Evaluation of Various Parameter Sets in Spoken Digits Recognition

AKIRA ICHIKAWA, YASUAKI NAKANO, and
KAZUO NAKATA

Abstract—Various parameter sets—including a spectrum envelope, cepstrum, autocorrelation function, linear predictive coefficients, and partial autocorrelation coefficients (PAC's)—are evaluated experimentally to determine which constitutes the best parameter in spoken digit recognition.

The principle of recognition is simple pattern matching in the parameter space with nonlinear adjustment of the time axis.

The spectrum envelope and cepstrum attain the best recognition score of 100 percent for ten spoken digits of a single-male speaker.

PAC's seem to be preferable because of their ease of extraction and theoretical orthogonalities; however, these PAC's tend to suffer from computation errors when computed by fixed-point arithmetic with a short accumulator length. We find two effective means to improve the errors; one is variable use of the PAC dimensions controlled by computation accuracy, and the other is smoothing along the time axis.

With these improvements the PAC's offer almost 100 percent recognition.

I. Objectives

The main difficulties of speech recognition in general are caused by the variation of physical characteristics of speech due to a change of speakers and phonetic environments.

For recognition of a limited vocabulary, pattern matching of word units can be workable. Variation due to a change of speakers is avoided by the use of a speaker's own utterances as the standard, and variation due to a change of phonetic environments is also avoided by the use of word unit pattern matching.

A nonuniform change of word duration and syllable duration appears as the most difficult obstacle that remains, but this obstacle has recently been removed by the introduction of a nonuniform time pattern matching algorithm based on dynamic programming (DP) techniques [1]-[3].

Under these situations, it becomes very important to solve the problem of determining which is the best speech parameter in practical application of speech recognition.

This paper describes the results of our experimental research on this problem and their further developments.

II. Various Parameter Sets Being Tested

Filter banks are the most popular means used to analyze speech. But this method has certain defects, one of which is a variation of filter output by a change of pitch frequency.

The power spectrum of speech is represented as

$$P(\omega) = |S(\omega)|^2 = |G(\omega)|^2 \, |H(\omega)|^2 = |V(\omega)|^2 \, |E(\omega)|^2 \tag{1}$$

where $S(\omega)$, $G(\omega)$, and $H(\omega)$ represent spectra of speech, excitation function, and transmission characteristics of a vocal tract, $V(\omega)$ is the idealized excitation spectrum (periodic line spectrum for voiced and uniform for unvoiced), and $E(\omega)$ is the spectrum envelope, respectively. Ordinarily, phonemic information—except voice-unvoice distinction—is considered to be conveyed mainly by the spectrum envelope $V(\omega)$; however, it is difficult to extract exactly the spectrum envelope $V(\omega)$ or transfer function $H(\omega)$.

If the power spectrum is smoothed along the frequency axis, the effect of excitation is reduced to be less apparent and a spectrum envelope can be extracted approximately. However, this method of envelope extraction is considered insufficient compared with cepstrum techniques described in Section II, B. Since an approximate spectrum envelope and output of filter banks are similar to a cepstrum in their information content, we omitted them from the parameter sets being tested. The tested parameters are described briefly later.

A. Autocorrelation Function

It is well known that an autocorrelation function is obtained from a power spectrum by the inverse Fourier transform relation. Autocorrelation functions have the same defects as does the power spectrum.

B. Cepstrum and Smoothed Logarithmic Power Spectrum

Quefrency limited cepstrum $C(t)$ is defined as

$$C(t) = l(t) \cdot C^*(t)$$
$$= \begin{cases} \displaystyle\int_{-\infty}^{\infty} \exp(j\omega t) \log P(\omega) d\omega & (t \leqq T), \\ 0 & (t > T), \end{cases} \tag{2}$$

where

$$l(t) = \begin{cases} 1 & (t \leqq T) \\ 0 & (t > T). \end{cases} \tag{3}$$

and $P(\omega)$ is power spectrum.

A spectrum envelope $E(\omega)$ is approximated by smoothing the logarithmic power spectrum.

Manuscript received April 30, 1972.

The authors are with the Central Research Laboratory, Hitachi Limited, Tokyo, Japan.

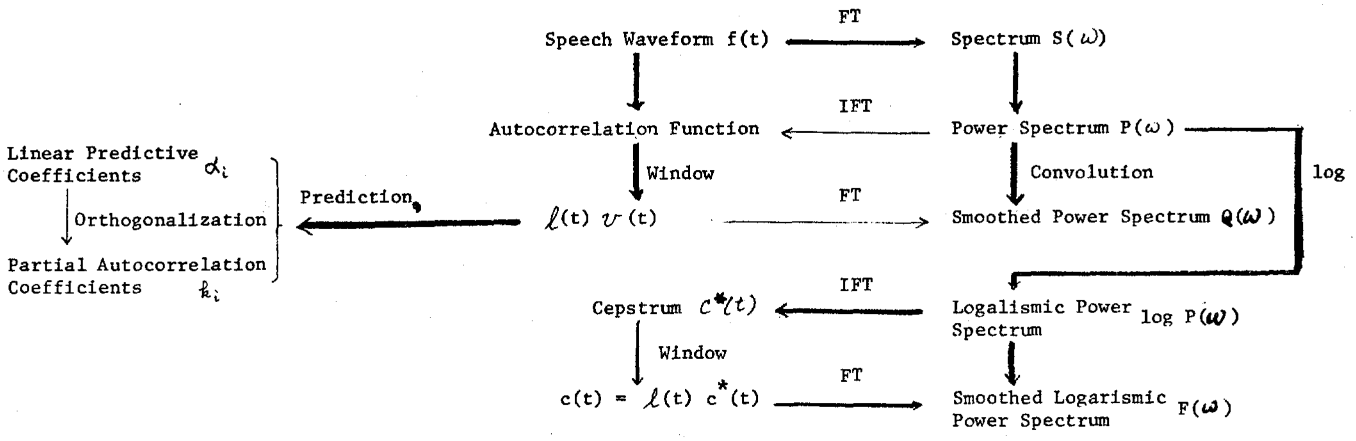

Fig. 1. Relation between parameter sets considered: Fourier transform (FT) and inverse Fourier transform (IFT).

$$F(\omega) = \int_0^\infty L(\omega - \xi)\log P(\xi)\,d\xi \quad (0 \leqq \omega \leqq \infty), \quad (4)$$

and $L(\omega)$ is the Fourier transform of the window function $l(t)$. $C(t)$ and $F(\omega)$ constitute a Fourier transform pair.

Quefrency limited cepstrum $C(t) = l(t)\,C^*(t)$ and $F(\omega)$ constitute a Fourier transform pair. $C(t)$ suffers from less affect of pitch frequency changes than does the power spectrum $Q(\omega)$, because the logarithmic operation transforms the multiplication of G and H frequency filtering or E and V in (1) into addition and can eliminate high quefrency components that correspond to the excitation source; by the same reason of cepstrum, $F(\omega)$ suffers from less affect of pitch frequency changes than the power spectrum $P(\omega)$ or $Q(\omega)$ as the simple smoothed $P(\omega)$.

$$Q(\omega) = \int_0^\infty L(\omega - \xi)P(\xi)\,d\xi. \quad (5)$$

C. Linear Predictive Coefficients

Letting the sampled speech waveforms be

$$f_i = f(i\Delta) \quad (\Delta\text{: sampling period}) \quad (6)$$

and assuming that $f(t)$ has a spectrum of the rational function of ω, the linear predictive coefficients set $\{\alpha_i\}$ that renders the squared predictive error ϵ^2 minimum are obtained by solution of the following matrix equation [4].

$$\begin{bmatrix} v_0 & v_1 & \cdots & v_{p-1} \\ v_1 & v_0 & \cdots & v_{p-2} \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ v_{p-1} & v_{p-2} & & v_0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_p \end{bmatrix} = - \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_p \end{bmatrix} \quad (7)$$

where

$$v_i = \frac{1}{N}\sum_{j=0}^{N-1-i} f_j f_{j+i}. \quad (8)$$

The power spectrum envelope $|E(\omega)|^2$ of $f(t)$ is estimated as

$$|E(\omega)|^2 = \frac{K}{\left|\sum_{i=0}^{p} \alpha_i \zeta^i\right|^2} \quad [\zeta = \exp(j\omega\Delta)]. \quad (9)$$

D. Partial Autocorrelation Coefficients (PAC's)

It has been reported that the most efficient decomposition of speech is attained by PAC's [6].

The PAC's of delay $(n+1)\Delta$ is defined as the correlation coefficient between samples f_t and $f_{t-(n+1)}$ after removing the influence of samples between the two,

$$f_{t-n}, f_{t-n+1}, \cdots, f_{t-1}.$$

The influence of these samples are linearly estimated using α_i's in (7) of order n [written as $\alpha_i^{(n)}$] as

$$\hat{f}_t = -\sum_{i=0}^{n} \alpha_i^{(n)} f_{t-i} \quad \text{(forward prediction)},$$

$$\hat{f}_{t-(n+1)} = -\sum_{i=1}^{n+1} \alpha_{n+1-i}^{(n)} f_{t-i} \quad \text{(backward prediction)}. \quad (10)$$

Using these estimations, the PAC of delay $(n+1)\Delta$ is given by

$$k = \frac{E[(f_t - \hat{f}_t)(f_{t-(n+1)} - \hat{f}_{t-(n+1)})]}{\{E[(f_t - \hat{f}_t)^2]\,E[(f_{t-(n+1)} - \hat{f}_{t-(n+1)})^2]\}^{1/2}} \quad (11)$$

where $E[\ \ ]$ represents the expected values.

On the other hand, it can be shown that PAC's k_i are obtained by orthogonalization of the sets of parameter $\alpha_i^{(n)}$ of the order 1 to p. Note that the values of $\alpha_i^{(n)}$ $(i \leqq p)$ change according to the change of order p, while the values $\alpha_p^{(n)}$ $(i \leqq p)$ are independent of the order p.

Fig. 1 shows the relation of various parameter sets stated previously. The process of extracting parameters is rendered by heavy lines.

III. Recognition Algorithm

A simple pattern matching with nonuniform transformation of time axis is adopted as the recognition algorithm.

A. Nonuniform Transformation of Time Axis

Matching between two vector functions $x(t)$ and $y(t')$ given as time series along the time mapping function

$$t' = \rho(t) \tag{12}$$

is evaluated by

$$R_\rho = \sum_t D(x(t), y(\rho(t))) \tag{13}$$

where D represents a scalar function evaluating the similarity between two vectors.

If $\rho(t)$ represents parallel translation or uniform expansion (or compression), the procedure to maximize R_ρ and obtain

$$R = \max_{\rho(t)} R_\rho \tag{14}$$

becomes relatively easy.

When $\rho(t)$ is the nonlinear transforming function, with the condition of monotonic nondecreasing, the optimum path (along which R_ρ becomes maximum) searching algorithm is formulated and executed by DP techniques [1], [2].

In our case, boundary conditions of DP are fixed-starting point and free-ending point, and diagonal paths of the matching are admitted (see Fig. 2). Details of DP formulation of the problem and programming procedure are found in [1] and [2] and not repeated here.

B. Range of Transformation of Time Axis

The range of nonlinear transformation of time is the area in (t, t')-plane of Fig. 2 determined by the relation

$$t - d \leqq t' \leqq t + d \tag{15}$$

where d is the range of time axis transformation. Any transformation $t' = \rho(t)$ satisfying (15) and monotonic nondecreasing relation represented by

$$\frac{d\rho(t)}{dt} \geqq 0 \tag{16}$$

is taken into account.

If d is set as large, the transformation has wide flexibility but consumes much time for matching calculation, and in some cases, forced false matching occurs. Then d must be kept as small as possible in the range within which the affect of nonlinear transformation is effective.

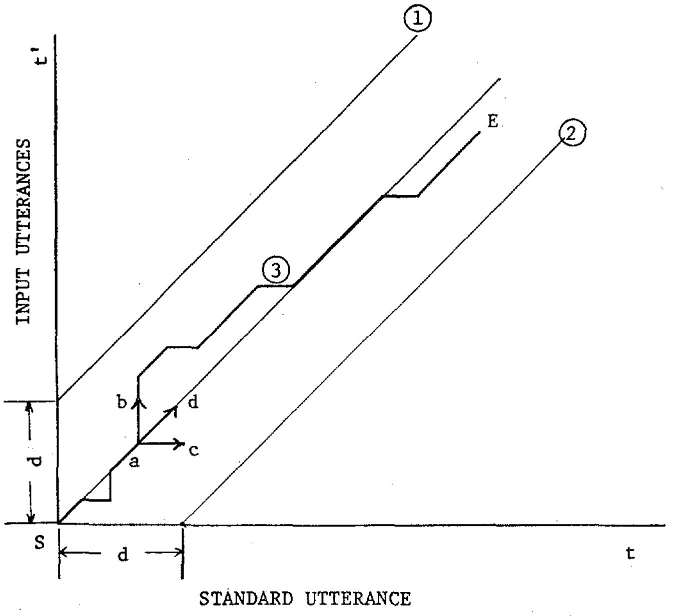

Fig. 2. Range of transformation of time axis. Matching path ③ of input and standard utterances is permitted between lines ① and ②. And the path from point a is only admitted to points b, c (straight paths), or d (diagonal path). S is the starting point and E is the ending point.

C. Similarity Evaluation Function

The similarity of two vectors of dimension N,

$$\begin{aligned} x &= (x_0, x_1, \cdots, x_{N-1}), \\ y &= (y_0, y_1, \cdots, y_{N-1}), \end{aligned} \tag{17}$$

is evaluated in two ways. One of them, used for normalized parameters such as PAC's, is the difference between 1 and the simple sum of squared differences of the respective coefficients

$$D_1 = 1 - \frac{1}{p} \sum_i^p (x_i - y_i)^2. \tag{18}$$

The other used for nonnormalized parameters is the inner product of x and y after normalization

$$D_2 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{19}$$

where $\bar{x}$ and $\bar{y}$ are the means of x_i and y_i. We set the similarity D between silent segments of speech to 1 and that between silent and nonsilent segments to 0.

To avoid variations of similarity evaluation due to unequal length of the matching path, we adopted $R'_\rho = R_\rho / T_{\max}$, as R_ρ of (14), where $T_{\max}$ is the longer duration of two words being matched.

IV. Speech Utterances and Analyzing Conditions

A. Speech Utterances

Speech utterances used in the recognition experiment are ten Japanese digits uttered by a professional male announcer. Each digit is uttered ten times in random order.

Each utterance is band-limited over a range from 150–4200 Hz and converted to 11-b digital samples with a sampling frequency of 10 kHz. The sampling period written as Δ is then 100 μs.

From 100 spoken utterances, 10 utterances (one for each digit) are selected at random as standards and the remainders are used for test utterances.

The experiment of recognition is carried out with a HITAC 5020F computer.

B. Condition of Analysis

The condition of analysis for extracting parameter sets are as follows.

Analyzing Time Window Length: the number of samples processed at one time is taken to 256 that that corresponds to 25.6 ms. This value of the time window length is selected by preliminary experiments and unchanged throughout the experiment.

Analyzing interval: speech signals are analyzed every 10 ms.

Liftering Quefrency: the logarithm of power spectrum is low-passfiltered with an equivalent liftering quefrency of 1.7 ms.

Liftering in the Time Domain: cepstrum is liftered to pass the band (a, b)

$$c_i = \begin{cases} c^*(i\Delta) & (a \leqq i \leqq b), \\ 0 & (\text{otherwise}) \end{cases} \qquad (20)$$

in the quefrency domain.

The cutoff quefrencies are multiples of the sampling period in the quefrency domain that corresponds with Δ. Noticing the pitch periods (usually male pitchs periods are greater than 50), the values of a and b are taken in the range of

$$0 \leqq a \leqq 4,$$
$$5 \leqq b \leqq 11. \qquad (21)$$

Correlation Function: the correlation function is calculated only in the range of delay (a, b)

$$v_i = \begin{cases} \sum_{j=0}^{v-1-i} f_j f_{j+i} & (a \leqq i \leqq b), \\ 0 & (\text{otherwise}), \end{cases} \qquad (22)$$

where f_i are sampled values of speech waves weighted by the Hamming window of duration N. The range of delay is selected equivalent to the liftering quefrency of cepstrum. The a and b are taken in the range

$$0 \leqq a \leqq 1,$$
$$3 \leqq b \leqq 11. \qquad (23)$$

Linear Predictive Coefficients: the order of prediction p is changed 3–12. v_i's are calculated from $i = 0\text{–}p$.

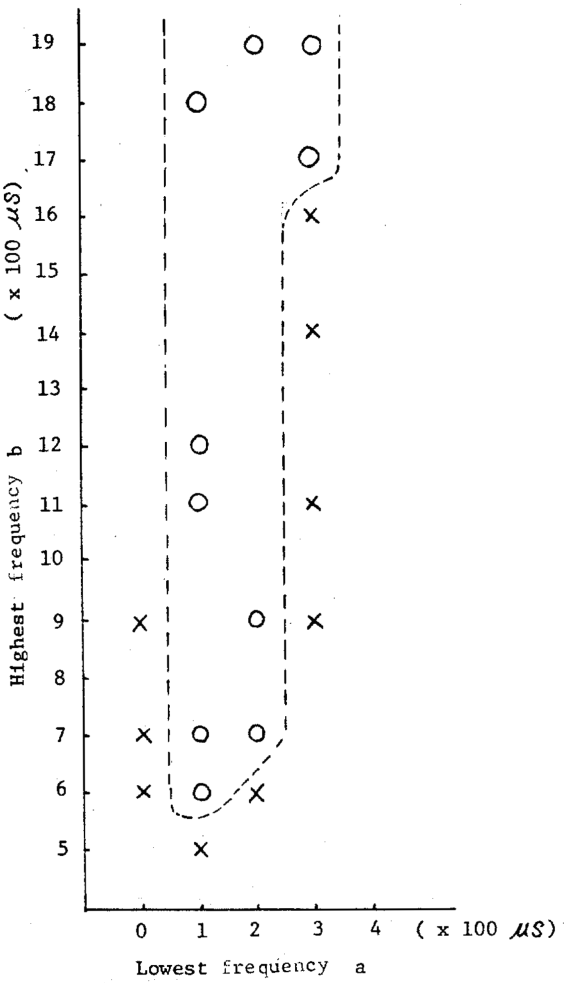

Fig. 3. Recognition score using cepstrum. Limited range of cepstrum is used for recognition. Full circles represent the recognition score of 100 percent and crosses represent that of less than 100 percent.

V. Experimental Results

A. Smoothed Logarithmic Power Spectrum

Smoothed logarithmic power spectrum $F(\omega)$ is resampled at every 200-Hz sampling interval in order to obtain 25 channel equivalents.

The recognition score of 100 percent is obtained by using this spectrum envelope approximation, when the range of nonlinear transformation of time in (15) is greater than 3. Accordingly, d is fixed to 4 throughout subsequent experiments.

B. Cepstrum

To determine the optimum quefrency range, parameter a and b are changed in the range indicated in (21).

Fig. 3 shows the region of frequency range that gives the recognition score of 100 percent.

Fig. 3 shows that cutoff frequency must be greater than 0. It may be explained as due to the fact that the average spectrum level (dc component of logarithmic spectrum) is less effective than the higher ones for speech recognition. Fig. 3 also shows that

TABLE I
Comparison of Parameter Sets

Parameter Set	Dimension[a]	a	b[b]	Recognition Score	D[c]	Time[d]
Smoothed logarithmic power spectrum $F(\omega)$	25	—	—	100 percent	D_2	2.0
Cepstrum $C(t)$	6	1 2	6 7	100 percent	D_2	1.0
Autocorrelation function $v(t)$	4	1	4	92 percent	D_2	0.4
Linear predictive coefficients α_i	7	1	7	77 percent	D_2	0.5
Partial autocorrelation coefficients k_i	5	1	5	98 percent	D_1	0.5

[a] Minimum dimension of parameter vector that gives the best result.
[b] Range of dimension.
[c] Type of similarity evaluation function.
[d] Relative time required for recognition (normalized to 1 for cepstrum).

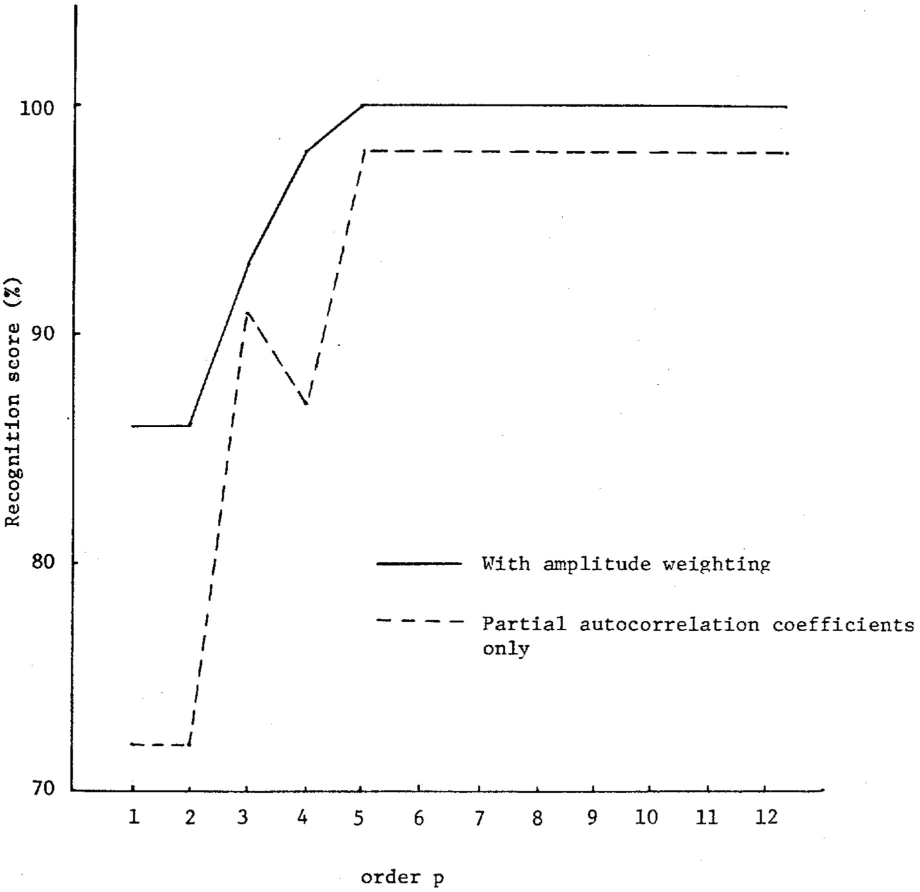

Fig. 4. Recognition score using partial autocorrelation coefficients.

only six parameters are required for 100 percent recognition of ten spoken digits.

C. Autocorrelation Function and Linear Predictive Coefficients

The recognition experiments using autocorrelation functions v_i, or linear predictive coefficients α_i produce unexpectedly inferior results that are tabulated in Table I.

D. PAC's

The reason why the parameters α_i produce such a low recognition rate may be explained by the statistical dependency of parameters on each other. The PAC's k_i are promising in this aspect because of their coordinate orthogonality.

In Fig. 4, the recognition score using k_i $(i \leqq p)$ with a change in the order p are shown by the dotted line. It is obvious that the parameters k_i are much superior to parameters α_i. The error rate, however, is still relatively high and inferior to the cepstrum.

E. Comparison of the Parameter Sets

The recognition scores using various parameter sets are tabulated in Table I. And also the table indicates the relative time (the time normalized to 1 for cepstrum) required for each parameter extraction.

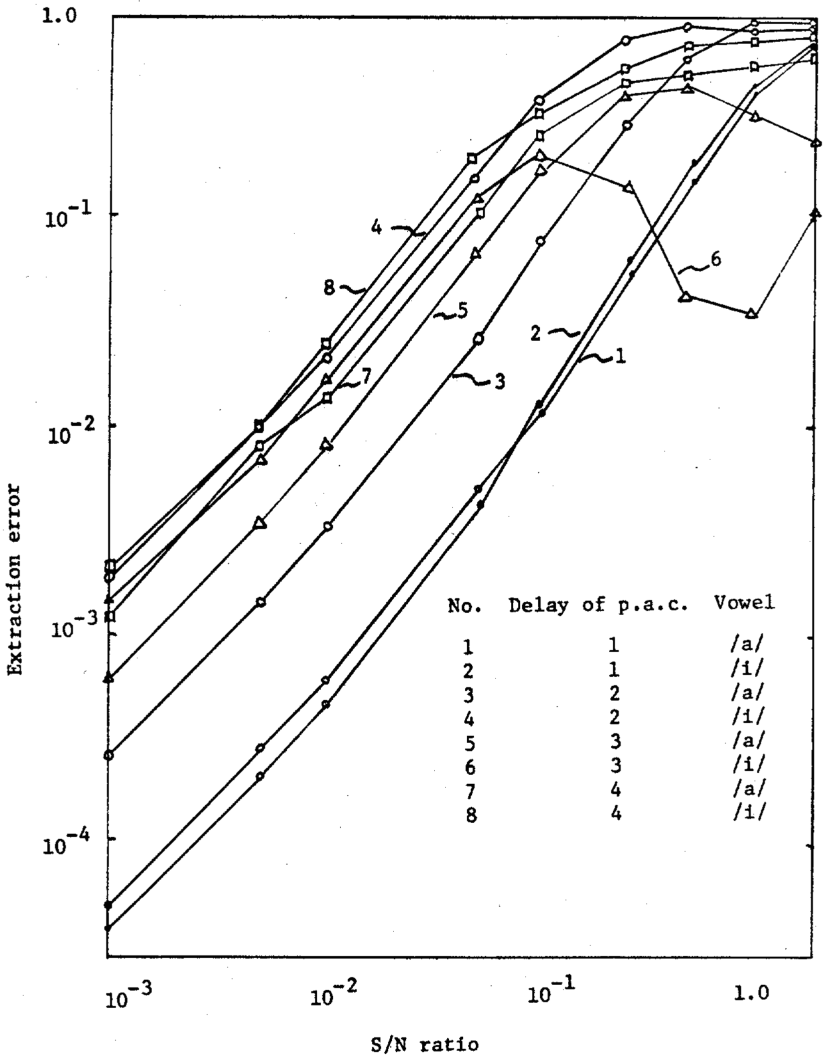

Fig. 5. Extraction error of partial autocorrelation coefficients due to S/N ratio. Error means the absolute difference between the value of PAC from noisy synthetic vowels and that without noise. Noise is a synthetic Gaussian one.

Table I reveals that the spectrum envelope and cepstrum attain the best score of 100 percent.

PAC's attain the promising result of 98 percent and seem preferable because of the ease and relatively short time required in extracting them. Accordingly, certain efforts to improve the recognition score is exerted.

VI. Improvements of Recognition Using PAC's

The main reason for the recognition error by using PAC's k_i seems to originate from the deterioration of accuracy at a small signal amplitude.

The PAC's k_i are derived from normalized correlation coefficients and are independent of signal amplitude. Accordingly, the contribution of PAC's are uniform throughout the word. The accuracy of computation of PAC, however, becomes low when the signal amplitude becomes small because of a decrease of the S/N ratio. Fig. 5 shows the change of extraction error of PAC when a random noise is added to the speech signals and S/N ratio has been changed. Three methods of improvement are described in the following.

A. Weighting by Signal Amplitude

To avoid the influence of deterioration of extraction accuracy, the contribution of PAC's to recognition is weighted by the signal amplitude. In this sense (13) is rewritten as

$$R_\rho = \sum_t w(t) D(x(t), y(\rho(t))) \qquad (13')$$

where $w(t)$ is the weighting function such as

$$w(t) = 1 + c \cdot q(t) \qquad (24)$$

and $q(t)$ represents the normalized relative amplitude (the maximum amplitude in a given word is normalized to 1). The constant c is assumed as 1 in this experiment.

The results of recognition using amplitude weighting is also shown in Fig. 4 by the solid line. The best score of 100 percent is obtained when the order of estimate p is greater than 5. The minimum number of parameters required to attain the best score of 100 percent is 6, counting the amplitude as one of them, and it is identical with that of the cepstrum.

B. Variable Use of Number of Parameters Controlled by their Reliabilities

The basic idea of this improvement is that the recognition score must improve when only the reliable parameters that have small extraction errors are used.

What characteristics does the extraction error of the PAC have? In the case of an easily predictable simple waveform or small-signal amplitude, the predictive residue, $f - \hat{f}$ of (11) becomes small and the S/N ratio of the predictive residue decreases. Then predictive residue becomes small in general as the order of PAC increases. Furthermore, the calculating method of PAC's is considered to be a process for solving linear simultaneous equations, and every method of solving linear simultaneous equations must use some recursive calculation process to calculate the next order coefficient from the residue of the present order. Then the calculating error propagates and increases as the order of recursive calculation increases, once the error has arisen.

From a practical viewpoint, it is desirable that these coefficients be calculated by fixed-point calculations with a small accumulator length but the higher order coefficients become unreliable more often.

Then, we must check the extraction error or reliability of calculated coefficients at every order, and if it is considered unreliable with a certain threshold reference, exclude it from the reliable effective parameters.

It is confirmed that the coefficients, calculated by processors that have floating-point arithmetic of a 24-b mantissa part and 8-b exponent part, have sufficient accuracy. Absolute differences between these coefficients and those calculated by 16-b fixed-point arithmetic are regarded as errors. Input is the Japanese vowel /e/. (The same digitized signal is used in both calculations.) The errors propagate roughly in the form of

$$\epsilon(n) = 0.00005 \times 10^{0.2n} \tag{25}$$

where n is the order of PAC's.

To ensure using only reliable parameters and to exclude unreliable ones from recognition use, the extraction of parameters is stopped when the predictive residue become smaller than a certain threshold.

Two types of threshold are considered. The one is a constant threshold and the other is a variable-type such as formula (25),

$$u(n) = \theta \times 10^{\delta n} \tag{26}$$

where θ and δ are predetermined constants.

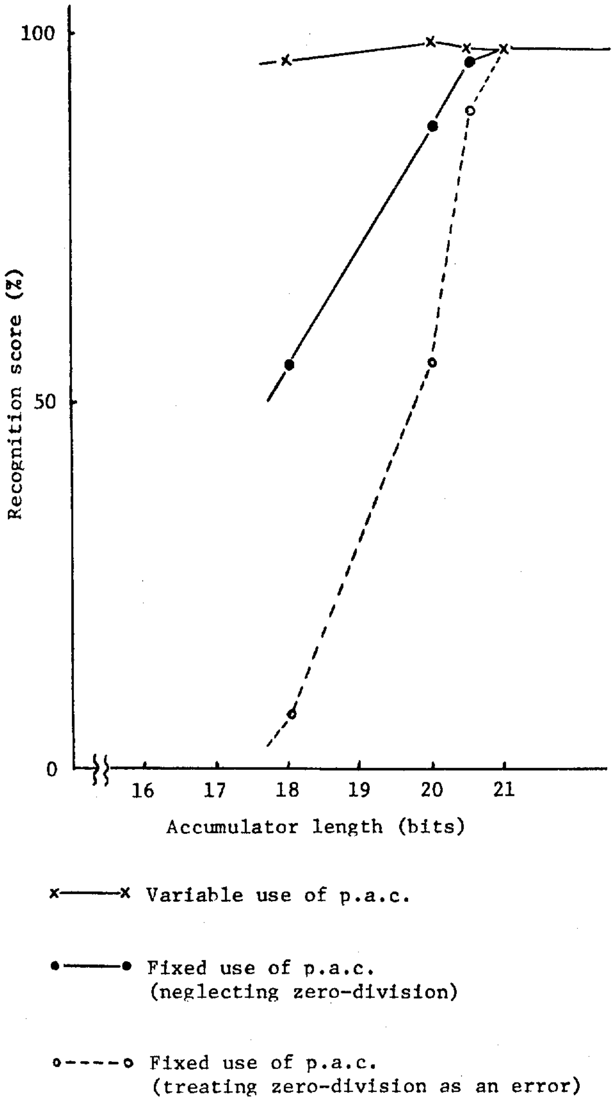

Fig. 6. Change of recognition score due to accumulator length. Fixed-point arithmetics. $k_1 \sim k_6$ are used as parameters.

Fig. 6 shows the recognition scores obtained by implementing the previously mentioned method on the previous test utterances. The computation method is the conjugate gradient-type and the highest order of coefficient is limited to 6 even if the coefficients have still enough accuracies. When using a short-length accumulator, this method can remarkably improve the recognition score. For the small amplitude speech waveforms, the effective accumulator length becomes small, and this method can stabilize the recognition score at a high level. The threshold is a variable-type, and θ and δ are chosen optimally by experiments. As an example, optimum pairs of θ and δ for 19-b processors are shown in Fig. 7.

To solve the linear simultaneous equations, we tested Crout's method and also obtained good results.

C. Smoothing of Parameters Along Time (Low-Pass Filtering)

Regarding speech recognition, it is useful to elim-

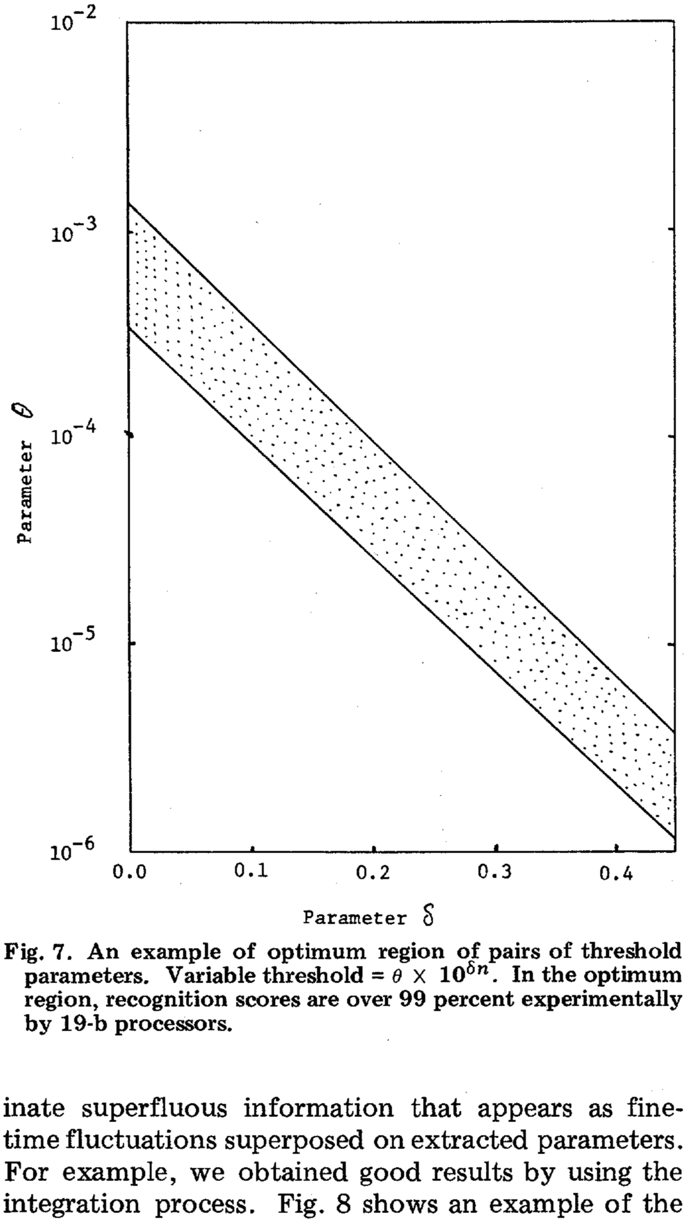

Fig. 7. An example of optimum region of pairs of threshold parameters. Variable threshold = $\theta \times 10^{\delta n}$. In the optimum region, recognition scores are over 99 percent experimentally by 19-b processors.

inate superfluous information that appears as fine-time fluctuations superposed on extracted parameters. For example, we obtained good results by using the integration process. Fig. 8 shows an example of the effect of low-pass filtering for parameters using 16-b fixed-point accumulators.

VII. Conclusion and Remarks

Among the various tested parameter sets, the cepstrum and spectrum envelope give the best results, and PAC's give the next best score.

The number of parameters necessary to attain the best score is 6 for both cepstrum and PAC's; however, the latter is far more easy to extract from speech waves.

By intensive experimental research, two methods to improve the recognition score by PAC's are detected. One of them, more effective when computation accuracy is adequate, is the combination of amplitude weighting of the matching process and the smoothing of time of each parameter. The other, especially effective when computation is less accurate (for example, in the case of small amplitude input utterances), is a combination of the variable use of PAC's depending on their extraction accuracy and smoothing of each parameter in the time axis.

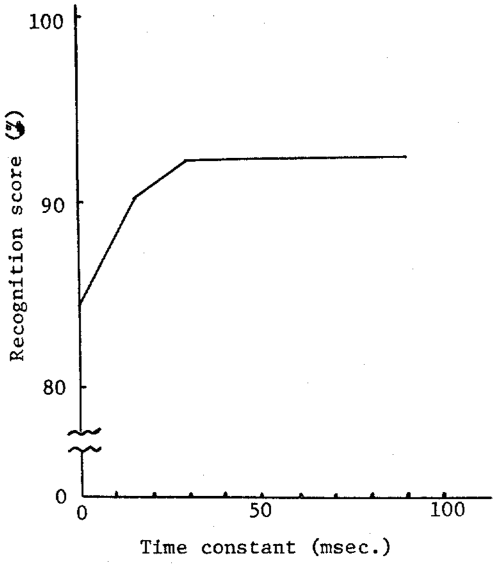

Fig. 8. Effect of time smoothing of partial autocorrelation coefficients. . . . Fixed-point arithmetics. Accumulator length is 16 b.

With these improvements, almost perfect recognition is obtained by PAC's.

The previous conclusion is derived from not only ten digits uttered by one single-male speaker but from numerous speakers. The choice of a standard from every 10 utterances takes on a slight change in the recognition score but confirms the conclusion. Efforts to substantiate this conclusion through experiments with larger vocabularies and speakers are necessary.

Acknowledgment

The authors wish to thank both H. Yoshine, Director of the 5th Research Division of the Central Research Laboratory, Hitachi Limited, and Dr. T. Miura, Chief Researcher of the Laboratory for their encouragement and advice.

References

[1] V. M. Velichko *et al.*, "Automatic recognition of 200 words," *Int. J. Man-Machine Studies*, vol. 2, p. 223, July 1970.

[2] H. Sakoe *et al.*, "A dynamic programming approach to continuous speech recognition," in *Proc. 7th ICA*, Aug. 1971, 10-c-13.

[3] M. Kohda *et al.*, "Learning of reference patterns and its application of spoken digits," presented at the 1969 ASJ Meeting, paper 2-2-22 (in Japanese).

[4] F. Itakura *et al.*, "A statistical method for estimation of speech spectral density and formant frequencies," *J. Inst. Electron. Commun. Eng. Jap.*, vol. 53-A, p. 25, Jan. 1970.

[5] B. Atal *et al.*, "Adaptive predictive coding of speech signals," *Bell Syst. Tech. J.*, vol. 49, p. 1973, 1970.

[6] F. Itakura *et al.*, "Digital filtering techniques for speech analysis and synthesis," in *Proc. 7th ICA*, Aug. 1971, 25-c-1, p. 261.