

SPOKEN WORD RECOGNITION USING STATISTIC AND DYNAMIC INFORMATION
 OBTAINED BY TWO-DIMENSIONAL CEPSTRUM ANALYSIS

Y. Arikai*, S. Mizuta*, M. Nagata* and T.Sakai*

ABSTRACT

In this paper, Two-Dimensional Cepstrum (TDC) analysis and its application to word recognition are described. The TDC can represent two different kinds of information contained in speech wave forms simultaneously: static and dynamic information, global and fine frequency structure. Noise reduction filtering or speech enhancement filtering is easily established on this TDC. It is shown that the TDC is an effective parameter for word recognition by both DP-matching and linear matching. Through the word recognition experiments, it is confirmed that the global static information and slow dynamic information are effective for that recognition.

INTRODUCTION

In a conventional speech recognition system, speech data is usually represented as a time sequence of parameter vectors produced through a stationary analysis like LPC. The static information about frequency components of the speech within a short time period (frame) is explicitly represented. The dynamic information about the time variation, however, is implicitly represented as a time sequence of the static information. Since the dynamic information provides important cues for the perception of consonants or vowel transitions, it is necessary to develop an explicit representation of the dynamic information and its application to speech recognition(ref 1). As a representation of this dynamic information, we employ a Two Dimensional FFT-Cepstrum (TDC) analysis which can represent as well as the static information the time varying and frequency varying information simultaneously. Word and monosyllable recognition are carried out on this TDC by two different methods. One uses a DP-matching to absorb the time difference. The other uses, for high speed processing, a linear matching of spoken word matrixes produced by the TDC analysis. Finally, the recognition accuracy is evaluated by the recognition rate under a noisy environment as well as a quiet environment. It is shown that the recognition rate increases when the dynamic information is enhanced, while the recognition rate of the noisy speech increases when the dynamic information is suppressed.

ANALYSIS OF DYNAMIC INFORMATION BY A TWO-DIMENSIONAL CEPSTRUM

A one-dimensional FFT-cepstrum (ODC) analysis applies one-dimensional FFT to the short time log spectrum. On the other hand, a two-dimensional FFT-cepstrum (TDC) analysis applies two-dimensional FFT to a time sequence of the log spectrum, and converts it to the two-dimensional cepstrum. In the analysis, therefore, several consecutive frames are grouped as a block and processed. It is desired that the time duration of a block almost equals that of human perception, for example, 70 ms for vowels. The TDC is formalized as follows. Let us denote the frame length by N, the number of frame in a block by M. The speech wave form x_{nm} sampled at the nth point in the mth frame in a block is represented as:

$$x_{nm} = x_{n+mL} \quad (0 \leq n \leq N-1, 0 \leq m \leq M-1, 0 < L < N) \quad (1)$$

where L indicates the frame period. The log spectrum S_{km} of the mth frame in a block is expressed as:

$$S_{km} = 10 \cdot \log |\sum x_{nm} W_1^{-nk}|^2 \quad (W_1 = \exp(j2\pi/N), 0 \leq k \leq N-1) \quad (2)$$

The TDC coefficients C_{qp} are obtained by applying the two-dimensional FFT to a time sequence of the log spectrum S_{km} as to the frequency k and the time m as follows:

$$C_{qp} = (1/NM) \sum \sum S_{km} W_1^{-kq} W_2^{-mp} \quad (W_2 = \exp(j2\pi/M), 0 \leq q \leq N-1, 0 \leq p \leq M-1) \quad (3)$$

* Dept. of Information Science, Faculty of Engineering, Kyoto University, Kyoto, Japan, 606.

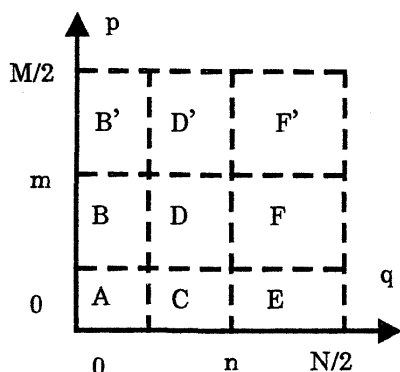


Fig.1 Nine different regions on the TDC.

Table1 Analysis and recognition condition.

	condition
analysis	12bits, 10kHz sampling frame length: 25.6 m frame period: 6.4 ms block length: 8 frames block period: 4 frames
recognition	parameter: TDC coefficient { $C_{qp}, 0 \leq q \leq 14,$ $0 \leq p \leq m, (q,p) \neq (0,0)$ } DP-matching with end point free TDC distance measure

Table 2 Result of word and syllable recognition(%) .
(): the third candidates are included.

speech data	parameter	
	TDC	ODC
word	93.0(97.0)	94.5(97.0)
monosyllable	87.5(97.0)	85.0(94.5)

The axis "q" is the quefreny and has the time dimension. The axis "p" is the time frequency and has the frequency dimension. The number of the TDC coefficients C_{qp} is actually reduced to 1/4 of the total, $(1 + M/2)(1 + N/2)$. A higher component on the q-axis, including the pitch information, corresponds to a fine structure of the spectrum, and the lower component to the spectral envelope. This has the same meaning in the conventional one-dimensional FFT-cepstrum. On the other hand, a higher component on the p-axis corresponds to the local time variation and the lower component to the global time variation. According to the meaning of the axes p and q, the TDC coefficients are divided into nine regions A to F as shown in Fig. 1. The meaning of each region is as follows:

- A: Averaged value of the log spectrum in a block
- B: Global time variation of the averaged log spectrum
- B': Local time variation of the averaged log spectrum
- C: Spectral envelope
- D: Global time variation of the spectral envelope
- D': Local time variation of the spectral envelope
- E: Spectral fine structure
- F: Global time variation of the spectral fine structure
- F': Local time variation of the spectral fine structure

The features of the TDC are:

- (1) The static information(A, C, E) and the dynamic information(B, D, F) are represented.
- (2) The global time variation(B, D, F) and the local time variation(B', D', F') are represented.
- (3) The spectral envelope (C) and the fine structure (E) are separately represented.

WORD AND SYLLABLE RECOGNITION ON THE TDC BY DP-MATCHING

Analysis and Recognition Condition

Table 1 shows the condition for the analysis and recognition. A block consists of eight frames and block length is then about 70 ms (6.4 ms x 7 + 25.6 ms) which almost corresponds to the perceptual time duration. The range of the quefreny q is determined as $0 \leq q \leq 14$ according to the experimental results and the knowledge that the spectral envelope is important for speech recognition. The parameter C_{00} is not used because of the power normalization. The DP matching is applied to a time sequence of the TDC coefficients which include the dynamic information.

Range of the Dynamic Information Effectlive for Recognition

A word recognition experiment was carried out to investigate the effective range of the dynamic information by changing the value of the parameter m on the time frequency p . The speech data is 100 Japanese city names (words) spoken by two females. As a result, the recognition rate reaches almost the upper bound at $m=1$ (about 20Hz). This means that the spectral envelope ($p=0$, $1 \leq q \leq 14$) and its global time variation ($p=1$, $0 \leq q \leq 14$) is effective for recognition. Using this range of the dynamic information, the word and monosyllable recognition were carried out. The result is shown in Table 2 with the result by one-dimensional FFT cepstrum(ODC). The frame length and frame period in the ODC analysis are 25.6 ms and 12.8 ms respectively with $1 \leq q \leq 14$. The speech data is 101 Japanese monosyllables. From Table 2, the TDC is effective for monosyllables whose time duration is shorter than words. This is because block length of the TDC is longer than frame length of the ODC so that a DP-matching of the word decreases the accuracy.

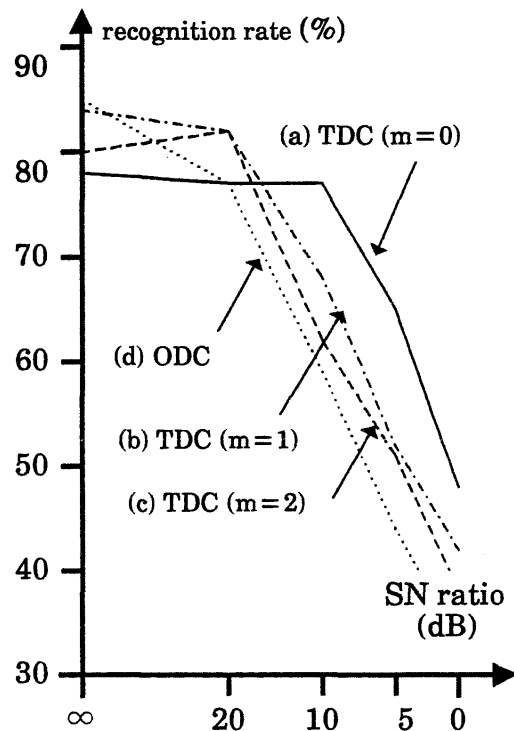


Fig.2 Recognition results for noisy monosyllables by three different filters.

Speech Recognition on TDC with High-enhancement Filter

Since the time variation provides important cues for the perception of consonants or vowel transitions, it is expected that the recognition rate should be improved by enhancing the dynamic information. We carried out the monosyllable recognition experiment by changing the value of the weight α on the TDC coefficient at $p=1$, holding the weight as 1 at $p=0$. The highest recognition rate is achieved when $\alpha=2$ at $p=1$. Weights greater than 2 decrease the recognition rate because it reduces the contribution of the static information comparatively.

Noisy Speech Recognition on the TDC with a Low-pass Filter

Since white noise dominates mainly on the regions E, F, F', B' and D' on the TDC in Fig.1, it is expected that the recognition rate of noisy speech should be improved by utilizing the TDC coefficients only on the regions B, C and D. This low-pass filter on the TDC enables two-dimensional smoothing on a time sequence of the spectrum(ref 2). Fig.2 shows the result of noisy monosyllable recognition as a function of the SN ratio ∞ , 20, 10, 5 and 0 dB by applying three different low-pass filters: (a) $m=0$, (b) $m=1$ and (c) $m=2$. The result by the ODC is also depicted. In recognition, the SN ratio is set to the same level between the input patterns and the standard patterns. From Fig.2, a high recognition rate is obtained by utilizing the dynamic information ($m=1$) at the high SN ratio, and by removing the dynamic information ($m=0$) at the low SN ratio. The utilization of the fine dynamic information ($m=2$) decreases the recognition rate at both the high SN ratio and the low SN ratio. Since the recognition rate by the TDC is greater than that by the ODC, a low-pass filter on the TDC is confirmed to be effective for spectral smoothing.

WORD RECOGNITION ON THE TDC WITHOUT DP-MATCHING

Linear Matching of Words on the TDC

The TDC can be extended so that each word has one block (one TDC). The lower component of the TDC, the region B, C, and D in Fig.1, includes information about the spectral envelope and

Table 3 Analysis condition for linear matching.

	linear	DP
analysis	frame length: 25.6ms the number of frames: 128	frame length: 25.6ms frame period: 12.8ms
recognition	parameter: TDC {C _{qp} , 0 ≤ q ≤ 14, 0 ≤ p ≤ m, (q,p) ≠ (0,0)}	parameter: ODC {C _q , 1 ≤ q ≤ 14}

Table 4 Recognition result by linear matching.

	recognition		CPU time	memory amount
	∞ dB	0dB		
linear	92%	79%	3 min	62 kB
DP	96%	50%	75 min	375 kB

its global time variation which are important features for word recognition. We use this lower component of the TDC, which is called a word matrix henceforth, as the standard pattern of the word because it contains the inherit features of the word. Recognition is performed by computing the distance between the word matrixes of the input pattern and output pattern (linear matching). The name of the standard pattern whose distance to the input pattern is minimum is the recognition result. The advantages of this method are:

- (1) Memory amount required for storing the standard pattern is small because the word feature is represented on the word matrix.
- (2) Recognition rate is high because the dynamic information is contained as well as the static information in the word matrix.
- (3) Recognition time is short because words are recognized by linear matching between the word matrixes.

Analysis Condition and Recognition Results

Table 3 shows the analysis condition of speech data. In order to obtain the TDC for each word, the frame length and the number of frames must be fixed for all words. Here, we fixed the frame length at 25.6 ms and the number of frames as 128. The word matrix is produced as the lower component of the TDC with $0 \leq q \leq 14$, $0 \leq p \leq 4$. In the Table, the analysis condition for a DP-matching is also depicted for comparison. Table 4 shows the recognition results by linear matching and DP-matching for the 100 Japanese city names spoken by two males and females. The linear matching shows a high recognition rate, though it does not exceed that by DP-matching. Under a noisy environment, the linear matching shows a high recognition rate. The recognition time and memory amount are reduced to 1/25 and 1/6 respectively compared to DP-matching.

CONCLUSION

In this paper, we clarified the effective range of the dynamic information analyzed by the TDC for word and monosyllable recognition. As a result, it is shown that the spectral envelope and its global time variation are important for recognition. It is also shown that the syllable recognition rate increases when the dynamic information is enhanced, and the recognition rate of noisy monosyllables increases with a low-pass filter on the TDC. For high speed recognition with a small memory amount, we proposed linear matching of the word matrix on the TDC, and confirmed its high recognition rate.

References:

1. S Furui, IEEE Trans. on ASSP, vol.ASSP-34, no.1, (1986)
2. Y Ariki, ICASSP, (1986)