. of the Committee on Speech Research, ASJ, S82-12, May 1982.

(79) M.Sugiyama, K.Shikano : "LPC Peak Weighted Spectral Matching Measure", IECE Trans., Vol.J64-a, No.5, pp.409-416, May 1981.

(80) K.Shikano, M.Sugiyama : "Evaluation of LPC Spectral Matching Measure for Spoken Word Recognition", IECE Trans., Vol.J65-D, No.5, pp.535-541, May 1982.

(81) H.Sakoe, S.Chiba : "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. on ASSP, Vol.26, pp.43-49, Feb. 1978.

(82) S.E.Levinson, L.R.rabiner, A.E.Rosenberg, J.G.Wilpon : "Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates for Isolated Word Recognition", IEEE Trans. on ASSP, Vol.27, No.2, Apr. 1979.

(83) L.R.Rabiner, S.E.Levinson, A.E.Rosenberg, J.G.Wilpon : "Speaker Independent Recognition of Isolated Words Using Clustering Techniques", IEEE Trans. on ASSP, Vol.27, No.4, Aug. 1977.

IECE : Institute of Electrical and Communication Engineers of Japan.

ASJ : Acoustical Society of Japan

APPENDIX 1 Proof of theorem.

The theorem is proved by introducing the following rational function of z .

$$R_p(z) = \frac{z^{p+1} A_p(1/z)}{A_p(z)} = z \prod_{i=1}^{p} \frac{1 - \overline{z_i}\, z}{z - z_i} \qquad (A1.1)$$

$$|z_i| > 1$$

where z_i denote the zeros of $A_p(z)$.

When z moves round on the unit circle, namely $z=e^{j\omega}$, $-\pi<\omega<\pi$, $R_p(z)$ circulates (p+1) times on the unit circle. Denoting values of ω at which $R_p(e^{-j\omega})=1$ is satisfied by $\omega_1, \omega_2, \omega_3$ --------, and denoting values of θ at which $R_p(e^{-j\omega})=-1$ is satisfied by $\theta_1, \theta_2, \theta_3$ -------, the the following inequality holds

$$0=\omega_0<\theta_1<\omega_1< \text{----------} <\theta_p<\omega_p<\theta_{p+1}=\pi \qquad (A1.2)$$

Since $R_p(z) = \pm 1$ leads to P (z)=0 or Q (z)=0, respectively, the theorem is proved.

APPENDIX 2

Opinion score for each speech sample added multiplicative random noise

SPEAKER	TS(F)	EN(F)	SM(M)	SM(M)	TS(F)	EN(F)	EN(F)	SM(M)	YG(M)	AVERAGE
CONTENTS	A	B	C	A	B	C	A	B	C	
S / N	T1			T2			T3			
0 dB	0.139	0.222	0.278	0.194	0.278	0.306	0.111	0.056	0.139	0.191
3	0.222	0.250	0.250	0.667	0.417	0.417	0.167	0.222	0.111	0.303
7	0.528	0.694	0.611	0.750	0.806	0.917	0.583	0.333	0.444	0.630
10	0.583	1.028	0.917	1.056	1.361	1.278	0.861	0.694	0.917	0.966
13	0.556	1.083	1.278	1.444	1.417	1.667	0.889	0.611	0.944	1.10
15	0.889	1.583	1.583	1.583	1.444	1.861	1.583	0.889	1.083	1.39
17	1.389	1.639	1.583	1.528	1.444	2.167	1.583	1.361	1.306	1.56
20	1.389	1.972	1.806	1.972	2.056	2.417	1.889	1.611	1.500	1.85
23	1.750	2.389	2.250	2.278	2.056	2.611	2.194	1.722	2.306	2.17
27	1.694	2.972	2.917	2.528	2.694	3.028	2.694	2.500	2.028	2.56
30	2.389	2.889	3.111	2.889	2.861	2.972	3.278	2.889	3.306	2.95
40	3.194	3.111	3.389	3.722	3.389	3.528	3.500	3.694	3.472	3.44
∞	3.222	3.139	3.444	3.556	3.639	3.611	3.667	3.611	3.667	3.51

APPENDIX 2

Opinion score for each synthesized speech sample

SAMPLING METHOD	SPEAKER		TS(F)	EN(F)	SM(M)	SM(M)	TS(F)	EN(F)	EN(F)	SM(M)	YG(M)	AVERAGE
	CONTENTS		A	B	C	A	B	C	A	B	C	
	BPS	FRAME LENGTH	T1			T2			T3			
UFR	1800	30	1.222	1.278	1.028	1.139	0.806	1.667	1.472	0.972	1.389	1.219
	1500	30	1.056	0.833	1.083	0.889	0.750	1.750	1.583	1.028	1.278	1.139
	1500	40	0.806	0.222	1.083	0.361	0.722	0.944	1.528	0.722	0.944	0.815
	1500	50	0.444	0.306	0.778	0.361	0.194	0.444	0.361	0.639	0.528	0.451
	1200	30	0.611	0.833	0.972	0.444	0.500	1.111	0.889	0.639	1.028	0.781
	1200	40	0.611	0.556	1.000	0.472	0.583	0.833	0.972	0.667	0.750	0.716
	1200	50	0.278	0.417	0.528	0.306	0.194	0.556	0.222	0.333	0.444	0.364
VFR	1500	30	1.000	1.222	0.778	1.444	1.472	1.556	1.583	1.028	1.361	1.272
	1500	40	1.111	0.861	0.417	0.500	1.139	1.472	0.917	1.083	1.139	0.960
	1500	50	0.889	0.556	0.111	0.500	0.778	1.444	0.750	0.556	1.167	0.750
	1200	30	0.750	0.750	1.167	0.583	0.778	0.861	1.250	0.778	1.194	0.901
	1200	40	0.806	0.556	0.361	0.361	1.028	1.556	1.389	0.500	0.639	0.800
	1200	50	0.861	0.444	0.111	0.417	0.389	0.917	0.778	0.306	0.500	0.525

APPENDIX 3 Dynamic time warping algorithm [80].

Dynamic time warping [81] is modified into an unconstrained endpoint type. In this modification, only two direction paths are permitted at the starting point, considering the cost function. Assume that input word is B, word template is A and D(i,j) is defined as the spectral distance between the i-th frame of B and j-th frame of A. G(i,j) denotes the cost function at (i,j) point. The dynamic time warping algorithm is shown in Fig. 1A

Initial conditions and iterations are given as follows.

(1) Initialization: $G(1,1) = 2\,D(1,1)$

$$\begin{aligned} G(1,j) &= G(1,j-1)+D(1,j): \quad 2 \leqq j \leqq r/2 \\ G(i,1) &= G(i-1,1)+D(i,1): \quad 2 \leqq i \leqq r/2 \\ G(1,j) &= \infty : r/2 \leqq j \leqq r \\ G(i.1) &= \infty : r/2 \leqq i \leqq r \end{aligned} \tag{A2.1}$$

where r/2 is a permissible width for starting.

(2) Iterations

$$\begin{aligned} G(i,j) = \min[&G(i-1,j-2)+2\,D(i,j-1), \\ &G(i-1,j-1)+D(i,j), \\ &G(i-2,j-1)+2\,D(i-1,j)] \\ &+D(i,j) \end{aligned} \tag{A2.2}$$

$$2 \leqq i \leqq iend,\ 2 \leqq j \leqq jend$$

where iend is the frame number of word template and jend is the frame

number of input word.

(3) Decision: Final matching score S(A:B) is defined as follows.

$$S(A:B) = \min_{(i,j)} (G(i,j)/(i+j)) \qquad (A2.3)$$

$$(i,j) \in \text{end region}$$

In this dynamic time warping, (iend-r/2～iend,jend) and (iend,jend-r/2～jend) are free regions. Then, even if (iend,jend) is out of the adjustment window, S(A:B) can be calculated under conditions satisfying $|\,iend-jend\,| < r+r/2$. In other cases, the input word is rejected.

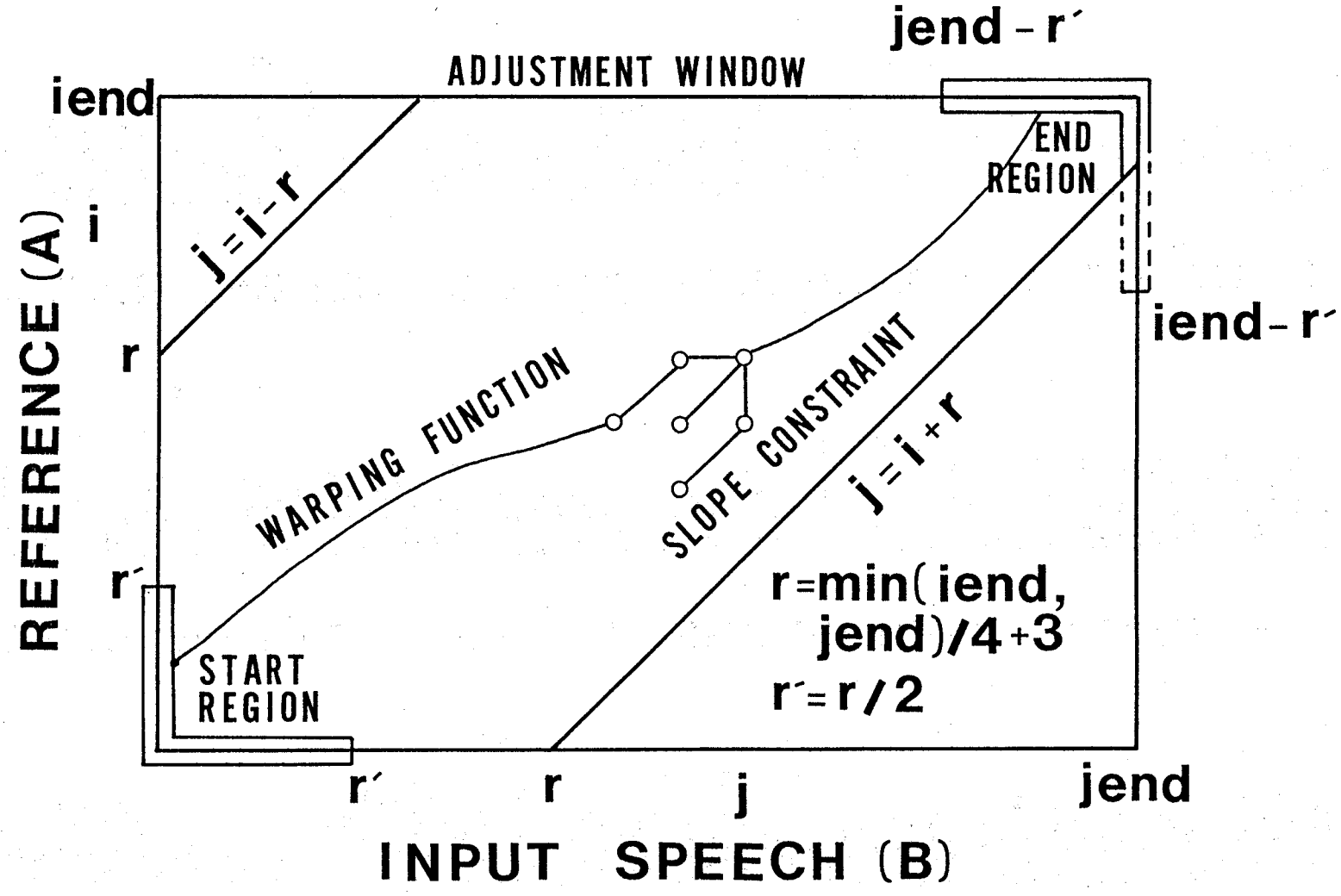


Fig. 1A Dynamic time warping algorithm.

APPENDIX 4

641 city names list for speaker-dependent word recognition.

1	SAPPORO	81	NAGAI	161	IRUMA
2	HAKODATE	82	TEN-DO	162	HATOGAYA
3	OTARU	83	HIGASHINE	163	ASAKA
4	ASAHIKAWA	84	OBANAZAWA	164	SHIKI
5	MURORAN	85	NAN-YO	165	WAKO:
6	KUSHIRO	86	FUKUSHIMA	166	NI:ZA
7	OBIHIRO	87	AIZUWAKAMATSU	167	OKEGAWA
8	KITAMI	88	KO:RIYAMA	168	KUKI
9	YU:BARI	89	IWAKI	169	KITAMOTO
10	IWAMIZAWA	90	SHIRAKAWA	170	YASHIO
11	ABASHIRI	91	HARAMACHI	171	FUJIMI
12	RUMOI	92	SUKAGAWA	172	KAMIFUKUOKA
13	TOMAKOMAI	93	KITAKATA	173	MISATO
14	WAKKANAI	94	SO:MA	174	HASUDA
15	BIBAI	95	NIHON-MATSU	175	CHIBA
16	ASHIBETSU	96	MITO	176	CHO:SHI
17	EBETSU	97	HITACHI	177	ICHIKAWA
18	AKABIRA	98	TSUCHIURA	178	FUNABASHI
19	MON-BETSU	99	KOGA	179	TATEYAMA
20	SHIBETSU	100	ISHIOKA	180	KISARAZU
21	NAYORO	101	SHIMODATE	181	MATSUDO
22	MIKASA	102	YU:KI	182	NODA
23	NEMURO	103	RYU:GASAKI	183	SAWARA
24	CHITOSE	104	NAKAMINATO	184	MOBARA
25	TAKIKAWA	105	SHIMOTSUMA	185	NARITA
26	SUNAGAWA	106	MITSUKAIDO:	186	SAKURA
27	UTASHINAI	107	HITACHIO:TA	187	TO:GANE
28	FUKAGAWA	108	KATSUTA	188	YO:KAICHIBA
29	FURANO	109	TAKAHAGI	189	ASAHI
30	NOBORIBETSU	110	KITAIBARAKI	190	NARASHINO
31	ENIWA	111	KASAMA	191	KASHIWA
32	DATE	112	TORIDE	192	KATSUURA
33	AOMORI	113	IWAI	193	ICHIHARA
34	HIROSAKI	114	UTSUNOMIYA	194	NAGAREYAMA
35	HACHINOHE	115	ASHIKAGA	195	YACHIYO
36	GOSHOGAWARA	116	TOCHIGI	196	ABIKO
37	TOWADA	117	SANO	197	KAMOGAWA
38	MISAWA	118	KANUMA	198	KAMAGAYA
39	MUTSU	119	NIKKO:	199	KIMIZU
40	MORIOKA	120	IMAICHI	200	FUTTSU
41	MIYAKO	121	OYAMA	201	HACHIO:JI
42	O:FUNATO	122	MAOKA	202	TACHIKAWA
43	MIZUSAWA	123	O:TAWARA	203	MUSASHINO
44	HANAMAKI	124	YAITA	204	MITAKA
45	KITAKAMI	125	KUROISO	205	O:ME
46	KUJI	126	MAEBASHI	206	FUCHU:
47	TO:NO	127	TAKASAKI	207	AKISHIMA
48	ICHINOSEKI	128	KIRYU:	208	CHO:FU
49	RIKUZEN-TAKATA	129	ISESAKI	209	MACHIDA
50	KAMAISHI	130	O:TA	210	KOGANEI
51	ESASHI	131	NUMATA	211	KODAIRA
52	NINOHE	132	TATEBAYASHI	212	HINO
53	SEN-DAI	133	SHIBUKAWA	213	HIGASHIMURAYAMA
54	ISHINOMAKI	134	FUJIOKA	214	KOKUBUN-JI
55	SHIOGAMA	135	TOMIOKA	215	KUNITACHI
56	FURUKAWA	136	AN-NAKA	216	TANASHI
57	KESEN-NUMA	137	KAWAGOE	217	HO:YA
58	SHIROISHI	138	KUMAGAYA	218	FUSSA
59	NATORI	139	KAWAGUCHI	219	KOMAE
60	KAKUDA	140	URAWA	220	HIGASHIYAMATO
61	TAGAJO:	141	O:MIYA	221	KIYOSE
62	IZUMI	142	GYO:DA	222	HIGASHIKURUME
63	IWANUMA	143	CHICHIBU	223	MUSASHIMURAYAMA
64	AKITA	144	TOKOROZAWA	224	TAMA
65	NOSHIRO	145	HAN-NO	225	INAGI
66	YOKOTE	146	KAZO	226	AKIKAWA
67	O:DATE	147	HON-JO:	227	YOKOHAMA
68	HON-JO:	148	HIGASHIMATSUYAMA	228	YOKOSUKA
69	OGA	149	IWATSUKI	229	HIRATSUKA
70	YUZAWA	150	KASUKABE	230	KAMAKURA
71	O:MAGARI	151	SAYAMA	231	FUJISAWA
72	KAZUNO	152	HANYU:	232	ODAWARA
73	YAMAGATA	153	KO:NOSU	233	CHIGASAKI
74	YONEZAWA	154	FUKAYA	234	ZUSHI
75	TSURUOKA	155	AGEO	235	SAGAMIHARA
76	SAKATA	156	YONO	236	MIURA
77	SHIN-JO:	157	SO:KA	237	HADANO
78	SAGAE	158	KOSHIGAYA	238	ATSUGI
79	KAMINOYAMA	159	WARABI	239	YAMATO
80	MURAYAMA	160	TODA	240	ISEHARA

241 EBINA
242 ZAMA
243 MINAMIASHIGARA
244 NI:GATA
245 NAGAOKA
246 SAN-JO:
247 KASHIWAZAKI
248 SHIBATA
249 NI:TSU
250 OJIYA
251 KAMO
252 TO:KAMACHI
253 MITSUKE
254 MURAKAMI
255 TSUBAME
256 TOCHIO
257 ITOIGAWA
258 ARAI
259 GOSEN
260 RYO:TSU
261 SHIRONE
262 TOYOSAKA
263 JO:ETSU
264 TOYAMA
265 TAKAOKA
266 SHIN-MINATO
267 UOZU
268 HIMI
269 NAMERIKAWA
270 KUROBE
271 TONAMI
272 OYABE
273 KANAZAWA
274 NANAO
275 KOMATSU
276 WAJIMA
277 SUZU
278 KAGA
279 HAKUI
280 MATSUTO:
281 FUKUI
282 TSURUGA
283 TAKEFU
284 OBAMA
285 O:NO
286 KATSUYAMA
287 SABAE
288 KO:FU
289 FUJIYOSHIDA
290 EN-ZAN
291 TSURU
292 YAMANASHI
293 O:TSUKI
294 NIRASAKI
295 NAGANO
296 MATSUMOTO
297 UEDA
298 OKAYA
299 IIDA
300 SUWA
301 SUZAKA
302 KOMORO
303 INA
304 KOMAGANE
305 NAKANO
306 O:MACHI
307 IIYAMA
308 CHINO
309 SHIOJIRI
310 KO:SHOKU
311 SAKU
312 GIFU
313 O:GAKI
314 TAKAYAMA
315 TAJIMI
316 SEKI
317 NAKATSUGAWA
318 MINO
319 MIZUNAMI
320 HASHIMA
321 ENA
322 MINOKAMO
323 TOKI
324 KAKAMIGAHARA
325 SHIZUOKA
326 HAMAMATSU
327 NUMAZU
328 SHIMIZU
329 ATAMI
330 MISHIMA
331 FUJINOMIYA
332 ITO:
333 SHIMADA
334 FUJI
335 IWATA
336 YAIZU
337 KAKEGAWA
338 FUJIEDA
339 GOTEN-BA
340 FUKUROI
341 TEN-RYU:
342 HAMAKITA
343 SHIMODA
344 SUSONO
345 KOSAI
346 NAGOYA
347 TOYOHASHI
348 OKAZAKI
349 ICHINOMIYA
350 SETO
351 HAN-DA
352 KASUGAI
353 TOYOKAWA
354 TSUSHIMA
355 HEKINAN
356 KARIYA
357 TOYOTA
358 AN-JO:
359 NISHIO
360 GAMAGO:RI
361 INUYAMA
362 TOKONAME
363 KO:NAN
364 BISAI
365 KOMAKI
366 INAZAWA
367 SHIN-SHIRO
368 TO:KAI
369 O:BU
370 CHITA
371 CHIRYU:
372 OWARIASAHI
373 TAKAHAMA
374 IWAKURA
375 TOYOAKE
376 TSU
377 YOKKAICHI
378 ISE
379 MATSUSAKA
380 KUWANA
381 UENO
382 SUZUKA
383 NABARI
384 OWASE
385 KAMEYAMA
386 TOBA
387 KUMANO
388 HISAI
389 O:TSU
390 HIKONE
391 NAGAHAMA
392 O:MIHACHIMAN
393 YO:KAICHI
394 KUSATSU
395 MORIYAMA
396 KYO:TO
397 FUKUCHIYAMA
398 MAIZURU
399 AYABE
400 UJI
401 MIYAZU
402 KAMEOKA
403 JO:YO:
404 MUKO:
405 NAGAOKAKYO:
406 O:SAKA
407 SAKAI
408 KISHIWADA
409 TOYONAKA
410 IKEDA
411 SUITA
412 IZUMIO:TSU
413 TAKATSUKI
414 KAIZUKA
415 MORIGUCHI
416 HIRAKATA
417 IBARAKI
418 YAO
419 IZUMISANO
420 TON-DABAYASHI
421 NEYAGAWA
422 KAWACHINAGANO
423 MATSUBARA
424 DAITO:
425 IZUMI
426 MINOO
427 KASHIWARA
428 HABIKINO
429 KADOMA
430 SETTSU
431 TAKAISHI
432 FUJIIDERA
433 HIGASHIO:SAKA
434 SEN-NAN
435 SHIJO:NAWATE
436 KATANO
437 KO:BE
438 HIMEJI
439 AMAGASAKI
440 AKASHI
441 NISHINOMIYA
442 SUMOTO
443 ASHIYA
444 ITAMI
445 AIOI
446 TOYOOKA
447 KAKOGAWA
448 TATSUNO
449 AKO:
450 NISHIWAKI
451 TAKARAZUKA
452 MIKI
453 TAKASAGO
454 KAWANISHI
455 ONO
456 SAN-DA
457 KASAI
458 NARA
459 YAMATOTAKADA
460 YAMATOKO:RIYAMA
461 TEN-RI
462 KASHIHARA
463 SAKURAI
464 GOJO:
465 GOSE
466 IKOMA
467 WAKAYAMA
468 KAINAN
469 HASHIMOTO
470 ARIDA
471 GOBO:
472 TANABE
473 SHIN-GU:
474 TOTTORI
475 YONAGO
476 KURAYOSHI
477 SAKAIMINATO
478 MATSUE
479 HAMADA
480 IZUMO

481 MASUDA
482 O:DA
483 YASUGI
484 GO:TSU
485 HIRATA
486 OKAYAMA
487 KURASHIKI
488 TSUYAMA
489 TAMANO
490 KASAOKA
491 IBARA
492 SO:JA
493 TAKAHASHI
494 NI:MI
495 BIZEN
496 HIROSHIMA
497 KURE
498 TAKEHARA
499 MIHARA
500 ONOMICHI
501 IN-NOSHIMA
502 FUKUYAMA
503 FUCHU:
504 MIYOSHI
505 SHO:BARA
506 O:TAKE
507 HIGASHIHIROSHIMA
508 SHIMONOSEKI
509 UBE
510 YAMAGUCHI
511 HAGI
512 TOKUYAMA
513 HO:FU
514 KUDAMATSU
515 IWAKUNI
516 ONODA
517 HIKARI
518 NAGATO
519 YANAI
520 MINE
521 SHIN-NAN-YO:
522 TOKUSHIMA
523 NARUTO
524 KOMATSUSHIMA
525 ANAN
526 TAKAMATSU
527 MARUGAME
528 SAKAIDE
529 ZEN-TSU:JI
530 KAN-ON-JI
531 MATSUYAMA
532 IMABARI
533 UWAJIMA
534 YAWATAHAMA
535 NI:HAMA
536 SAIJO:
537 O:ZU
538 KAWANOE
539 IYOMISHIMA
540 IYO
541 HO:JO:
542 TO:YO
543 KO:CHI
544 MUROTO
545 AKI
546 NAN-KOKU
547 TOSA
548 SUSAKI
549 NAKAMURA
550 SUKUMO
551 TOSASHIMIZU
552 KITAKYU:SHU:
553 O:MUTA
554 KURUME
555 NO:GATA
556 I:ZUKA
557 TAGAWA
558 YANAGIGAWA
559 YAMADA
560 AMAGI
561 YAME
562 CHIKUGO
563 O:KAWA
564 YUKUHASHI
565 BUZEN
566 NAKAMA
567 OGO:RI
568 CHIKUSHINO
569 KASUGA
570 O:NOJO:
571 FUKUOKA
572 SAGA
573 KARATSU
574 TOSU
575 TAKU
576 IMARI
577 TAKEO
578 KASHIMA
579 NAGASAKI
580 SASEBO
581 SHIMABARA
582 ISAHAYA
583 O:MURA
584 FUKUE
585 HIRADO
586 MATSUURA
587 KUMAMOTO
588 YATSUSHIRO
589 HITOYOSHI
590 ARAO
591 MINAMATA
592 TAMANA
593 HON-DO
594 YAMAGA
595 USHIBUKA
596 KIKUCHI
597 UTO
598 O:ITA
599 BEPPU
600 NAKATSU
601 HITA
602 SAIKI
603 USUKI
604 TSUKUMI
605 TAKETA
606 BUN-GOTAKADA
607 KITSUKI
608 USA
609 MIYAZAKI
610 MIYAKONOJO:
611 NOBEOKA
612 NICHINAN
613 KOBAYASHI
614 HYU:GA
615 KUSHIMA
616 SAITO
617 EBINO
618 KAGOSHIMA
619 SEN-DAI
620 KANOYA
621 MAKURAZAKI
622 KUSHIKINO
623 AKUNE
624 NAZE
625 IZUMI
626 O:KUCHI
627 IBUSUKI
628 KASEDA
629 KOKUBU
630 NISHINOOMOTE
631 TARUMIZU
632 NAHA
633 ISHIKAWA
634 GUSHIKAWA
635 GINOWAN
636 HIRARA
637 ISHIGAKI
638 URASOE
639 NAGO
640 ITOMAN
641 OKINAWA

APPENDIX 5
31 words list for speaker-independent word recognition

1. HAI
2. IIE
3. SYU:RYO: JIKAN
4. SYORI JIKAN
5. UNYO: KEIKAKU
6. ONSEI TO:ROKU
7. KAIGI SHITSU
8. SENTA: BATTI
9. EFU BATTI
10. ICHI GO:
11. NI GO:
12. YOKOSUKA
13. UNTEN KEIKAKU
14. RENRAKU JIKO:
15. GYO:JI YOTEI
16. KON SYU:
17. RAI SYU:
18. GETSU YO:
19. KA YO:
20. SUI YO:
21. MOKU YO:
22. KIN YO:
23. DO YO:
24. ICHI MARU NI
25. GO: MARU GO:
26. TEREBI KAIGI SHITSU
27. GOZEN
28. GOGO
29. ICHI NICHI
30. YAMADA TARO:
31. SUZUKI HANAKO