論文 / 著書情報
Article / Book Information

Title	Spectral Movement Function and its Application to Speech Recognition
Author	Kiyoaki Aikawa, Sadaoki Furui
Journal/Book name	IEEE ICASSP1988, Vol. , No. , pp. 223-226
発行日 / Issue date	1988, 4
権利情報 / Copyright	(c)1988 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.

Reprint from PROCEEDINGS FROM ICASSP — INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, New York, N.Y., April 11-14, 1988

S5.11

SPECTRAL MOVEMENT FUNCTION AND ITS APPLICATION TO SPEECH RECOGNITION

Kiyoaki AIKAWA and Sadaoki FURUI

NTT Human Interface Laboratories
Musashino-Shi, Tokyo 180, Japan

ABSTRACT

In this paper we propose a new mathematical method for extracting spectral movement in a time-sequence of speech spectrum. The spectral movement is characterized by the time and frequency derivative of a time-sequence of log spectrum envelopes. Spectral movement direction, the movement toward a higher or a lower frequency region, can be identified by the sign of the proposed function. A parameter which can be used for speech segmentation is derived from this function. A distance measure for speech recognition is also derived as the Euclidean distance between two spectral movement patterns extracted by the proposed function. This distance is easily calculated using cepstrum coefficients. Speech recognition results using Dynamic Time Warping (DTW) template matching with this new distance measure indicate that recognition error rate can be reduced to less than half compared with the conventional Euclidean cepstrum distance measure.

INTRODUCTION

Of the various speech signal features, spectral movement, especially at the spectral peaks, provides significant information in auditory perception of phonemes [1]. If this spectral movement can be extracted efficiently and reliably, it will be useful in enhancing automatic speech recognition performance. Conventional methods for extracting spectral movement information are mainly based on formant trajectory estimation. However, tracking formant trajectories is usually difficult and error-prone. Therefore, it is desirable to extract spectral movement without resorting to formant tracking. It is also desirable that the algorithm to extract spectral movement reflects an auditory neural network model.

One of the authors previously proposed a combined spectral distance measure where instantaneous spectral information was used in conjunction with transitional spectral information for both speaker verification[2] and speech recognition [3]. In that proposed distance measure, the dynamic aspects of spectral movement were emphasized without tracking formant explicitly.

In the same spirit, Oka used a vector-field approach for speech recognition. He used a two-dimensional (time and frequency) spectral gradients for characterizing spectral movements [4].

In this paper we propose a new function based on a functional model of auditory perception. The goal is to extract spectral movement from a time sequence of speech spectrum without using formant tracking. We named this Spectral Movement Function (SMF).

The proposed SMF is applied to speech segmentation and template-matching-based speech recognition. For the former purpose, a segmentation parameter is introduced using the SMF which can detect phoneme boundaries even if there is only slow spectral movement. For the latter purpose, a distance measure between two spectral movement patterns extracted by SMF is proposed. The effectiveness of this measure is shown.

SPECTRAL MOVEMENT FUNCTION

In this paper, spectral movement is defined as the amount of spectral energy which passes through a point on frequency axis per unit time. The spectral movement value should be zero when the spectral shape is invariant even if its energy is changing. Spectral movement is characterized by a mathematical function which meets the following conditions.

(a) The function should only respond to the spectral peak movement and not to the spectral energy change.
(b) The function should discriminate between spectral movement toward higher and lower frequencies.
(c) The function should reflect the auditory neural network models.

Since the perceptual intensity is proportional to the logarithm of input energy, the output of the lower level auditory neural network is modeled by a log spectrum $S(\omega, t)$ where ω is frequency and t is time. The instantaneous energy $q(t)$ is given by the average of $S(\omega, t)$ over the whole frequency range as

$$q(t) = \frac{1}{\pi} \int_0^{\pi} S(\omega, t)\, d\omega \,. \qquad (1)$$

and the average spectrum $g(\omega)$ over the maximum time window $[0, t_{max}]$ of interests is defined as

$$g(\omega) = \frac{1}{t_{max}} \int_0^{t_{max}} S(\omega, t)\, dt \qquad (2)$$

$S(\omega, t)$ can be represented as follows.

$$S(\omega, t) = R(\omega, t) + q(t) + g(\omega) \qquad (3)$$

where $R(\omega, t)$ corresponds to the residual spectral component after the "frequency-stationary" and "time-stationary" spectral components, $q(t)$ and $g(\omega)$, have been removed. Since $q(t)$ and $g(\omega)$ are insensitive to spectral movement, our objective function for spectral movement extraction, denoted as $f\{\ \}$, should meet the equation

$$f\{\ S(\omega, t)\ \} = f\{\ R(\omega, t)\ \} \qquad (4)$$

for arbitrary spectrum $S(\omega, t)$.

One solution which meets equation(4) under the conditions (a), (b), and (c), is the time and frequency differentiation

$$f\{\ \} = \frac{\partial^2}{\partial\omega\ \partial t} . \qquad (5)$$

This function is adopted as the Spectral Movement Function (SMF) and the extracted movement, denoted as

$$F(\omega, t) = f\{\ S(\omega, t)\ \} , \qquad (6)$$

is named "spectral movement pattern" hereafter.

Let us consider the spectral movement pattern of speech spectrum passed through a time-invariant linear filter. The SMF of the filtered log spectrum has the property of

$$f\{\ S(\omega, t) + \ln(\ H(\omega)\ )\ \} = f\{\ S(\omega, t)\ \} \qquad (7)$$

where $H(\omega)$ is the transmission function of the linear filter. This equation indicates that the SMF has a capability of removing the effect of a time-invariant transmission channel. It is also expected that the SMF is relatively invariant with respect to the variation of average spectrum due to vocal excitation source. In Fig. 1, a schematic diagram illustrates the spectral movement extraction mechanism. An equivalent auditory neural network model is shown in Fig. 2.

It should be noted that the SMF responds not only to the spectral peak movement, but also to the spectral valley movement. In addition, the function is accompanied by sidelobes as shown in Fig. 1. To cope with these problems, LPC (linear predictive coding)-based spectral envelope is used as $S(\omega, t)$. Consequently, the spectral movement pattern takes a much higher amplitude for spectral peaks than for spectral valleys. An example of the spectral movement pattern is shown in Fig. 3.

SMF CHARACTERISTICS

Spectrum moving speed $V(\omega, t)$ (rad/sec) is defined as the spectrum moving angle on the frequency axis per unit time. In Fig. 4, a, b, and their ratio are given by

$$a = \frac{\partial\ S(\omega, t)}{\partial\ t}\ \Delta t , \qquad (8)$$

$$b = V(\omega, t)\ \Delta t , \qquad (9)$$

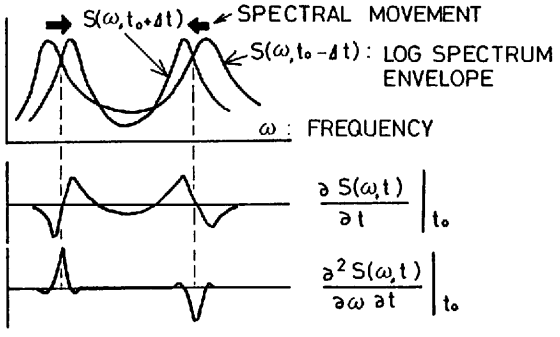

Fig. 1 The mechanism of spectral movement extraction by time and frequency differentiation.

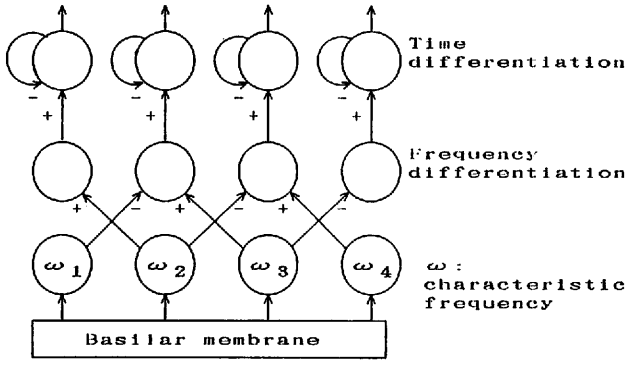

Fig. 2 A neural network model for spectral movement extraction.

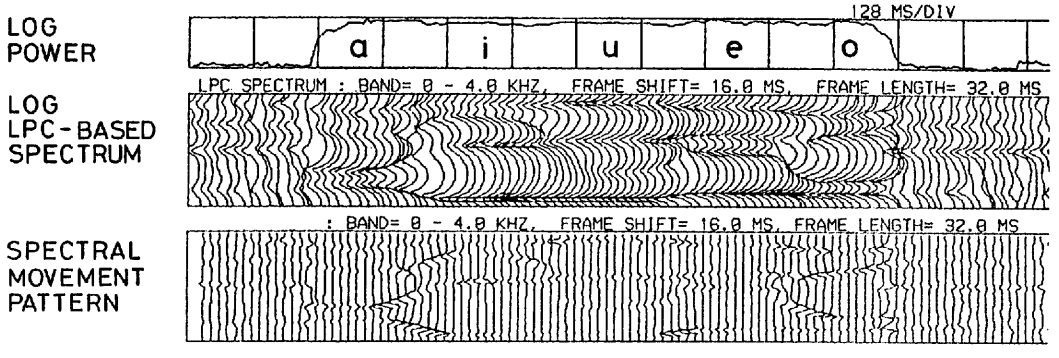

Fig. 3 Spectral movement pattern extracted by SMF. (speech : /aiueo/)

and

$$a/b = -\frac{\partial S(\omega, t)}{\partial \omega} . \quad (10)$$

Substituting equation (8) and (9) for a and b in equation (10), $V(\omega, t)$ is

$$V(\omega, t) = -\frac{\partial S(\omega, t)}{\partial t} \Big/ \frac{\partial S(\omega, t)}{\partial \omega} . \quad (11)$$

After the denominator $\frac{\partial S(\omega, t)}{\partial \omega}$ is multiplied to both sides of equation (11), both sides are differentiated with respect to ω. Then, the spectral movement pattern $F(\omega, t)$ is given by

$$F(\omega, t) = \frac{\partial^2 S(\omega, t)}{\partial \omega \, \partial t} \quad (12)$$

$$= -\frac{\partial V(\omega, t)}{\partial \omega}\frac{\partial S(\omega, t)}{\partial \omega} - V(\omega, t)\frac{\partial^2 S(\omega, t)}{\partial \omega^2} . \quad (13)$$

If the spectrum moving speed $V(\omega, t)$ is assumed to be a constant function of ω around a spectral peak, the $F(\omega, t)$ is then the spectrum moving speed multiplied by the second order frequency derivative of the log spectrum.

PHONEME SEGMENTATION PARAMETER

Since SMF takes a positive or negative value depending upon whether the spectral energy moves toward a higher or lower frequency, the total ascending and descending spectral movement over the entire frequency range can be obtained by

$$A(t) = \frac{1}{\pi}\int_0^{\pi} p[\, F(\omega, t) \,] \, d\omega , \quad (14)$$

and

$$D(t) = \frac{1}{\pi}\int_0^{\pi} p[\, -F(\omega, t) \,] \, d\omega \quad (15)$$

where

$$p[x] = x , \quad x \geq 0 \quad (16)$$
$$= 0 , \quad x < 0 .$$

These parameters are applicable to speech segmentation at phoneme boundaries.

SPECTRAL-MOVEMENT-PATTERN DISTANCE MEASURE

A log spectrum envelope can be written using cepstrum coefficients $c_k(t)$ $(k=1, 2, \ldots, K)$ as

$$S(\omega, t) = c_0 + 2\sum_{k=0}^{K} c_k(t) \cos k\omega \quad (17)$$

where K is the maximum quefrency. The spectral movement pattern $F(\omega, t)$ can then be written as

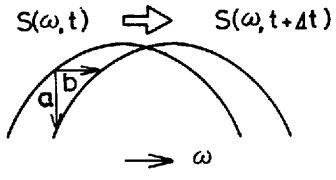

Fig. 4 Definition of a and b parameters related to spectral movement.

$$F(\omega, t) = 2\sum_{k=1}^{K} \frac{d\, c_k(t)}{dt} k \sin k\omega . \quad (18)$$

The well-known cepstral distance measure is defined as the Euclidean distance between two log spectra $S_1(\omega, t)$ and $S_2(\omega, t)$ represented by cepstrum coefficients as

$$d_{CEP}(t) = \frac{1}{\pi}\int_0^{\pi} (S_1(\omega, t) - S_2(\omega, t))^2 \, d\omega . \quad (19)$$

A Euclidean distance between two spectral movement patterns $F_1(\omega, t)$ and $F_2(\omega, t)$ can be defined by

$$d_{SM}(t) = \frac{1}{\pi}\int_0^{\pi} (F_1(\omega, t) - F_2(\omega, t))^2 \, d\omega . \quad (20)$$

By replacing $F(\omega, t)$ with equation (18), $d_{SM}(t)$ can be written as

$$d_{SM}(t) = \frac{4}{\pi}\int_0^{\pi} \{ \sum_{k=1}^{K} (\frac{d\, c_{1k}(t)}{dt} - \frac{d\, c_{2k}(t)}{dt}) k \sin k\omega \}^2 \, d\omega \quad (21)$$

$$= 2\sum_{k=1}^{K} \{ (\frac{d\, c_{1k}(t)}{dt} - \frac{d\, c_{2k}(t)}{dt}) k \}^2 . \quad (22)$$

Because even a small fluctuation will be emphasized by the time and frequency differentiations, the log spectrum sequence should be reasonably smooth. Frequency domain smoothing can be accomplished using quefrency liftering of cepstral coefficients. One of the authors proposed weighted cepstrum and weighted differential cepstrum distance measures for speaker verification [2]. Tohkura investigated a weighted cepstrum distance measure defined by

$$d_{WCEP}(t) = \sum_{k=1}^{K} \frac{1}{v_l} (c_{1k}(t) - c_{2k}(t))^2 \quad (23)$$

where $l = k , \quad k < L$
$= L , \quad k \geq L$

and where L is the quefrency where the weighting saturates [5]. This weighting saturation is effective for obtaining reliable distance values. In [5] intra-speaker variance of k-th order cepstrum coefficient, v_l, is reported to be approximately proportional to the inverse of l^2. In the experiments using this measure, which will be described later, l^2 is used instead of $1/v_l$.

For obtaining a reasonable estimate of the time differentiation without introducing a large amount of noise, the use of the best linear fitting is effective [2-3]. Soong et al. have reported the usefulness of a distance measure d_{WDCEP} defined by

$$d_{WDCEP}(t) = \sum_{k=1}^{K} \frac{1}{v_k} (c_{1k}(t)' - c_{2k}(t)')^2 \quad (24)$$

for speaker recognition [6]. Here, $c_k(t)'$ means the best linear fitting slope of the k-th order cepstrum coefficient time series.

Based on these investigations, the distance measure d_{SM} is finally defined by

$$d_{SM}(t) = 2\sum_{k=1}^{K} l^2 (c_{1k}(t)' - c_{2k}(t)')^2 \quad (25)$$

where $l = k , \quad k < L$
$= L , \quad k \geq L .$

In equation (24), if k^2 is used instead of v_k, d_{WDCEP} is the same as the distance measure d_{SM}.

It is thought that both instantaneous spectrum information and transitional information are

necessary for speech recognition. Therefore, a joint distance measure d(t) defined as

$$d(t) = r\, d_{SM}(t) + (1-r)\, d_{CEP}(t) \qquad (26)$$

is expected to show a better performance than either of the two distance measures. For an LPC log spectrum, cepstrum coefficients can be derived from the corresponding LPC coefficients through the standard recursion formula.

EXPERIMENTS

First, segmentation experiments were performed. One of the authors proposed a segmentation parameter based on an onset-sensitive auditory neuron model [7]. Although this parameter is very efficient, segmenting /kj/ from a following vowel was very difficult, and the segmentation error rate was as high as 87.5 %. When the new segmentation parameter based on SMF was used, the error rate was reduced to 12.5 %.

Next, speech recognition experiments were performed. The performance of the proposed distance measure d_{SM} was evaluated through the recognition of 50 Japanese city names using DTW. The test speech sets were spoken by ten male speakers and the reference template set was spoken by a different male speaker. Each speaker uttered each word once.

The optimal combination of the weighting factor r and the weighting saturation quefrency L were examined in various conditions. The minimum error rate (4.4 %) was achieved in case where r=0.5 and L=3, and is shown in Fig. 5. This error rate is less than half of that obtained using the cepstrum distance measure (9.5 %). In Fig. 5, L=1 means that d_{SM} does not include frequency differentiation and is composed of only the time differentiation. The minimum error rate obtained by this distance measure at the optimal condition (r=0.9) was 5.5 %. The minimum error rate obtained using the weighted cepstrum distance measure d_{WCEP} with the weighting saturated at L=3 was 5.9 %.

These experimental results indicate that the spectral movement pattern provides significant information for speech recognition. One possible reason why the proposed spectral movement pattern distance measure have achieved the best result is that the spectral tilt of the speech spectrum caused by an excitation source, such as the effect of variable vocal effort, is alleviated by SMF.

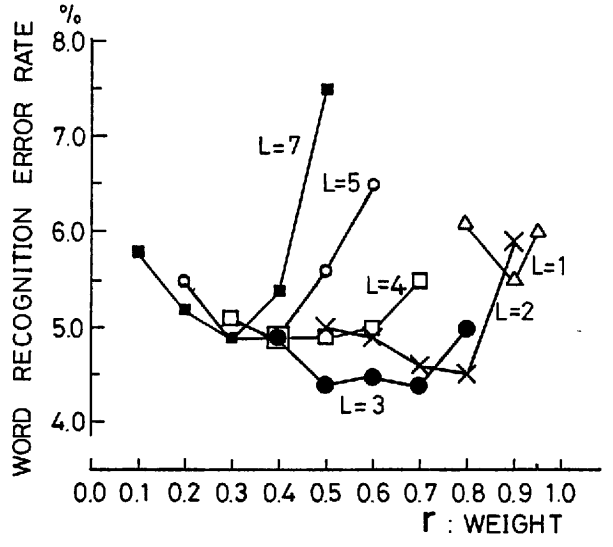

Fig. 5 The word recognition error rates obtained by the combined distance measure d(t).

CONCLUSION

In this paper, a function for extracting spectral movement has been proposed based on an auditory neural network model. The function is defined as the time and frequency derivative of a log LPC-based spectrum envelope. The proposed function is used to quantify spectral energy movement, to separate the movement toward higher and lower frequencies. A distance measure between two spectral movement patterns has been proposed. The speech recognition performance of the new measure was tested and compared with other existing cepstrum-based distance measures. Experimental results show that the proposed distance measure can reduce the speech recognition error rate to less than half compared with the conventional cepstrum distance measure. The proposed measure is also better than a weighted cepstrum distance measure. A segmentation parameter based on the spectral movement function has also been proposed and its effectiveness for detecting phoneme boundaries where the spectral pattern changes very slowly has been shown. These results indicate that transitional spectral information is important for speech recognition.

ACKNOWLEDGMENT

The authors wish to thank Kazuhiko KAKEHI, Masaki KOHDA, and Shigeki SAGAYAMA for their valuable suggestions. The authors also wish to thank Frank K. Soong at AT&T Bell Laboratories for the revision of this paper.

REFERENCES

[1] A. Cole (editor), Perception and Production of Fluent Speech, Lawrence Erlbaum Associates, Publishers, (1980).

[2] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," IEEE Trans., ASSP-29, No. 2, pp. 254-272, (1981-4).

[3] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," IEEE Trans., ASSP-34, No. 1, pp. 52-59, (1986-2).

[4] R. Oka, "Comparison Studies on the Effectiveness Between the Vector Field of Spectrum and Spectrum for Speech Recognition," IECE Trans., J69-D, No. 11, pp. 1704-1713, (1986-11).

[5] Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition," IEEE Trans., ASSP-35, No. 10, pp. 1414-1422, (1987-10).

[6] F. K. Soong and A. E. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," Proc. of ICASSP86, pp. 877-880, (1986-4).

[7] K. Aikawa, "Segmentation Parameter Based on the Positive Spectrum Change Detection," Trans. on IECE meeting, Vol. 86, No. 93, pp. 9-16, (1986-7).