論文 / 著書情報
Article / Book Information

Title	Speaker-Independent Isolated Word Recognition Based on Emphasized Spectral Dynamics
Author	SADAOKI FURUI
Journal/Book name	IEEE ICASSP1986, Vol. , No. , pp. 1991-1994
発行日 / Issue date	1986, 4
権利情報 / Copyright	(c)1986 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Reprinted from PROCEEDINGS OF IEEE-IECEJ-ASJ INTERNATIONAL CONFERENCE
ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, April 7—11, 1986

SPEAKER-INDEPENDENT ISOLATED WORD RECOGNITION BASED ON EMPHASIZED SPECTRAL DYNAMICS

SADAOKI FURUI

NTT Electrical Communication Laboratories

Musashino-shi, Tokyo, 180 Japan

ABSTRACT

A new speech analysis technique applicable to speech recognition is proposed considering the auditory mechanism of speech perception which emphasizes spectral dynamics and which 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.

I. 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 to 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][2]. The other is the context-dependent hearing mechanism of the contrast effect.

A model for the speech perception mechanism which compensates the spectral undershoot based on its dynamic features has already been proposed[3]. In the model, the first- and second-derivatives as well as the instantaneous values for formant transition were integrated with time-axis weighting. 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 [4].

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 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 [5], no effective method has yet been proposed for improving recognition performance which can be applied to 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 [6] and in speech recognition [7]. The difference between the present and previous methods lies in the technique of combining dynamic and instantaneous features.

II. 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 taken out 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 [8] and to the minimum length for syllable perception [9]. Additionally, this length seems adequate for preserving the transitional information associated with the changes

ICASSP 86, TOKYO CH2243-4/86/0000-1991 $1.00

occurring from one phoneme to another.

III. EMPHASIS OF SPECTRAL TRANSITION USING CEPSTRUM POLYNOMIAL EXPANSION COEFFICIENTS

Ten-dimensional cepstrum vectors for the adjacent seven frames are denoted by C_j (j = 1,...,7), and the first- and second-order polynomial expansion coefficients for their time functions are denoted by C' and C" respectively. In this paper, the following orthogonal polynomial representations are used:

$$P_{0j} = 1,$$
$$P_{1j} = j - 4,$$
$$P_{2j} = j^2 - 8j + 12. \quad (1)$$

C' and C" can then be represented by

$$C' = \sum_{j=1}^{7} C_j P_{1j} / \sum_{j=1}^{7} P_{1j}^2,$$

$$C'' = \sum_{j=1}^{7} C_j P_{2j} / \sum_{j=1}^{7} P_{2j}^2. \quad (2)$$

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,

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

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

$$\tilde{C} = C + k_1 C' - k_2 C''. \quad (4)$$

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

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

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

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,

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

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

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

This indicates that the spectral transition is emphasized and that spectral undershoot in continuous speech can be compensated for by the method proposed in this paper.

An auditory model including the above-mentioned process which emphasizes the spectral transition is shown in Fig. 1. An example of a short-term variation of the logarithmic spectral envelope and its modification by emphasizing the transition is shown in Fig. 2.

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.

IV. 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. 3. The beginning and end of the actual sample utterances are determined through a short-term energy calculation. The cepstrum coefficient and energy time functions are extracted through LPC analysis for the speech interval. 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 insignificant with respect to phoneme perception, and because the performance of a recognition system using the energy level is vulnerable to its variation among speakers and speaking times.

In order to reduce the number of distance calculations at the time registration stage, the frame

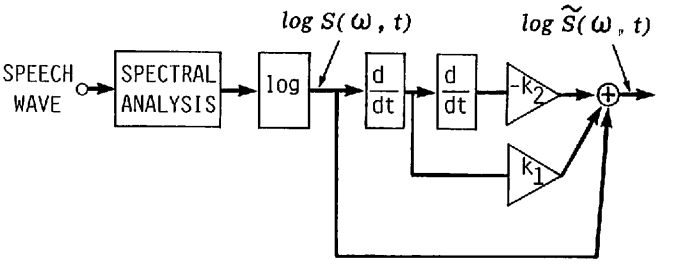

Fig. 1 - Block diagram of an auditory model employing spectral dynamics-emphasis mechanism.

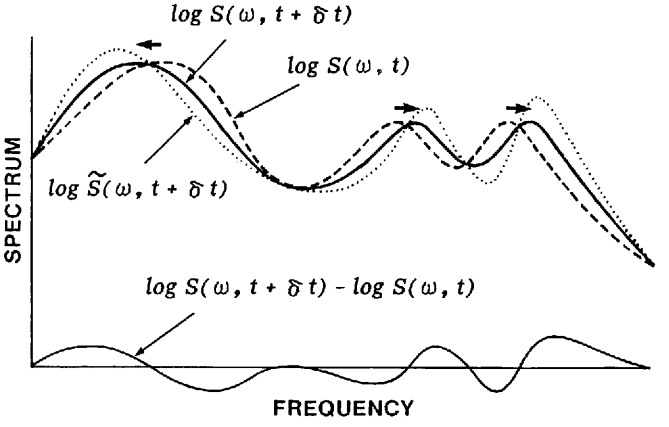

Fig. 2 - Example of time variation of logarithmic spectral envelope and its modification by dynamics-emphasis.

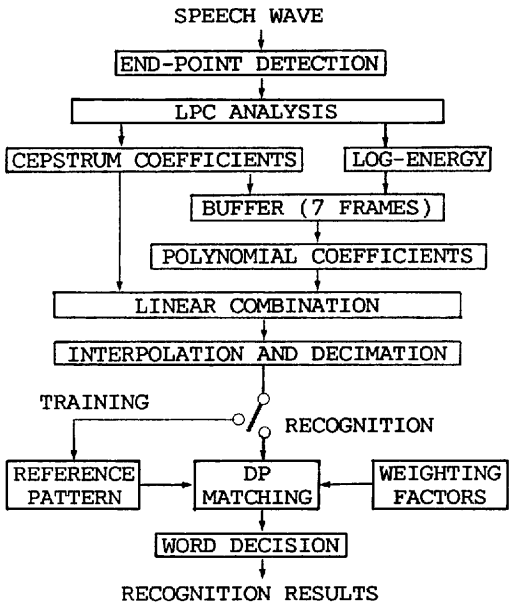

Fig. 3 - Block diagram indicating principal operations of spoken word recognition system.

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.

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 [7].

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 at 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 defined 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 at a preliminary experiment.

V. 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 [7] 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). 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 [10]. 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 different male speakers were used as test utterances. That is, the comparisons for this utterance 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

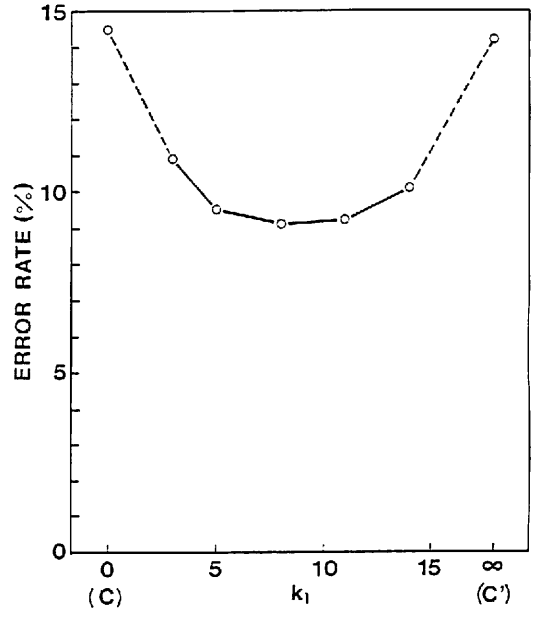

Fig. 4 - Error rate for word recognition using utterance set 1 as a function of weighting factor, k_1, for 1st-order cepstrum polynomial coefficients.

results indicated in Fig. 4 show that error rate fluctuations as a function of the weighting factor, k_1, are small and that the optimum condition can be realized for a wide range of k_1.

Using the cepstrum coefficients modified by emphasizing the dynamics using 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 5 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.

Figure 5 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 using second polynomial coefficients (second derivatives) in addition to the summation of the cepstrum and its first polynomial coefficients.

The error rate for the combination of the polynomial coefficient of the energy and the dynamics-emphasized cepstrum coefficients produces an error rate of 2.5 % for utterance set 2. This value is 2/5 of the error rate of 6.2 % obtained using the original cepstrum. Furthermore it is 2/3 of the error rate of 3.8 % obtained using the combination of the energy polynomial coefficient and the cepstrum coefficients without dynamics emphasis.

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. Additionally, the new method is advantageous in practical terms in that it does not increase the number of parameters necessary for the distance calculation.

Distribution of major confusible word pairs indicates that, although the results using the original cepstrum coefficients include several confusible word pairs which are unlikely to occur in human speech perception, the process of emphasizing the dynamics to clarify the phonetic features removes these unreasonable confusible word pairs.

VI. CONCLUSION

Based on the auditory mechanism of speech perception which emphasizes spectral dynamics and compensates for the spectral undershoot associated with coarticulation, a new speech analysis technique for speech recognition was proposed in this paper. In this technique, the spectral dynamics are emphasized by the linear combination of the polynomial expansion coefficients, that is, the derivatives, and the instantaneous cepstrum values. 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. 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 [11].

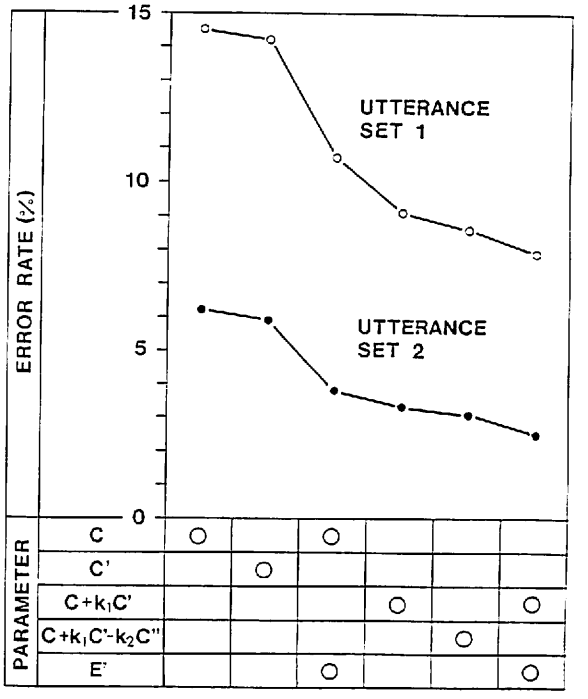

Fig. 5 - Error rates for word recognition using utterance sets 1 and 2 under various feature parameter conditions; ○ indicates parameter set used.

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., vol.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., vol.42,pp.830-843,1967

[3] K.Tanaka, "Construction of a phonetic feature space and dynamic processing in the feature space," Trans. Tech.Group on Speech of Acoust.Soc.Jpn.,S74-20, 1974

[4] 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.,S81-35,1981

[5] H.Kuwabara, "An approach to normalization of coarticulation effects for vowels in connected speech," J.Acoust.Soc.Am.,vol.77,pp.686-694,1985

[6] S.Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-29,pp.254-272,1981

[7] S.Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-34,1986 (to be published)

[8] S.Namba(Ed.), "Handbook of hearing," Nakanishiya Shuppan,Kyoto,Japan,1984

[9] S.Furui, "Physical characteristics of essential speech intervals for phoneme perception," Fall Meeting of Acoust.Soc.Jpn., 1-3-15,1985

[10] 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

[11] S.Furui and M.Akagi, "On the role of spectral transition in phoneme perception and its modeling," 12th ICA, 1986 (to be published)