Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum

SADAOKI FURUI, MEMBER, IEEE

Abstract—**This paper proposes a new isolated word recognition technique based on a combination of instantaneous and dynamic features of the speech spectrum. This technique is shown to be highly effective in speaker-independent speech recognition. Spoken utterances are represented by time sequences of cepstrum coefficients and energy. Regression coefficients for these time functions are extracted for every frame over an approximately 50 ms period. Time functions of regression coefficients extracted for cepstrum and energy are combined with time functions of the original cepstrum coefficients, and used with a staggered array DP matching algorithm to compare multiple templates and input speech. Speaker-independent isolated word recognition experiments using a vocabulary of 100 Japanese city names indicate that a recognition error rate of 2.4 percent can be obtained with this method. Using only the original cepstrum coefficients the error rate is 6.2 percent.**

I. Introduction

DYNAMIC spectral features (spectral transition) as well as instantaneous spectral features are believed to play an important role in human speech perception [1]. Our recent perceptual experiment using isolated syllables truncated at the initial or final end has demonstrated that the portion of the utterance where the spectral variation is locally maximum bears the most important phonetic information in all syllables [2], [3].

However, there have been only a few attempts to use dynamic spectral features directly in speech recognition. In a previous experiment in speaker verification using the first and second polynomial expansion coefficients extracted from each time sequence of cepstrum coefficients, the effectiveness of these dynamic features was demonstrated [4], [5]. The polynomial coefficients were extracted every 10 ms over 90 ms periods (nine frames).

This paper describes a new method for multitemplate speaker-independent word recognition using dynamic spectral features. In principle, it is a modification of the previous method which was applied to speaker verification. The differences between the present word recognition method and the previous speaker verification method lie in the features selected for analysis and in the length of the period for extracting the dynamic features.

II. System Operation

A. Speech Analysis

A block diagram indicating the principal operations of a word recognition system using the new techniques based on the dynamic features is presented in Fig. 1 [6]. The speech wave, passed through a low-pass filter whose cutoff frequency is 4 kHz, is sampled at 8 kHz. The digitized speech is then scanned forward from the beginning of the recording interval and backward from the end to determine the beginning and end of the actual utterance. Also, endpoint detection is incorporated in the unconstrained endpoint DP matching algorithm applied at a later stage. It is therefore desirable that short silent intervals, that is, background noise intervals, be added to both the beginning and end of the actual utterance rather than strictly perform the endpoint detection. Here, the endpoint detection is accomplished by means of an energy calculation. A 32 ms Hamming window is applied to the speech every 8 ms, and the first- to the tenth-order linear predictor coefficients are extracted from each of these frames by the autocorrelation method. These coefficients are transformed into cepstrum coefficients [7], and a logarithmic transformation is applied to the energy to approximate the perceptual loudness scale.

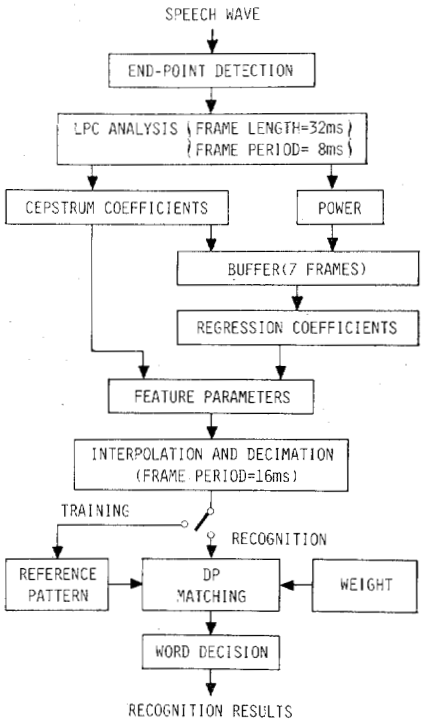

Fig. 1. Block diagram of an isolated word recognition system using dynamic spectrum features.

Manuscript received April 1, 1985; revised July 19, 1985.

The author is with the Musashino Electrical Communication Laboratory, N.T.T., 3-9-11, Midoricho, Musashino-shi, Tokyo, 180 Japan.

IEEE Log Number 8406034.

B. Regression Coefficients

Regression analysis is applied to each time function of the cepstrum coefficients and to the log-energy over several frames every 8 ms. The linear regression coefficient, namely, the first-order orthogonal polynomial coefficient, is

$$a_m(t) = \left(\sum_{n=-n_0}^{n_0} x_m(n) \cdot n\right) \Big/ \left(\sum_{n=-n_0}^{n_0} n^2\right) \qquad (1)$$

where $x_m(-n_0 \leqq n \leqq n_0)$ is the time function of the mth parameter within the segment being measured; $x_0(n)$ is the log-energy and $x_m(n)$ $(1 \leqq m \leqq 10)$ is the mth cepstrum coefficient. These coefficients represent the slope of the time function of each parameter in each segment, respectively. The length of the interval was set to seven frames (56 ms), based on the preliminary experiment described in Section V. Accordingly, n_0 is equal to 3. The 56 ms interval length seemed adequate for preserving transitional information associated with changes from one phoneme to another.

The utterance is then represented by time functions of the log-energy $x_0(t)$, cepstrum coefficients $\{x_m(t)\}_{m=1}^{10}$, and the regression coefficients $\{a_m(t)\}_{m=0}^{10}$, where t is the frame number. These time functions, excepting the log-energy $x_0(t)$ itself which is sensitive to the speech level, are used for the recognition. Since the coefficient $a_m(t)$ cannot be defined within three frame intervals at the beginning and end of speech period, these three frame intervals are eliminated from the speech period. It does not cause the elimination of the actual utterance because of the short noise interval added to the utterance period at the endpoint detection stage.

In order to reduce the number of calculations at the time registration stage, the frame interval is converted from 8 to 16 ms by averaging the time functions of adjacent frames.

C. Time Registration

A sample utterance is brought into time registration with reference templates to calculate the overall distance between them. This is accomplished by the staggered array DP matching algorithm [8], a new time warping algorithm employing a dynamic programming technique. This algorithm reduces the calculation points along the diagonal axis on the DP matching plane which is determined by a reference template and a sample utterance. This algorithm requires fewer distance calculations than conventional algorithms while realizing a complete unconstrained endpoint matching.

Fig. 2 shows the constraints for the warping function. The warping function is required to increase monotonically with a maximum slope of two and a minimum slope of $\frac{1}{2}$. The DP matching calculation is performed only at the points indicated by the symbol "⊙" which constitute every third point along the diagonal axis. This reduces the number of DP matching calculations to $\frac{1}{3}$ of that of conventional algorithms. Precision in the accumulated distance calculation is maintained by using the distance values at the neighboring points indicated by the symbol "•."

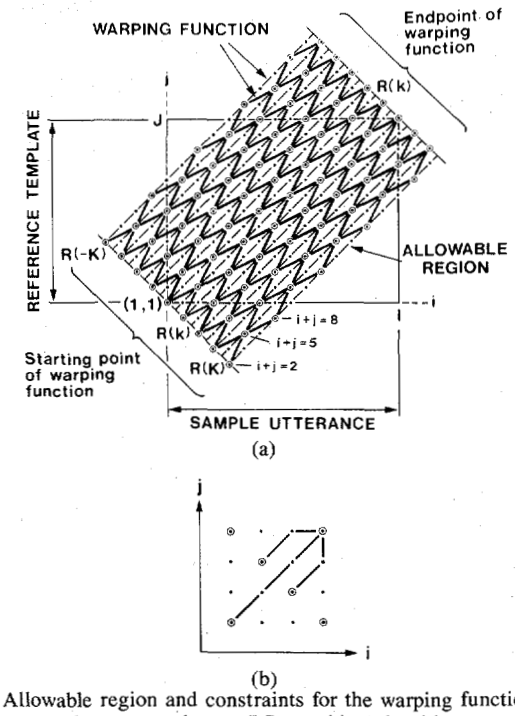

Fig. 2. Allowable region and constraints for the warping function path in the staggered array DP matching algorithm.

An actual calculation is performed at (i, j) points which satisfy the condition

$$i + j = 3l + 2 \qquad (l = 0, 1, 2, \cdots, l_{\max};$$

$$l_{\max} = \text{int}\,[(I + J - 2)/3]). \qquad (2)$$

Distance calculations are obtained for successive values of l, within the allowable region of the warping function path. Here, I and J are the lengths of the sample utterance and the reference template, respectively, and int [ ] is the integral number calculation. The intermediate distance values are stored in the register $R(k) = R(i - j)$ indicated in Fig. 2(a), where

$$R(k) = \min \begin{bmatrix} R(k-1) + d(i-1, j) + d(i, j) \\ R(k) + \frac{4}{3}\{d(i, j) + d(i-1, j-1) \\ \qquad + d(i-2, j-2)\} \\ R(k+1) + d(i, j-1) + d(i, j) \end{bmatrix}. \qquad (3)$$

Points of the distance values $d(i, j)$ used for each distance accumulation are indicated in Fig. 2(b). K, the half-width of the allowable region of the warping function, namely, the half-length of the register $R(k)$, is set to

$$K = \text{int}\,[\{\min(I, J)\}/4 + 3]. \qquad (4)$$

The unconstrained endpoint condition at both the beginning and end of the utterance is provided by using the spectral values before and after the actual speech period. The warping function can start from any frame of the first $R(k)$ register $[-K \leqq k \leqq K, R(0)$ is located at $(1, 1)]$, and end at any frame of the last $R(k)$ register indicated in

Fig. 2(a). The unconstrained endpoint technique is highly effective in coping with uncertainty in the location of both the initial and final frames due to breath noise, etc. The overall distance accumulated over the optimum warping function is obtained by

$$D(F) = \{ \min_{\substack{(i,j)\in i+j=I+J \\ k=i-j}} R(k)\}/(I+J). \quad (5)$$

D. Distance Measure

The distance measure $d(i, j)$ is defined as

$$d(i,j) = \Bigg\{ w_1 \sum_{m=1}^{10} (x_m^R(i) - x_m^I(j))^2 + w_2(a_0^R(i) - a_0^I(j))^2 + w_3 \sum_{m=1}^{10} (a_m^R(i) - a_m^I(j))^2 \Bigg\} \Bigg/ \left(\sum_{r=1}^{3} w_r \right), \quad (6)$$

where R and I indicate the reference template and sample utterance, respectively. The weighting factors $\{w_r\}_{r=1}^3$ are set *a priori* based on the effectiveness of each parameter set indicated by preliminary experiments.

The overall distance obtained using the DP matching between the sample utterance and each reference template is transferred to the word decision stage. The recognized utterance is then selected to be the word whose reference template has a smaller distance than any of the other reference templates.

III. Sample Utterances

In order to evaluate the effectiveness of the new recognition techniques in speaker-independent conditions, a vocabulary of 100 Japanese city names shown in Table I was selected. Two kinds of utterance sets used in the recognition experiments were uttered in a computer room whose background noise level was about 70 dB(A). They were recorded through a dynamic microphone.

1) Utterance Set 1: This utterance set consisted of the 100 words uttered twice each by four male speakers. These speakers, considered to represent the individual range of male voices, were selected from 30 male speakers. The selection was based on clustering results obtained in the course of a speech recognition experiment using the SPLIT method [9], [10]. In the SPLIT experiment, the utterances of these four speakers were most frequently used to construct multiple word templates.

In the recognition experiments of this paper using utterance set 1, the second utterances from each speaker were recognized using the first utterances from each of the four speakers as reference templates. That is, comparisons for this utterance set were always between test utterances and single templates. The number of reference-test speaker combinations was 16, in which four combinations were intraspeaker conditions and 12 were interspeaker conditions. The total number of test utterances were 400 and 1200 in the former and the latter conditions, respectively.

TABLE I
100 Japanese City Names in the Vocabulary

1. SaQporo	26. Kawaguchi	51. Numazu	76. Akashi
2. Hakodate	27. Urawa	52. Shimizu	77. Nishinomiya
3. Asahikawa	28. Omiya	53. Fuji	78. Kakogawa
4. Kushiro	29. Tokorozawa	54. Nagoya	79. Nara
5. Aomori	30. Koshigaya	55. Toyohashi	80. Wakayama
6. Hachinohe	31. Chiba	56. Okazaki	81. Okayama
7. Morioka	32. Ichikawa	57. Ichinomiya	82. Kurashiki
8. Sendai	33. Funabashi	58. Kasugai	83. Hiroshima
9. Akita	34. Matsudo	59. Toyota	84. Kure
10. Yamagata	35. Kashiwa	60. YoQkaichi	85. Fukuyama
11. Fukushima	36. Ichihara	61. Otsu	86. Shimonoseki
12. Koriyama	37. Tokyo	62. Kyoto	87. Tokushima
13. Iwaki	38. Hachioji	63. Osaka	88. Takamatsu
14. Niigata	39. Fuchu	64. Sakai	89. Matsuyama
15. Toyama	40. Machida	65. Toyonaka	90. Kochi
16. Kanazawa	41. Yokohama	66. Higashiosaka	91. Kitakyushu
17. Fukui	42. Kawasaki	67. Suita	92. Fukuoka
18. Nagano	43. Yokosuka	68. Takatsuki	93. Kurume
19. Matsumoto	44. Hiratsuka	69. Hirakata	94. Nagasaki
20. Mito	45. Fujisawa	70. Ibaraki	95. Saseho
21. Hitachi	46. Sagamihara	71. Yao	96. Kumamoto
22. Utsunomiya	47. Kofu	72. Neyagawa	97. Oita
23. Maebashi	48. Gifu	73. Kobe	98. Miyazaki
24. Takasaki	49. Shizuoka	74. Himeji	99. Kagoshima
25. Kawagoe	50. Hamamatsu	75. Amagasaki	100. Naha

2) Utterance Set 2: The first utterances from all four speakers of utterance set 1 were stored as multiple templates, and utterances from 20 male speakers, who were different from the four speakers, were used as test utterances. That is, the comparisons for this utterance set were with four templates. The total number of test utterances was 2000, where one utterance from each speaker was tested for each word.

IV. Differential Spectrum

The cepstrum is defined as the inverse Fourier transformation of the log spectrum. The Fourier transformation of regression coefficients derived from lower order cepstrum coefficients, therefore, produces differential log-spectrum envelopes, which represent log-spectrum envelope transitions in unit time periods.

Fig. 3 illustrates the time sequences of log-energy, the log-spectrum envelope, and the differential log-spectrum envelope for the Japanese word /saQporo/ uttered by two male speakers. The differential envelopes explicitly represent the dynamics of the spectrum envelopes. For example, the dynamic characteristics of the /r/ sound, which are difficult to observe in conventional spectrum envelope sequences, are clearly shown in this figure.

Although the regression coefficient for the energy contour is not shown in this figure, it is reasonable to assume that its time function is effective in representing the speaker-independent macroscopic characteristic or broad outline for each word. Additionally, it has an advantage over the energy contour itself. Since the energy regression coefficient is independent of the absolute speech level, it is not necessary to normalize it by maximum and minimum speech levels which can be obtained only after the end of an utterance. This calculation causes a decision

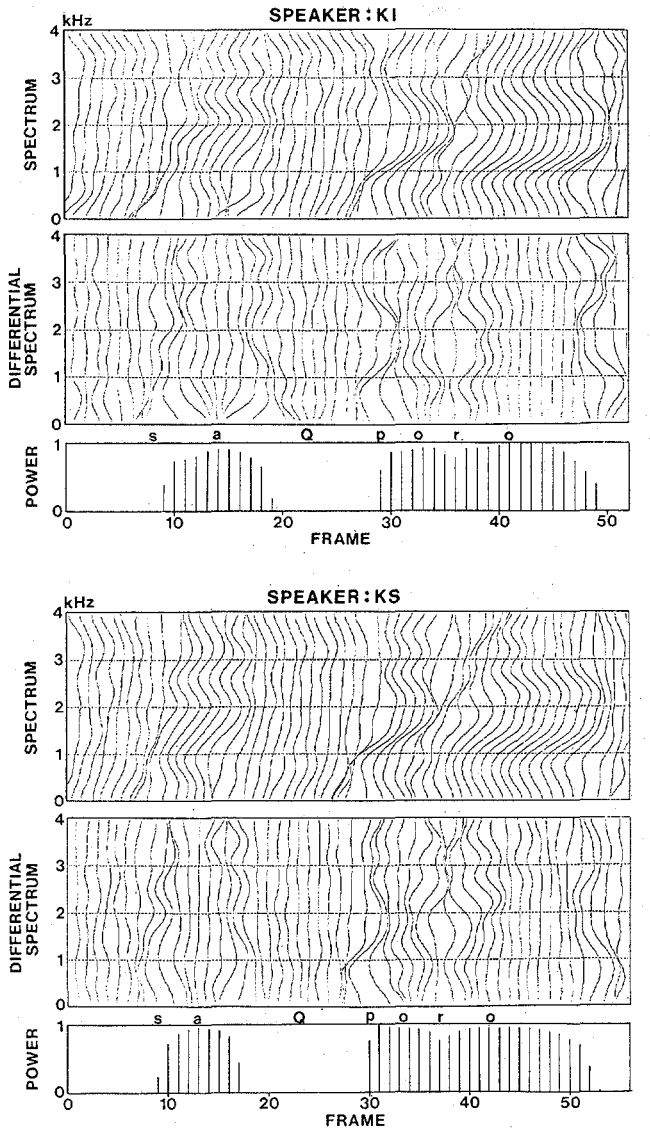

Fig. 3. Time sequences of spectrum envelope, differential spectrum envelope and energy for a word /saQporo/ uttered by two male speakers.

delay in conventional methods [11], [12]. Strictly speaking, the method of this paper requires a three-frame (24 ms) delay for calculating the regression coefficients. However, this delay is negligible.

V. Effectiveness of Cepstral Regression Coefficients

A. Recognition by Regression Coefficients Only (Utterance Set 1)

As a preliminary experiment, the effectiveness of regression coefficients for cepstrum sequences in word recognition was tested by varying the number of frames for extracting the regression coefficients between 3 and 11 frames. Cepstrum coefficients and energy regression coefficients were not used in this experiment. This condition corresponds to $w_1 = w_2 = 0$ and $w_3 = 1$ in (6). Plots of recognition results using utterance set 1 are shown in Fig. 4. Recognition error rates for interspeaker and intraspeaker conditions are shown separately. These results indicate that the three- and five-frame interval lengths (24–40 ms) are too short to derive effective regression coefficients, whereas the nine-frame length (72 ms) is optimum.

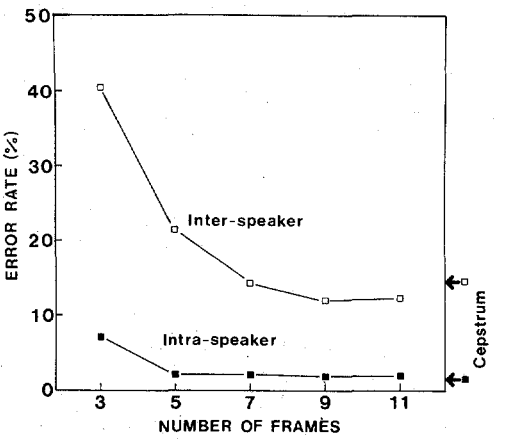

Fig. 4. Relation between number of cepstrum frames used for regression analysis and recognition error rate obtained with utterance set 1. On the outside right are shown the error rates for the recognition using only cepstrum coefficients.

The error rates for the recognition experiment using only the instantaneous cepstrum coefficients ($w_1 = 1$ and $w_2 = w_3 = 0$) are shown on the outside right of Fig. 4. Comparison of results using instantaneous cepstrum and cepstral regression coefficients indicates that regression coefficients extracted from nine-frame intervals are slightly more efficient than the instantaneous cepstrum coefficients. The regression coefficients extracted from seven-frame intervals are almost equal in efficiency to the instantaneous cepstrum coefficients. A part of each error rate, that is, 1.0 percent of the interspeaker error rate and 0.5 percent of the intraspeaker error rate, is attributable to the excessive length discrepancies between test utterances and reference templates, which make DP matching ineffective.

B. Combination of Cepstrum and Regression Coefficients (Utterance Set 1)

Fig. 5 presents the recognition results for the combination of cepstrum coefficients and their regression coefficients. In this experiment, the value of the weighting factor for the cepstral regression coefficients w_3 was varied, whereas the other weighting factors w_1 and w_2 were set to 1 and 0, respectively. This figure shows the results for two regression analysis interval length conditions of seven and nine frames.

These results indicate the following important features.

1) The results for the two regression analysis interval length conditions are almost the same when the cepstrum coefficients and their regression coefficients are used in combination.

2) For interspeaker recognition, using appropriately combined cepstrum and regression coefficients extracted from seven-frame intervals, a mean recognition error rate of 8.2 percent can be obtained. This error rate is $\frac{3}{5}$ of that obtained using the cepstrum (14.5 percent) or regression coefficients only (14.2 percent).

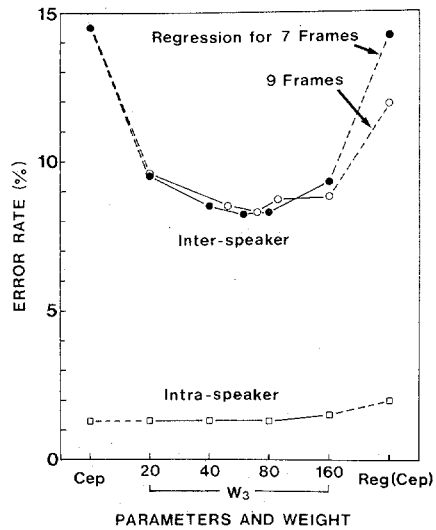

Fig. 5. Error rates for various feature parameter conditions. Left: cepstrum (Cep); right: regression coefficient for cepstrum [Reg (Cep)]; middle: their combination, weighting factor being varied. Reg (Cep) is extracted from 7 or 9 frames, and dimensions of Cep and Reg (Cep) are both fixed at 10.

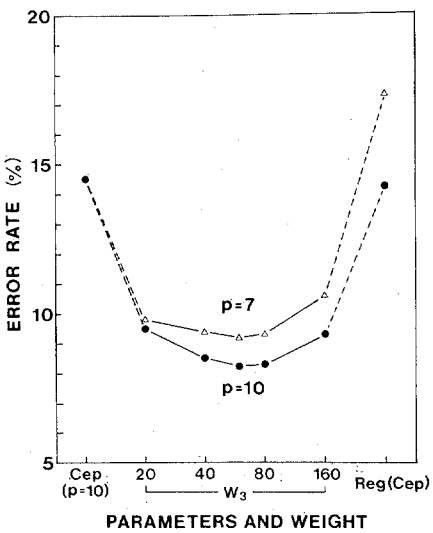

Fig. 6. Effects of the dimension of regression coefficients for cepstrum (p) on the word recognition error rates under the interspeaker matching condition.

3) Fluctuation of error rate as a function of the weighting factor w_3 is small and the optimum condition can be realized for a wide range of values of w_3.

4) Combining regression and cepstrum coefficients provides no improvement in intraspeaker recognition. The error rate using cepstrum only is already very small.

Fig. 6 indicates the results of a supplementary interspeaker recognition experiment where regression coefficients were extracted only for the first- to seventh-order cepstrum coefficients. This extraction ensures that dynamic features are restricted to relatively broad spectral characteristics. Cepstrum coefficients were analyzed up to the 10th order as in the above experiments. The results demonstrate that including regression coefficients up to 10th order, relating to the relatively fine spectral characteristics, is important to preserve phonetic information and maintain performance.

Based on these results, it was decided to apply the regression analysis to seven-frame intervals for all ten cepstrum coefficients. The optimum condition of the weighting factors for the cepstrum regression coefficients ($w_1 = 1$ and $w_3 = 60$) approximately satisfies the following equation:

$$w_1 \cdot E\left[\sum_{m=1}^{10} (x_m^R - x_m^I)^2\right] \simeq w_3 \cdot E\left[\sum_{m=1}^{10} (a_m^R - a_m^I)^2\right] \tag{7}$$

where $E[\ ]$ represents the long-term average. This equation means that the weighted distances obtained from the cepstrum and their regression coefficients are comparable.

VI. Effectiveness of Energy Regression Coefficient

The effectiveness of the regression coefficient extracted from the energy contour was tested by varying the weighting factor w_2, and setting the other weighting factors at their optimum values ($w_1 = 1$ and $w_3 = 60$) which were estimated previously. Recognition results obtained using utterance set 1 are plotted in Fig. 7. In the left portion of the figure, results are shown for the combination of the cepstrum coefficients and their regression coefficients ($w_1 = 1$, $w_2 = 0$, and $w_3 = 60$). In the right portion is shown the result for the combination of the cepstrum coefficients and the energy regression coefficient when w_2 is set at the optimum value ($w_1 = 1$, $w_2 = 10$, and $w_3 = 0$). When the energy regression coefficient is used in combination with the cepstrum, an error rate of 10.7 percent is obtained. This rate is 3.8 percent lower than that obtained using only the cepstrum. However, this improvement is smaller than that obtained using the combination of the cepstral regression coefficients with the instantaneous cepstrum coefficients (6.3 percent improvement). A combination of the energy regression coefficient with the latter combination produces a 0.4 percent improvement in error rate.

VII. Multitemplate Speaker-Independent Word Recognition

A. Various Combinations of Parameters (Utterance Set 2)

Recognition experiments using utterance set 1 clarified the optimum weighting factor values $\{w_r\}_{r=1}^3$ for combining the cepstrum coefficients and the regression coefficients derived from cepstrum and energy contours. Based on these results, recognition experiments using utterance set 2 consisting of 20 male speakers' voices were carried out setting the weighting factors at the optimum values ($w_1 = 1$, $w_2 = 10$, and $w_3 = 60$). Four utterances from each of the male speakers of utterance set 1, different from the test utterance speakers, were stored as multiple templates for each word.

Fig. 8 compares mean recognition error rates obtained by varying the combination of parameter sets used for the

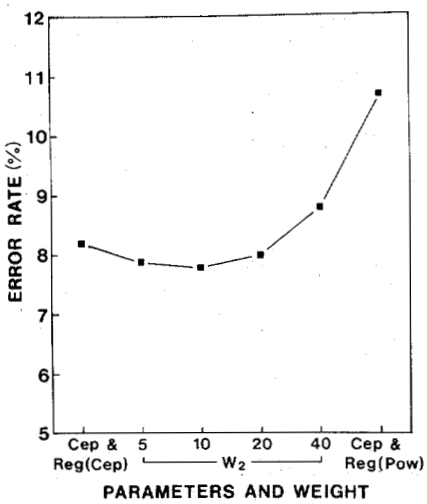

Fig. 7. Error rates for various feature parameter conditions. Left: cepstrum (Cep) and their regression coefficient [Reg(Cep)]; right: Cep and regression coefficient for energy [Reg(Pow)]; middle: combination of Cep, Reg(Cep) and Reg(Pow), weighting factor w_2 being varied (w_1 = 1 and w_3 = 60).

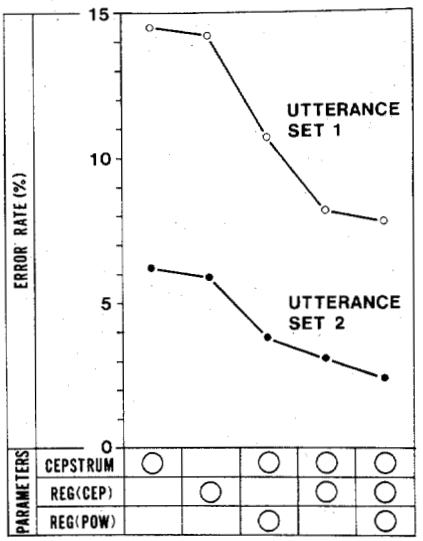

Fig. 8. Recognition results for five feature parameter combination conditions using utterance sets 1 and 2; ○ indicates the parameter set used.

recognition. A recognition experiment, using only the energy regression coefficient, was not attempted since it produces very large error rates. This figure includes the interspeaker recognition results obtained with utterance set 1. The results for the utterance set 2 are similar to those for the utterance set 1. The cepstral regression coefficients are slightly more efficient than the cepstrum coefficients, and the combination of these two kinds of parameter sets reduces the error rate from 6.2 percent (cepstrum) or 5.9 percent (cepstrum regression) to 3.1 percent. This rate is half of the error rate obtained by either one of the parameter sets.

The error rate obtained when the energy regression coefficient is combined with the cepstrum is 3.8 percent. This means that the cepstral regression coefficients are more efficient than the energy regression coefficient in combination with the cepstrum coefficients. This is similar to the result obtained for utterance set 1. The error rate obtained using the combination of the cepstrum, cepstral regression, and energy regression coefficient is 2.4 percent, which is $\frac{2}{5}$ of the error rate obtained using cepstrum coefficients only.

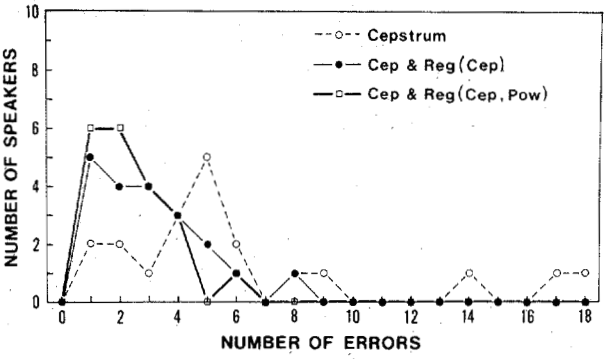

Fig. 9. Distribution of the number of errors out of 100 test utterances for 20 speakers.

B. Analysis of Error Rate Improvement

The distribution of the number of errors among speakers is shown in Fig. 9 for the three experimental conditions, namely, recognition by the cepstrum only; the combination of cepstrum and their regression coefficients; and the combination of cepstum, cepstral regression, and energy regression coefficients. Although there are three speakers in the first case for whom 14, 17, or 18 utterances are incorrectly recognized, the maximum number of errors is reduced to eight when cepstral regression coefficients are combined wih the cepstrum coefficients. Also, the addition of the energy regression coefficient reduces the errors even further to six. This means that the combination of instantaneous and dynamic spectral features is highly effective in reducing the so-called "sheep and goats phenomenon" [13].

Table II compares the distribution of major confusible word pairs for two feature parameter conditions, cepstrum only, and the combination of cepstrum, cepstral regression, and energy regression coefficients. Confusible word pairs occurring twice or more in all test utterances by 20 speakers are presented. Those which occur for both parameter conditions are given in the upper part of the table. Although the results using only the instantaneous features include several confusible word pairs which are unlikely to occur in human speech perception, the combination of dynamic and instantaneous features removes these unreasonable confusible word pairs. Detailed analysis of the confusions that occur indicates that the combination of instantaneous and dynamic features reduces serious DP matching errors such as phoneme insertion and deletion, and mapping between vowel and consonant. This outcome can be attributed to the fact that the transitional and steady parts of the speech spectrum are explicitly characterized by the regression coefficients.

VIII. Conclusion

A new speaker-independent spoken word recognition method which uses a combination of instantaneous and

TABLE II
COMPARISON OF THE FREQUENCY OF MAJOR CONFUSIBLE WORD PAIRS FOR TWO FEATURE PARAMETER CONDITIONS. *N*1: RECOGNITION USING INSTANTANEOUS FEATURES, AND *N*2: RECOGNITION USING THE COMBINATION OF INSTANTANEOUS AND DYNAMIC FEATURES *N*3 = *N*2 − *N*1. [ ]: DEVOCALIZED

WORD PAIRS	FREQUENCY		
	N1	N2	N3
Kawasa[ki] - Amagasa[ki]	6	10	+4
Nara - Naha	5	4	-1
Ichinomiya - Nishinomiya	8	3	-5
Nara - Urawa	2	3	+1
[To]kushima - [Fu]kushima	2	3	+1
Wakayama - Okayama	3	2	-1
[Fu]kuyama - Toyama	2	2	0
Ko[chi] - O[tsu]	2	2	0
[Fu]chu - [Ha]chioji	0	2	+2
Takasa[ki] - Nagasa[ki]	9	0	-9
Matsudo - Saseho	5	0	-5
Yamagata - Hirakata	3	0	-3
Kanagawa - Nagano	2	0	-2
[Fu]kui - Fuji	2	0	-2
I[chi]kawa - Fujisawa	2	0	-2
I[chi]kawa - I[chi]hara	2	0	-2
Funaba[shi] - Maeba[shi]	2	0	-2
Yokohama - [To]korozawa	2	0	-2
Hamama[tsu] - Takama[tsu]	2	0	-2
Toyoha[shi] - Takasa[ki]	2	0	-2
Aka[shi] - Taka[tsu]ki	2	0	-2
Kura[shi]ki - Taka[tsu]ki	2	0	-2
Shimonose[ki] - Naha	2	0	-2
OTHERS	55	17	
TOTAL	124	48	

dynamic spectral features has been proposed. Speech signals are analyzed by the LPC method every 8 ms, and converted into cepstrum and log-energy time functions. Regression analysis is applied to each time function over 56 ms intervals (seven frames) every 8 ms to extract dynamic spectral features. The time functions of regression coefficients extracted for the cepstrum and energy contours are used for recognition together with the cepstrum sequences. Following $\frac{1}{2}$-frame rate decimation, the time functions of the feature parameter set which consist of the cepstrum and the regression coefficients for cepstrum and energy are used for the recognition. The time functions of test utterance and reference templates are brought into time registration. This is accomplished by the staggered array DP matching algorithm.

Speaker-independent word recognition experiments using a vocabulary of 100 Japanese city names were carried out to evaluate the new method. When word utterances spoken by four male speakers representing the individual voice range were stored as multiple reference templates for each word, and utterances by another 20 male speakers were used as input speech, the recognition error rates using cepstrum coefficients, a combination of cepstrum and their regression coefficients, and a combination of cepstrum and regression coefficients for cepstrum and energy contours were 6.2, 3.1, and 2.4 percent, respectively. These results demonstrate the effectiveness of the new recognition method. The combination of temporal and dynamic features is effective in reducing the frequency of large individual speaker error rates and in reducing unlikely confusions from the perspective of human speech perception. These results are consistent with the knowledge concerning speech perception obtained from recent experiments indicating that dynamic spectral features play an important role in phoneme perception [3]. Some discussions on new techniques are also included in this paper.

In this recognition method, the frame interval for speech analysis is set to 8 ms, and that for DP matching is set to 16 ms, in order to obtain stable regression coefficients and to reduce the number of distance calculations. A recognition experiment using regression coefficients extracted through analysis with 16 ms frame interval still remains to be done.

Further investigations, current or projected, include a large-scale evaluation. For this purpose, it is desirable that this method be applied to the SPLIT word recognition system to produce pseudophoneme templates based on distances calculated by cepstrum and regression coefficients. The dynamic features represented by the regression coefficients have the advantage of being robust with respect to frequency response distortions introduced by transmission systems [5]. The differential spectrum envelope presented in Fig. 2 will become a useful method for observing the dynamic spectral characteristics.

Elenius *et al.* [14] have reported that adding the time derivative of critical band cepstral coefficients to the word patterns improves recognition performance. Comparison of effectiveness of the time derivative to the regression coefficients is currently being undertaken.

The optimum length of the speech segment for extracting the regression coefficients in this method is almost one-half of that adopted in the speaker verification system using short English sentences [5]. The variation might be caused by the difference in acoustic features used in word recognition and speaker verification (prosodic or phonetic) or by the difference of languages. This also remains to be investigated.

ACKNOWLEDGMENT

The author wishes to thank K. Kakehi, Head of Communication Principles Research Section 4, and Dr. M. Kohda for their guidance and stimulating discussions. The author also wishes to thank Dr. A. E. Rosenberg at AT&T Bell Laboratories and Dr. Y. Tohkura for the revision of this paper. The author also acknowledges Dr. K. Shikano for providing the DP matching program.

REFERENCES

[1] G. Ruske, "Auditory perception and its application to computer analysis of speech," in *Computer Analysis and Perception, Vol. II, Auditory Signals*, C. Y. Suen and R. De Mori, Eds. Boca Raton, FL: CRC Press, 1982.

[2] S. Furui, "On the role of dynamic characteristics of speech spectra for syllable perception," *Trans. Fall Meet. Acoust. Soc. Japan*, vol. 1-1-12, Oct. 1984.

[3] ——, "On the role of spectral transition for speech perception," *Trans. Tech. Group Hearing Acoust. Soc. Japan*, vol. H85-6, Jan. 1985.

[4] S. Furui and A. E. Rosenberg, "Experimental studies in a new automatic speaker verification system using telephone speech,"in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Denver, CO, Apr. 1980, pp. 1060–1062.
[5] S. Furui, "Cepstrum analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 254–272, Apr. 1981.
[6] ——, "Spoken word recognition using dynamic features of speech spectrum," *Trans. Tech. Group Speech Acoust. Soc. Japan*, vol. S84-65, Dec. 1984.
[7] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, June 1974.
[8] K. Shikano and K. Aikawa, "Staggered array DP matching," *Trans. Tech. Group Speech Acoust. Soc. Japan*, vol. S82-15, June 1982.
[9] N. Sugamura and S. Furui, "Large vocabulary word recognition using pseudo-phoneme templates," *Trans. Inst. Electron., Commun. Eng. Japan*, vol. J65-D, pp. 1041–1048, Aug. 1982
[10] N. Sugamura, K. Shikano, and S. Furui, "Isolated word recognition using phoneme-like templates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Boston, MA, Apr. 1983, pp. 723–726.
[11] K. Aikawa, K. Shikano, and S. Furui, "An isolated word recognition method using power-weighted spectral matching measure," *Trans. Tech. Group Speech Acoust. Soc. Japan*, vol. S81-59, Dec. 1981.
[12] M. K. Brown and L. R. Rabiner, "On the use of energy in LPC-based recognition of isolated words," *Bell Syst. Tech. J.*, vol. 61, pp. 2971–2987, Dec. 1982.
[13] G. Doddington, "Voice authentication gets the go-ahead for security systems," *Speech Technol.*, vol. 2, pp. 14–23, Sept./Oct. 1983.
[14] K. Elenius and M. Blomberg, "Effects of emphasizing transitional or stationary parts of the speech signal in a discrete utterance recognition system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Paris, France, May 1982, pp. 535–538.

Sadaoki Furui (M'79) was born in Tokyo, Japan, on September 9, 1945. He received the B.S., M.S., and Ph.D. degrees in mathematical engineering and instrumentation physics from Tokyo University, Tokyo, Japan, in 1968, 1970, and 1978, respectively.

After joining the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation, in 1970, he studied the analysis of speaker characterizing information in the speech wave, its application to speaker recognition and interspeaker normalization in speech recognition, and the vector-quantization-based speech recognition algorithm. He is currently a Senior Researcher at Musashino Electrical Communication Laboratory, working on research in speech perception and speech recognition. From December 1978 to December 1979 he was with the Staff of the Acoustics Research Department at Bell Laboratories, Murray Hill, NJ, as an exchange visitor working on speaker verification.

Dr. Furui is a member of the Acoustical Society of Japan and the Institute of Electronics and Communication Engineers of Japan.