論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Speaker-Indepedent Isolated Word Recognition Based on Dynamics-Emphasized Cepstrum
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	, Vol. 69, No. 12, pp. 1310-1317
Citation(English)	Trans. IECE Jpn, Vol. 69, No. 12, pp. 1310-1317
発行日 / Pub. date	1986, 12
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 1986 Institute of Electronics, Information and Communication Engineers.

THE TRANSACTIONS OF THE IECE OF JAPAN, VOL. E 69, NO. 12 DECEMBER 1986

PAPER

Speaker-Independent Isolated Word Recognition Based on Dynamics-Emphasized Cepstrum

Sadaoki FURUI†, *Member*

SUMMARY A new analysis technique applicable to speech recognition is proposed considering the auditory mechanism of speech perception which emphasizes spectral dynamics as well as compensates for the spectral undershoot associated with coarticulation. A speech wave is represented by the LPC cepstrum and logarithmic energy sequences, and the time sequences over short periods are expanded by the first- and second-order polynomial functions at every frame period. The dynamics of the cepstrum sequences are then emphasized by the linear combination of their polynomial expansion coefficients, that is, derivatives, and their instantaneous values. Speaker-independent word recognition experiments using time functions of the dynamics-emphasized cepstrum and the polynomial coefficient for energy indicate that the error rate can be largely reduced by this method. The experimental results are compared with those obtained by the previous method in which the polynomial coefficients for the cepstrum and energy time functions were used in combination with the original time functions of these parameters as independent parameters.

1. Introduction

The sounds that occur in continuous speech are characterized by time-varying spectral patterns which have almost no steady-state period. The spectral pattern of each phoneme is largely modified as a consequence of the coarticulation with the adjacent phonemes. Additionally, the spectrum starts changing into the succeeding phone spectrum before attaining the ideal spectrum (target) of the phoneme or the syllable which can be realized only when it is uttered in isolation. In other words, the spectrum of each phoneme in continuous speech undershoots the target. This is sometimes referred to as neutralization or reduction. The spectral modification is so large that when a short-length speech wave is presented, it sometimes sounds like a completely different phoneme. However, there is the perceptual tendency of phonemes to remain constant, and, as a result, continuous speech sounds as if the spectra of isolated phonemes or syllables were concatenated regardless of the modification.

This so-called categorical perception mechanism is thought to be realized by the combination of two separate mechanisms. One is the fundamental context-independent human hearing mechanism of spectral compensation or prediction, in which the target spectrum is perceived by overshooting the spectral dynamics[1]-[3]. The other is the context-dependent hearing mechanism of the contrast effect.

An auditory masking experiment using frequency-modulated (FM) sounds has demonstrated that the masking pattern shows considerable overshoot at the spectral change, which indicates high auditory sensitivity for transitional sounds[4]. The author has made clear that the most important phonetic information exists at the position of maximum spectral change through the investigation of the relationships between the results of identification tests using syllables modified in various ways and the features of spectral dynamics[5]. Physiological studies have pointed out that the higher the level of the ascending auditory nervous system is, the more the neurons respond to the transitional stimuli compared with those responding to steady-state stimuli. Even for the steady-state-type neurons, the greater part of them feature adaptation characteristics which produce high-frequency pulses at the onset of stimuli, with the number of pulses immediately decreasing as a function of time[6]. These facts suggest that the dynamic part of the spectrum is perceived after being emphasized in the auditory system.

A model for the speech perception mechanism which compensates for the spectral undershoot based on its dynamic features has already been proposed. In the model, the first- and second-derivatives as well as the instantaneous values for formant transition were integrated using time-axis weighting[7]. The propriety of this model was confirmed by experimentally comparing the hearing response to the concatenations of two vowels and sustained single vowels. In the experiment, the formant frequencies of the sustained vowels, which were perceptually equal to those of the transitional second vowels in the concatenations, were measured[8]. Another hearing experiment indicated that the perceptually equal quality corresponds to the spectral pattern which is produced by overshooting the dynamics of the formant frequencies to the extent of the DL (differential limen) of sound quality[9].

If speech is in fact perceived by humans based on the above-mentioned mechanism, it can be expected that the performance of an automatic speech recognition

Manuscript received June 12, 1986.
Manuscript revised August 30, 1986.
†The author is with NTT Electrical Communications Laboratories, Musashino-shi, 180 Japan.

algorithm will be improved by introducing a speech analysis method which incorporates this mechanism. Although an investigation from this point of view has indicated the improvement of vowel separability by emphasizing formant transition[10], no effective method has yet been proposed for improving recognition performance which can be applied continuous speech including consonants.

This paper proposes a method for compensating spectral undershoot by emphasizing the spectral dynamics using polynomial expansion coefficients (regression coefficients) for the cepstrum time sequences, and indicates its effectiveness in speaker-independent spoken word recognition. The polynomial coefficients for the cepstrum and energy time functions extracted from every short period were previously used by the author in combination with the original time functions of these parameters as independent parameters. The effectiveness of this method was ascertained for largely reducing recognition errors in speaker verification[11] and in speech recognition[12]. The difference between the present and previous methods lies in the technique of combining dynamic and instantaneous features.

2. Spectral Analysis

The speech spectrum of every short period is represented by the 1st through the 10th linear predictive coefficients as well as the energy both of which are extracted through LPC (linear predictive coding) analysis. Prior to the LPC analysis, the speech wave is band-limited at 4 kHz, sampled at 8 kHz, and windowed by a Hamming window of 32 ms-long at every 8 ms. The linear predictive coefficients and the energy are transformed into LPC cepstrum and logarithmic energy respectively, considering the auditory characteristics. LPC cepstrum coefficients can be directly related to a peak-weighted, smoothed spectral envelope via Fourier transformation. Time functions of the LPC cepstrum coefficients and the logarithmic energy (these will simply be called cepstrum coefficients and energy hereafter) over 56 ms (seven frames) intervals are extracted every 8 ms, and the dynamic characteristics indicated over the interval are analyzed. Based on the analysis results, cepstrum coefficients at the center of the interval are modified using the method which will be described in the next section.

The length of the interval (number of frames) for dynamic characteristic analysis was decided upon based on the spoken word recognition experiment using the combination of cepstrum coefficients and energy time functions and their polynomial expansion coefficients, as described in the Introduction. The optimum length of seven frames corresponds to the perceptual instant length[13] and to the minimum length for syllable perception [14]. Additionally, this length seems adequate for preserving the transitional information associated with the changes occurring from one phoneme to another.

3. Emphasis of Spectral Transition Using Cepstrum Polynomial Expansion Coefficients

Ten-dimensional cepstrum vectors, each of which consists of the 1st through the 10th cepstrum coefficients, for the adjacent seven frames are denoted by $\boldsymbol{C}_j(j=1, \cdots, 7)$, and the first- and second-order polynomial expansion coefficients for their time functions are denoted by $\boldsymbol{C}'$ and $\boldsymbol{C}''$ respectively. In this paper, the following orthogonal polynomial representations are used:

$$\begin{aligned} P_{0j}&=1, \\ P_{1j}&=j-4, \\ P_{2j}&=j^2-8j+12. \end{aligned} \tag{1}$$

$\boldsymbol{C}'$ and $\boldsymbol{C}''$ can then be represented by

$$\begin{aligned} \boldsymbol{C}'&=\sum_{j=1}^{7}\boldsymbol{C}_jP_{1j}/\sum_{j=1}^{7}P_{1j}^2, \\ \boldsymbol{C}''&=\sum_{j=1}^{7}\boldsymbol{C}_jP_{2j}/\sum_{j=1}^{7}P_{2j}^2. \end{aligned} \tag{2}$$

These representations correspond to the slope and to the curvature respectively, that is, to the first and second derivative coefficients averaged over the interval, for the cepstrum time function. Therefore,

$$\boldsymbol{C}'\simeq\frac{d}{dt}\boldsymbol{C}, \quad \text{and} \quad \boldsymbol{C}''\simeq\frac{d^2}{dt^2}\boldsymbol{C}. \tag{3}$$

When parameters k_1 and k_2 ($\geqq 0$) in the following equation are set to the appropriate values, the time function of $\tilde{\boldsymbol{C}}$ can be obtained in which the dynamic characteristics included in the time functions of $\boldsymbol{C}$ are emphasized:

$$\tilde{\boldsymbol{C}}=\boldsymbol{C}+k_1\boldsymbol{C}'-k_2\boldsymbol{C}''. \tag{4}$$

If $S(\omega, t)$ and $C_n(t)$ are the power spectral envelope and the n-th order cepstrum coefficient at time t,

$$\log S(\omega, t)=\sum_{n=-10}^{10}C_n(t)\mathrm{e}^{-jn\omega}, \tag{5}$$

$$\frac{d}{dt}\log S(\omega, t)=\sum_{n=-10}^{10}\frac{d}{dt}C_n(t)\mathrm{e}^{-jn\omega}. \tag{6}$$

This means that the Fourier transformation of the first derivative for the cepstrum coefficient corresponds to the first derivative for the logarithmic spectral envelope. A similar relationship is obtained for the second derivatives. Therefore, the logarithmic spectral envelope which corresponds to the dynamics-emphasized cepstrum can be written as follows:

$$\log \tilde{S}(\omega, t)=\sum_{n=-10}^{10}\tilde{C}_n(t)\mathrm{e}^{-jn\omega}$$

$$= \sum_{n=-10}^{10} (C_n(t) + k_1 \frac{d}{dt} C_n(t) - k_2 \frac{d^2}{dt^2} C_n(t)) e^{-jn\omega}$$

$$= \log S(\omega, t) + k_1 \frac{d}{dt} \log S(\omega, t) - k_2 \frac{d^2}{dt^2} \log S(\omega, t). \quad (7)$$

This indicates that spectral transition is emphasized and that spectral undershoot in continuous speech can be compensated for by the method proposed in this paper provided proper values are given to the parameters k_1 and k_2.

An auditory model includng the above-mentioned process which emphasizes the spectral transition is shown in Fig. 1. An example of cepstrum time functions before and after the transition emphasizing process is given in Fig. 2. The example is presented under the condition that parameters k_1 and k_2 are set to the optimum values which were obtained through the speech recognition experiments described in the following section. An example of a short-term variation of the logarithmic spectral envelope and its modification by emphasizing the transition is shown in Fig. 3.

The above-mentioned method can be regarded as an expansion of the auditory dynamic model for formant transition, proposed by Tanaka, to the model for spectral envelope transitions utilizing the polynomial expansion coefficients for cepstrum sequences.

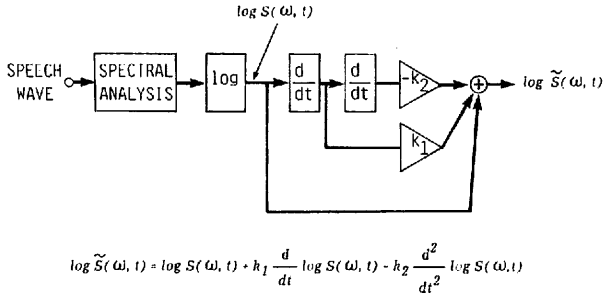

Fig. 1 Block diagram of an auditory model employing spectral dynamics-emphasis mechanism.

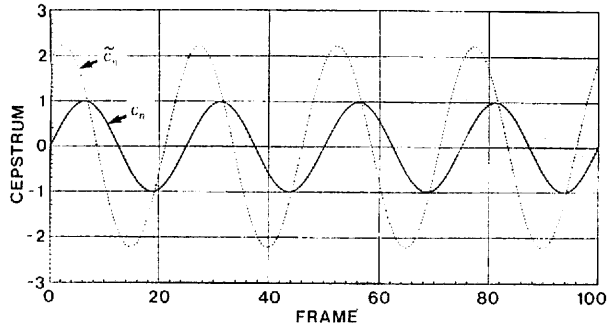

Fig. 2 Example of schematic time function of cepstrum coefficient and its modification by dynamics-emphasis.

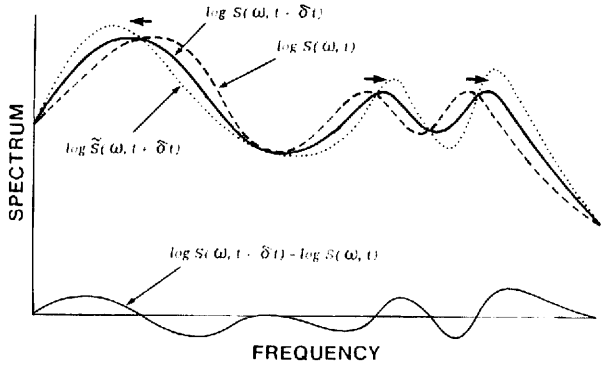

Fig. 3 Example of time variation of logarithmic spectral envelope and its modification by dynamics-emphasis.

4. Spoken Word Recognition Based on Emphasized Spectral Dynamics

4.1 Feature Extraction

A block diagram indicating the principal operations of the spoken word recognition system based on the above-mentioned model is shown in Fig. 4. The beginning and end of the actual sample utterances are determined through a short-term energy calculation. The cepstrum coefficients and energy time functions are extracted for the speech interval through LPC analysis. At the feature extraction stage, the time functions of the cepstrum coefficients modified by emphasizing the dynamics are obtained by the linear combination of cepstrum coefficients and their polynomial expansion coefficients. At the same time, the time function of the first-order polynomial coefficient is extracted for the energy, which is then directly used for the recognition without summation with the original energy value. This is because the absolute value of the energy level is

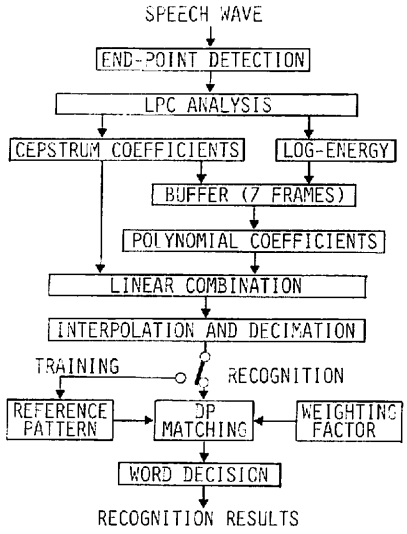

Fig. 4 Block diagram indicating principal operations of spoken word recognition system.

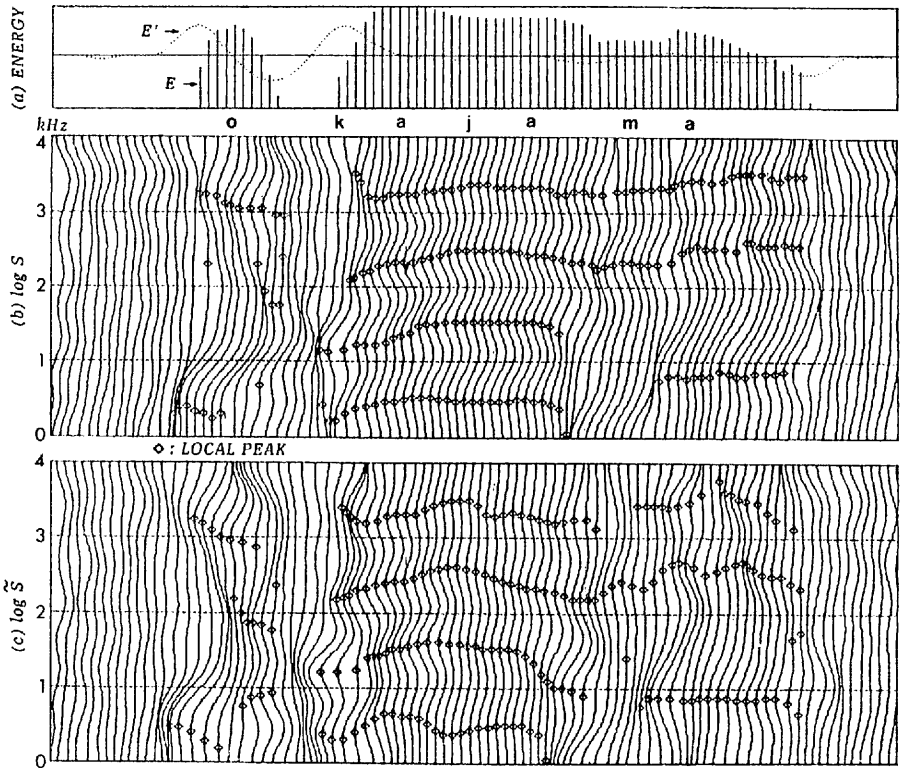

Fig. 5 Example of feature parameter extraction results for Japanese city name "Okayama" uttered by a male speaker. (a) Time function of short-time energy (E) and its 1st-order polynomial coefficient (E'), (b) time function of spectral envelope and its local peaks, and (c) time function of spectral envelope and its local peaks modified by dynamics-emphasis.

insignficant with respect to phoneme perception, and because the performance of a recognition system using the energy level is affected by speaker and speech variation.

In order to reduce the number of distance calculations at the time registration stage, the frame interval is converted from 8 ms to 16 ms by averaging the time functions of adjacent frames for both modified cepstrum and energy polynomial coefficients.

An example of the feature parameter extraction results for a Japanese city name, "Okayama", uttered by a male speaker is shown in Fig. 5. The time functions of the spectral envelope and its local peaks before and after modification by transition emphasis, as well as the first-order polynomial coefficient for energy, are indicated. These results show that transition of the spectral envelope as well as of its peaks are emphasized by the method proposed in this paper.

4.2 Word Recognition

A sample utterance represented by the parameters, which are extracted through the above-mentioned method, is brought into time registration with reference templates to calculate the overall distance between them. This is accomplished by the staggered-array DP (dynamic programming) matching algorithm, which requires fewer distance calculations than conventional algorithms while realizing complete unconstrained endpoint matching[12]. In order to cope with the speaker variability of feature parameters, multiple templates selected from the utterances by a large number of speakers are stored as references for each word during the training stage.

At the recognition stage, a sample utterance is then sequentially compared with the multiple reference templates, with the overall distance obtained using the DP matching between the sample utterance and each reference template being 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.

The distance measure, $D(k, l)$, between the k-th frame of the reference template and the l-th frame of the input speech is difined as

$$D(k, l)=w_1\sum_{n=1}^{10}(\tilde{C}_n^R(k)-\tilde{C}_n^I(l))^2 + w_2(E'^R(k)-E'^I(l))^2, \quad (8)$$

where R and I indicate the reference template and sample utterance respectively, and E' indicates the

polynomial expansion coefficient for energy.

The weighting factors, w_1 and w_2, are set a priori based on the magnitude of variation for each parameter which was calculated during a preliminary experiment.

5. Spoken Word Recognition Experiment

5.1 Sample Utterances

In order to evaluate the effectiveness of this method under speaker-independent conditions, a vocabulary of 100 Japanese city names[12] was selected. Two kinds of utterance sets used in the recognition experiments were uttered in a computer room whose background noine level was about 70 dB(A). The uttered sets 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 entire range of male voices, were selected from among 30 male speakers. The selection was based on the clustering results obtained in the course of a speaker-independent word recognition experiment using the SPLIT method[15]. In the SPLIT experiment, the utterances of these four speakers were most frequently used to construct multiple word templates.

In the recognition experiment using utterance set 1, the second utterances from each speaker were recognized using the first utterances from each of the three other speakers as reference templates (inter-speaker conditions only). That is, comparisons for the utterance set were always made between test utterances and single templates. The number of reference-test speaker combinations was 12, and the total number of test utterances was 1200.

(2) Utterance Set 2

The first utterances from all four speakers of utterance set 1 were stored as multiple templates, while the utterances from 20 differennt male speakers were used as test utterances. That is, the comparisons for this utterace set were made with four templates. The total number of test utterances was 2000, where one utterance from each speaker was tested for each word.

5.2 Recognition Results

(1) Effect of the Weighting Factor

The recognition rate as a function of the weighting factor for the first-order polynomial expansion coefficient, k_1, in Eq. (4) was examined by setting $k_2=0$, using utterance set 1. The energy polynomial coefficient was not used in this experiment (w_2 in Eq. (8) was set to 0). The results indicated in Fig. 6 show that error rate fluctuations as a function of the weighting factor, k_1, are small, and that the optimum condition producing the error rates close to the minimum value can be realized for a wide range of k_1.

Using the cepstrum coefficients modified by emphasizing the dynamics with an appropriate weighting factor, the recognition error rate can be reduced to roughly 2/3 of that obtained using the cepstrum without modification ($k_1=0$).

(2) Recognition Error Rate Improvement Using Various Parameter Combinations

Recognition experiments were performed by varying the combination of parameters using utterance sets 1 and 2. Figure 7 compares the mean recognition error rates for each parameter condition. Optimum values of the weighting factors ($k_1=k_2=8$) were decided based on the recognition experiments using utterance set 1, with these values then being applied to both utterance sets.

This figure confirms three important features. First, the cepstral polynomial coefficients are as effective as the instantaneous cepstrum coefficients. Second, the emphasis of spectral dynamics through the summation of both coefficients reduces the error rates to roughly 1/2 of those obtained before emphasizing the dynamics in the case of utterance set 2. Third, improvement is insignificant when second polynomial coefficients (second derivatives) are used in addition to the summation of the cepstrum and its first polynomial coefficients.

An error rate of 2.5% is produced for utterance set 2 by the combination of the energy polynomial coefficient and dynamics-emphasized cepstrum coefficients. This value is 2/5 of the 6.2% error rate obtained using the original cepstrum. Furthermore, it is 2/3 of the 3.8% error rate obtained employing the

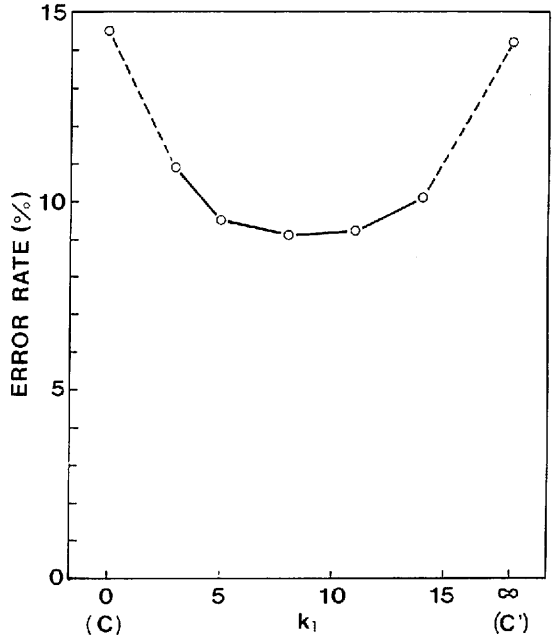

Fig. 6 Error rate for word recognition using utterance set 1 as a function of weighting factor, k_1, for 1st-order cepstrum polynomial coefficients.

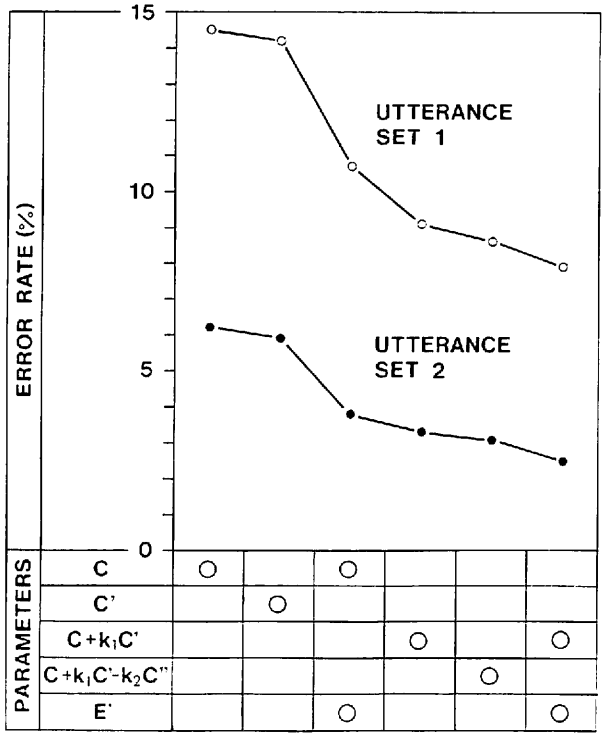

Fig. 7 Error rates for word recognition using utterance sets 1 and 2 under various feature parameter conditions: ○ indicates parameter set used.

combination of the energy polynomial coefficient and the cepstrum coefficients without dynamics emphasis.

An additional experiment revealed that when the energy polynomial coefficient is combined with the dynamics-emphasized cepstrum, there is no difference between the error rates associated with the conditions whether the dynamics emphasis of the cepstrum is performed using only the first-order polynomial coefficient or using both the first- and second-order polynomial coefficients.

The error rate for the summation of cepstrum coefficients and their first polynomial coefficients is slightly smaller than that obtained when both parameters were used for distance calculation as independent features. Although the difference of the error rates obtained by the method proposed in this paper and that obtained by our previous method in which the instantaneous and dynamic parameters were combined as independent parameters is insignificant, the new method is advantageous in practical terms in that it does not increase the number of parameters necessary for the distance calculation.

(3) Analysis of Error Rate Improvement

Table 1 compares the distribution of major confusable word pairs for utterance set 2 under the two principal parameter conditions, the original cepstrum and the cepstrum modified through dynamics-emphasis. Specifically presented are the confusable word pairs occuring two or more times in all test utterances by the 20 speakers under at least one of the parameter conditions. These word pairs account for 55% of all error under both parameter conditions. Those which occurs for the dynamics-emphasized cepstrum condition are given in the upper part of the table.

Table 1 Major confusable word pair frequencies compared under two feature parameter conditions: *N1*, recognition using instantaneous cepstrum, and *N2*, recognition using dynamics-emphasized cepstrum.

WORD PAIRS	FREQUENCY		
	N1	N2	N2-N1
Kawasaki - Amagasaki	6	6	0
Fukushima - Tokushima	2	5	+3
Ichinomiya - Nishinomiya	8	4	-4
Takasaki - Nagasaki	9	3	-6
Urawa - Nara	2	2	0
Ichihara - Ichikawa	2	2	0
Fuchu - Hachioji	0	2	+2
Kawasaki - Nagasaki	0	2	+2
Gifu - Fuchu	1	2	+1
Nara - Naha	5	2	-3
Hiroshima - Kashiwa	0	2	+2
Kochi - Otsu	2	2	0
Matsudo - Saseho	5	1	-4
Yamagata - Hirakata	3	0	-3
Wakayama - Okayama	3	1	-2
Toyama - Fukuyama	2	1	-1
Kanazawa - Nagano	2	0	-2
Fukui - Fuji	2	0	-2
Ichikawa - Fujisawa	2	1	-1
Funabashi - Maebashi	2	1	-1
Yokohama - Tokorozawa	2	0	-2
Hamamatsu - Takamatsu	2	0	-2
Toyohashi - Takasaki	2	0	-2
Akashi - Takatsuki	2	0	-2
Kurashiki - Takatsuki	2	0	-2
Shimonoseki - Naha	2	0	-2
OTHERS	54	23	
TOTAL	124	62	

Although the results using the original cepstrum coefficients include several confusable word pairs which are unlikely to occur in human speech perception, the process of emphasizing the dynamics to clarify the phonetic features removes these unnatural confusable word pairs.

The feature parameter extraction result presented earlier in Fig. 5 is an example of the utterance "Okayama", which was initially confused with "Wakayama" using the original cepstrum and was correctly recognized after emphasizing its spectral dynamics.

5.3 Comparison with the Previous Method

As was mentioned above in subsection 5.2, the difference of the mean error rate obtained by the method proposed in this paper and that obtained by our previous method is small. In order to investigate the mechanism of coincidence precisely, the relationship between the number of errors for the present and previous methods for each speaker is presented in the scatter diagram

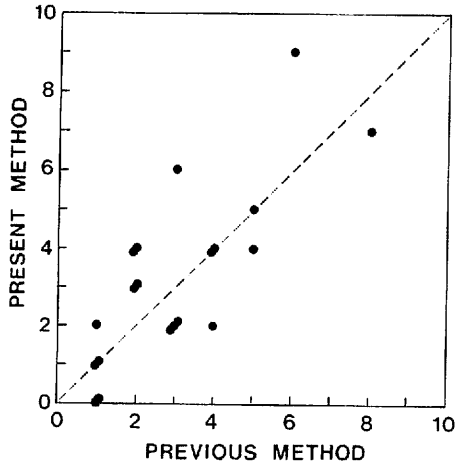

Fig. 8 Scatter diagram indicating the relationship between the number of errors by previous and present methods for 20 speakers.

labeled Fig. 8. The horizontal and vertical axes correspond to the number of errors produced by the previous and present methods respectively. These results indicate that a maximum difference of three samples exists between the two-method results for several speakers, and that the correlation between these two results is not high (correlation coefficient is 0.78). This suggests that the two methods use the same transitional information in different ways.

When the spectral dynamic characteristics are emphasized by the first derivative coefficients of the cepstrum in the present method, the distance measure between the k-th frame of the reference template and the l-th frame of the input speech can be written as

$$\begin{aligned} D_1(k, l) &= \sum_{n=1}^{10} (\tilde{C}_n^R(k) - \tilde{C}_n^I(l))^2 \\ &= \sum_{n=1}^{10} \{(C_n^R(k) + k_1 C'^R_n(k)) \\ &\quad - (C_n^I(l) + k_1 C'^I_n(l))\}^2 \\ &= \sum_{n=1}^{10} \{(C_n^R(k) - C_n^I(l)) \\ &\quad + k_1 (C'^R_n(k) - C'^I_n(l))\}^2 \\ &= \sum_{n=1}^{10} \{(C_n^R(k) - C_n^I(l))^2 + k_1^2 (C'^R_n(k) - C'^I_n(l))^2 \\ &\quad + 2k_1 (C_n^R(k) - C_n^I(l))(C'^R_n(k) - C'^I_n(l))\}. \end{aligned} \tag{9}$$

On the other hand, the distance measure for the previous method is

$$D_2(k, l) = \sum_{n=1}^{10} \{(C_n^R(k) - C_n^I(l))^2 + k_1^2 (C'^R_n(k) - C'^I_n(l))^2\}. \tag{10}$$

Hence, the mathematical relationship between the two distance measure is

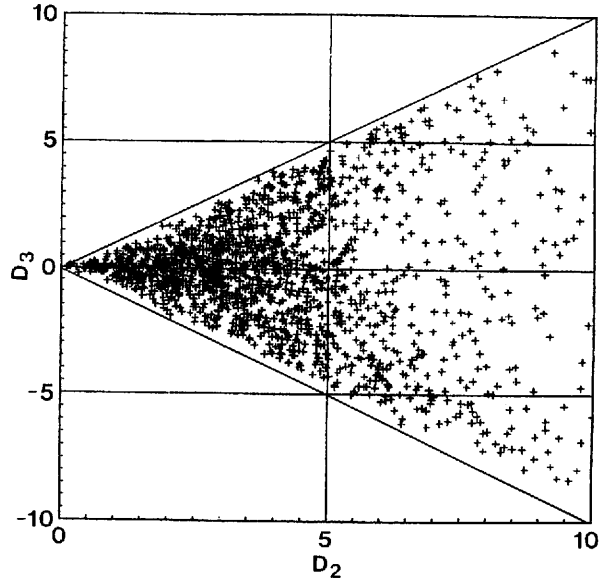

Fig. 9 Scatter diagram indicating the relationship between the two distance values. $D_2(k, l)$ and $D_3(k, l)$.

$$\begin{aligned} D_1(k, l) &= D_2(k, l) + D_3(k, l) \\ D_3(k, l) &= 2k_1 \sum_{n=1}^{10} (C_n^R(k) - C_n^I(l))(C'^R_n(k) - C'^I_n(l)), \end{aligned} \tag{11}$$

where $D_3(k, l)$ corresponds to the difference between the two measures.

The experimental results for the previous method indicate that the first and the second terms in the right side of Eq. (10) are nonredundant and complementary to each other for speech recognition. When the value of D_2 is small, both of these two terms must have small values. In this case, the value of D_3 also becomes small, and, therefore, D_1 has a similar value as D_2. However, on the other hand, when D_2 becomes large, D_1 does not necessarily become large, since D_3 can have both positive and negative values. The experimental relationship between D_2 and D_3 obtained using real speech is shown in the scatter diagram labeled Fig. 9. The D_3 values are clearly homogeneously distributed within the theoretical limit of $|D_3| \leqq D_2$. This means that no correlation exists between these two distance values. Therefore, there is no mathematical necessity that D_1 produces same results as D_2. The small experimental difference between the results produced by the previous and present methods can be concluded as occurring not in terms of mathematical necessity, but rather in terms of the same transitional information being used in both methods although in different ways.

6. Conclusion

Based on the auditory mechanism of speech perception which emphasizes spectral dynamics and compen-

sates for the spectral undershoot associated with coarticulation, a new speech analysis technique for speech recognition was proposed in this paper. In this technique, the speech wave is represented by LPC cepstrum sequences, and the time sequences over 56 ms were expanded by the first- and second-order polynomial functions every frame period. The dynamics of time sequences are then emphasized by the linear combination of the instantaneous cepstrum values and the polynomial expansion coefficients, that is, the derivatives.

The speaker independent word recognition experiment results indicate that the error rate using the dynamics-emphasized cepstrum is roughly 1/2 of that obtained using the original cepstrum coefficients. Confusable word pairs which are unlikely to occur in human speech perception were observed to disappear through phonetic feature clarification using spectral dynamics-emphasizing. The effectiveness of the time function of the polynomial coefficient for logarithmic energy was also ascertained. When the first-order polynomial coefficient for the energy sequence was combined with the dynamics-emphasized cepstrum, the error rate was found to be reducible to 2/5 of that using the original cepstrum.

Current investigations include psychological and physiological analyses of the mechanism of the human auditory system for time-varying speech signals and its modeling[16].

Acknowledgement

The author wishes to thank Kazuhiko Kakehi, former Head, and other members of Communication Principles Research Section 4 for their continual guidance and stimulating discussions.

References

(1) P. T. Brady, A. S. House and K. N. Stevens: "Perception of sounds characterized by a rapidly changing resonant frequency", J. Acoust. Soc. Am., **33**, pp. 1357-1362(1961).
(2) B. E. F. Lindblom and M. Studdert-Kennedy: "On the role of formant transition in vowel recognition", J. Acoust. Soc. Am., **42**, pp. 830-843(1967).
(3) H. Suzuki: "Mutually complementary effect between amount and rate of formant transition in perception of vowels, semivowels and voiced stops and a possible mechanism for their identification", J. Acoust. Soc. Jpn., **30**, pp. 169-180(1974).
(4) T. Ifukube: "Masking by frequency modulated tone", J. Acoust. Soc. Jpn., **29**, pp. 679-687(1973).
(5) S. Furui: "On the role of spectral transition for speech perception", Trans. Tech. Group on Hearing of Acoust. Soc. Jpn., H 85-6(1985).
(6) A. R. Muller: "Auditory physiology", Academic Press, New York (1983).
(7) K. Tanaka: "Construction of a phonetic feature space and dynamic processing in the feature space", Trans. Tech. Group on Speech of Acoust. Soc. Jpn., S 74-20(1974).
(8) H. Fujisaki, N. Higuchi, T. Futami and S. Shigeno: Perception of non-stationary sound stimuli and a model of the underlying processes", Trans. Tech. Group on Speech of Acoust. Soc. Jpn., S 81-35(1981).
(9) Y. Kanamori and K. Kido: "Recognition of vowel in connected speech by use of the characteristics on dynamic perception of vowel", Spring Meeting of Acoust. Soc. Jpn., 2-2-8(1975).
(10) H. Kuwabara: "An approach to normalization of coarticulation effects for vowels in connected speech", J. Acoust. Soc. Am., **77**, pp. 686-694(1985).
(11) S. Furui: Cepstral analysis technique for automatic speaker verification", IEEE Trans. Acoust., Speech & Signal Process., **ASSP-29**, pp. 254-272(1981).
(12) S. Furui: "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. Acoust., Speech & Signal Process., **ASSP-34**, pp. 52-59(1986).
(13) ed. S. Namba: "Handbook of hearing", Nakanishiya Shuppan, Kyoto, Japan(1984).
(14) S. Furui: "Physical characteristics of essential speech intervals for phoneme perception", Fall Meeting of Acoust. Soc. Jpn., 1-3-15(1985).
(15) N. Sugamura, K. Shikano and S. Furui: "Isolated word recognition using phoneme-like templates", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Boston, MA, pp. 723-726(1983).
(16) S. Furui and M. Akagi: "On the role of spectral transition in phoneme perception and its modeling", Proc. 12th Int. Cong-Acoust., A 2-6(1986).

Sadaoki Furui 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 The Tokyo University, Tokyo, Japan, in 1968, 1970, and 1978, respectively.

After joining the Electrical Communications Laboratories, Nippon Telegraph and Telephone Corporation in 1970, he studied the analysis of speaker characterizing information in the speech, its application to speaker recognition as well as interspeaker normalization in speech recognition, the vector-quantization-based speech recognition algorithm, and the analysis of speech perception mechanism.

He is currently the Head of Communication Principles Research Section 4 in charge of the fundamental research on speech signal processing and acoustics.

From December 1978 to December 1979 he joined the staff of the Acoustics Research Department at Bell Laboratories, Murray Hill, New Jersey, as an exchange visitor working on speaker verification.

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

He received the Yonezawa Prize for the paper "Analysis of Speaker Dependent Features in Spectral Transition in Speech Wave" from the IECEJ in 1975, and the Sato Paper Award for the paper "Isolated Word Recognition Using Strings of Phoneme-like Templates(SPLIT)" from the ASJ in 1985. He is the author of "Digital Speech Processing" (Tokai University Press, 1985).