

AN INFINITE-CONNECTED WORDS RECOGNITION SYSTEM FOR MALE SPEAKERS USING TIME-SPACE DYNAMIC PROGRAMMING

Ryu-ichi Oka
Pattern Processing Section
Information Sciences Division
Electrotechnical Laboratory
2-6-1 Nagata-cho, Chiyoda-ku
Tokyo 100, Japan

A new algorithm of pattern matching for time-space speech feature using Dynamic Programming is proposed for recognizing infinite-connected words spoken by male speakers without adaptation of speakers before the utterance. The new algorithm carries out frame-wisely two kinds of normalizations: the time normalization for removing speed variation of speech and the space normalization for removing speaker-dependence. This algorithm is realizable by a simple cellular automaton. The recognition rate 85.7% is obtained for 50 categories, 4 male speakers and 440 samples (connected words).

1. INTRODUCTION

The purpose of this paper is to propose a recognition system based on a new dynamic programming algorithm for recognizing continuous speech spoken by male speakers. This system is realizable as a simple cellular automaton. Recently, recognition systems dealing with isolated words spoken by many speakers are proposed [1], [2]. Other systems dealing with several connected words spoken by a speaker are also proposed [3], [4]. The following abilities are desirable for a system; 1) dealing with infinite connected words in an utterance, 2) allowing for many speakers without speaker adaptation before the utterance, 3) working in real-time. We propose a system with these abilities.

2. RECOGNITION PRINCIPLE

The principle is based on a new pattern matching algorithm using dynamic programming called Time-Space Dynamic Programming (TSDP). With regard to speech feature with the time and space axes such as spectrum and vocal tract area function, TSDP is applicable for normalizing two kinds of variations occurred in time-space axes of the feature. The one is time normalization for removing speed variation of speech and the other one is space normalization for removing speaker-dependence. Two normalization parts of TSDP are carried out frame-wisely without segmentation of a continuous speech. Segmentation problem of continuous speech is maintained to the final stage of recognition

process. Two kinds of normalizations of TSDP are separable.

3. OUTLINE OF THE SYSTEM

The recognition of words is carried out frame by frame. We can divide this frame-wise recognition process into three stages as follows.

- 1) Extraction of a vocal tract area function.
- 2) Application of Time-Space Dynamic Programming.
- 3) Frame-wise decision of category.

4. VOCAL TRACT AREA FUNCTION EXTRACTION

We use a vocal tract area function as a parameter. The reason is that a vocal tract area function is more stable for different male speakers than other parameters such as spectrum. The pre-experimental results revealed this conclusion.

The extraction technics of a vocal tract area function from speech waveform have already developed by Wakita [6] and Nakajima et al [7] by using inverse filter technic to remove glottal wave. We used an improved technic of Nakajima's one that is a frame-wise inverse filtering method of second order critical damping with a positive or a negative pole. The used one is a frame-wise inverse filtering method with the only positive pole. This method is more suitable because it does not carry out over inverse filtering for the frames in the consonant part of speech waveform.

The analysis condition is as follows ; 10 kHz

sampling, 20 msec window for analysis, 10 msec frame interval, 12 sections for a vocal tract area function and a vocal tract area function is normalized with regard to its volume.

5. TIME-SPACE DYNAMIC PROGRAMMING

5.1 Time Normalization Part

Let $\{f(t, x) : 1 \leq x \leq L\}$ be a vocal tract area function of the time t , where x indicates the space axis, i.e. vocal tract, and L is the length of vocal tract. Let $\{Z(\tau, x) : 1 \leq \tau \leq T, 1 \leq x \leq L\}$ be a standard pattern representing a word, where T indicates the length of the standard pattern and τ corresponds to the time axis of speech feature. We define a function $A(t)$ representing the temporary minimum distance between a past input pattern with the optimum length and a standard pattern, i.e.,

$$A(t) = \frac{1}{J^*(t)} \min \left\{ \sum_{j=1}^T K_j \left(\sum_{x=1}^L |f(t - \tau_j, x) - Z(\tau_j, x)| \right) \right\},$$

where $K_j = \sum_{j=1}^T \sum_{\tau_j=1}^T \sum_{x=1}^L |f(t - \tau_j, x) - Z(\tau_j, x)|$, $\tau_1 \geq \tau_2 \geq \dots \geq \tau_j \geq \dots \geq \tau_T \geq 0$, $\tau_1 \geq \tau_2 \geq \dots \geq \tau_j \geq \dots \geq \tau_T = 1$.

and $J^*(t)$ is the sum of K_j 's determined from the optimum pairs (τ_j^*, τ_j^*) from j equals one to j equals T . We will propose an algorithm for calculating $A(t)$ frame-wisely using Dynamic Programming. When we calculate $A(t)$ by the following method, we call it the time normalization part of TSDP. This idea is suggested by so-called Max-Product Parallel Machine proposed by Y. Isomichi [5]. Three variables are necessary for describing the time normalization part of TSDP. The variable $Q(t, \tau)$ indicates the local distance, $P(t, \tau)$ indicates the optimum value obtained from integration of the values of $Q(\cdot, \cdot)$'s, and $C(t, \tau)$ indicates an integrated value of K_j 's which are embedded in the value $P(t, \tau)$. The variable $C(t, \tau)$ is necessary for normalizing the length of the optimum pass.

$$P(0, \tau) = P(-1, \tau) = M \quad (1 \leq \tau \leq T)$$

$$P(t, 1) = 2Q(t, 1)$$

$$P(t, 2) = \min \begin{cases} P(t-2, 1) + 2Q(t-1, 2) + Q(t, 2) & (a) \\ P(t-1, 1) + 2Q(t, 2) & (b) \\ P(t, 1) + Q(t, 2) & (c) \end{cases}$$

$$P(t, \tau) = \min \begin{cases} P(t-2, \tau-1) + 2Q(t-1, \tau) + Q(t, \tau) & (d) \\ P(t-1, \tau-1) + 2Q(t, \tau) & (e) \\ P(t-1, \tau-2) + 2Q(t, \tau-1) + Q(t, \tau) & (f) \end{cases}$$

$$C(0, \tau) = C(-1, \tau) = 0 \quad (1 \leq \tau \leq T)$$

$$C(t, 1) = 2$$

$$C(t, 2) = \begin{cases} C(t-2, 1) + 3 & \text{if } (a) \\ C(t-1, 1) + 2 & \text{if } (b) \\ C(t, 1) + 1 & \text{if } (c) \end{cases}$$

$$C(t, \tau) = \begin{cases} C(t-2, \tau-1) + 3 & \text{if } (d) \\ C(t-1, \tau-1) + 2 & \text{if } (e) \\ C(t-1, \tau-2) + 3 & \text{if } (f) \end{cases}$$

where $Q(t, \tau) = \sum_{x=1}^L |f(t, x) - Z(\tau, x)|$ and M is the fixed value. Then the following equations hold.

$$P(t, T) = \min \left\{ \sum_{j=1}^T K_j \left(\sum_{x=1}^L |f(t - \tau_j, x) - Z(\tau_j, x)| \right) \right\}.$$

$$C(t, T) = J^*(t).$$

$$P(t, T) / C(t, T) = A(t).$$

The detail discussion about these derivations is described in a paper [8].

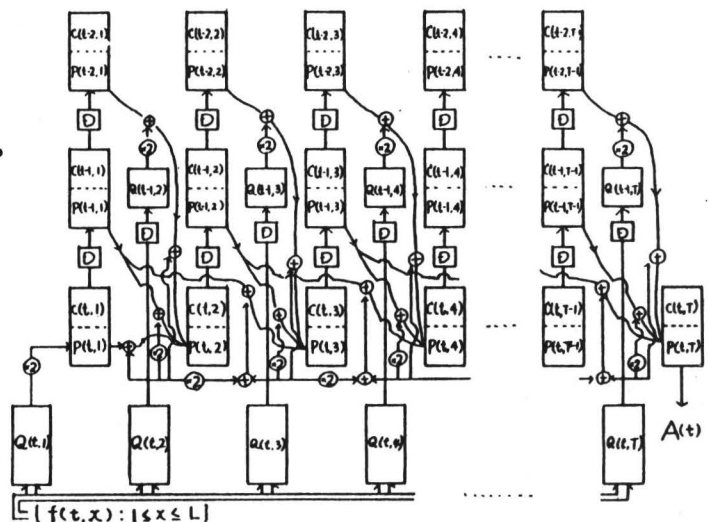


Fig. 1 Cellular Space Realization.

5.2 Space Normalization Part

The space normalization is applicable to the determination part of the value $Q(t, \tau)$ of the time normalization algorithm.

If we introduce a new variable $g(t, \tau, x, \xi)$ and suppose $1 \leq \xi-1 \leq x \leq \xi+1 \leq L$, then the space normalization algorithm is described as follows.

$$g(t, \tau, 1, 1) = 2 |f(x, 1) - Z(\tau, 1)|$$

$$g(t, \tau, x, \xi) = \min \begin{cases} g(t, \tau, x-1, \xi-2) + 2 |f(t, x) - Z(\tau, \xi-1)| \\ + |f(t, x) - Z(\tau, \xi)|, g(t, \tau, x-1, \xi-1) + 2 |f(t, x) - Z(\tau, \xi)| \\ , g(t, \tau, x-2, \xi-1) + 2 |f(t, x-1) - Z(\tau, \xi)| + |f(t, x) - Z(\tau, \xi)| \end{cases}$$

If the time normalization part has $Q(t, \tau)$ defined by $Q(t, \tau) = g(t, \tau, L, L) / 2L$, then the algorithm is called TSDP. The value $Q(t, \tau)$ indicates the local distance normalized with respect to the space x in $f(t, x)$. The following equation holds about the output of TSDP $A(t)$.

$$A(t) = \frac{1}{2L} \min_{j \in J^*(t)} \left\{ \sum_j K_j \left(\min_i \sum_l W_l |f(t - \zeta_j, x_i) - Z(\zeta_j, \xi_i)| \right) \right\}$$

$$(\zeta_j, \zeta_j)_{j=1,2,\dots,T}$$

where $K_j = \zeta_j - 1 - \zeta_j + \zeta_j - 1 - \zeta_j$, $W_l = x_l - x_{l-1} + \xi_l - \xi_{l-1}$, $x_1 = 1$, $x_L = L$, $\xi_1 = 1$, $\xi_L = L$.

6. CATEGORY DECISION ALGORITHM

Let $\{A_i(t): i = 1, 2, \dots, n\}$ be n values determined from TSDP, where i is the category number. Let t_F be the end time of a word in a continuous speech. We never use t_F in the recognition system. Section 5. reveals that the value of $A_i(t_F)$ is the smallest among those of $A_j(t_F)$, $j \neq i$, $A_i(t)$ ($t < t_F$), and $A_i(t)$ ($t > t_F$). We define $i_1^*(t)$, $i_2^*(t)$ so that $A_{i_1^*(t)}(t) = \min A_i(t) \leq \lambda$, $A_{i_2^*(t)}(t) = \min A_i(t) \leq A_{i_1^*(t)}(t) + h \leq \lambda$ for $i \neq i_1^*(t)$, where λ and h are the positive constants. Then we define the decision category number $i^*(t)$ by

$$i^*(t) = \begin{cases} i_1^*(t) & \text{if } L_{i_1^*(t)} \geq L_{i_2^*(t)} \\ i_1^*(t) & \text{if no } i_2^*(t) \\ i_2^*(t) & \text{if } L_{i_1^*(t)} < L_{i_2^*(t)} \end{cases}$$

where $L_{i_1^*(t)}$, $L_{i_2^*(t)}$ are the lengths of the standard pattern $i_1^*(t)$, $i_2^*(t)$. If $i_1^*(t)$ does not exist, we set $i^*(t) = 0$. The variable $i_2^*(t)$ is the candidate category number of the second minimum distance. In general, further algorithm is necessary since there is a time sequence whose frames indicate the same nonzero recognition category number.

7. EXPERIMENTAL RESULTS

Recognition experiments were carried out for 50 words making weather forecast sentences (25 Japanese-city names and 25 Japanese words about weather forecast). The standard patterns are isolated utterances of the words spoken by a male speaker who is out of the tested 4 speakers. The uttered connected words take a form of a sentence with natural speed. Three types of error existed, i.e., confusion: the system decides an incorrect category at the time when a word must be recognized, missing: the system decides no word at the time when a word must be recognized, ghost: the system decides a category at the time when no word must be recognized. Table 1 shows the two kinds recognition results of connected words (application of TSDP and the time normalization only of TSDP).

8. CONCLUSIONS

A new speech recognition system has been developed and evaluated from the viewpoint of continuous words recognition for male speakers

without segmentation and speaker adaptation. The recognition system is realizable as a simple cellular automaton. We obtained 85.7 % as the recognition rate for 440 connected samples, 4 speakers and 50 categories using the standard patterns spoken by a male speaker who was out of the tested speakers.

ACKNOWLEDGEMENTS

The author would like to thank Dr. S. Mori and Dr. T. Nakajima and the members of the Speech Recognition Section for supporting this work.

SPEAKER NAME	STAND. NAME	CONNECT SAMPLES	CORRECT	ERROR C M G	RATE
I.M	R.O	109	93	13 2 6	85.3 %
S.Ishi.	R.O	113	101	7 4 2	89.4 %
S.Ita.	R.O	111	97	10 3 3	87.4 %
T.S	R.O	107	86	11 8 3	80.4 %
TOTAL	R.O	440	377	41 17 14	85.7 %

TIME-SPACE DYNAMIC PROGRAMMING

R.O	R.O	500	466	15 19 28	93.2 %
9 males	R.O	48	30	5 12 7	62.5 %

ONLY TIME NORMALIZATION PART OF TSDP

Table 1

REFERENCES

- [1] Chiba, S., et al "A speaker-independent word recognition system." 4-th IJCPR, p.995, (1978).
- [2] Kido, K., et al "Spoken word system for unlimited speakers using gross pattern of spectrum and linguistic information." 4-th IJCPR, p.980, (1978).
- [3] Nakatsu, R., Kohda, M. "Speech recognition of connected words." 4-th IJCPR, p.1009, (1978).
- [4] Sakoe, H. "Recognition of continuously spoken words based on two level DP-matching." Paper of Tech. Com. of Speech, A. S. J., S75-28 (1975-11).
- [5] Isomichi, Y. "Segmentation-free recognition system for auditory pattern." Tech. Group of PRL, I.E.C.E.J., Paper PRL74-22 (1974).
- [6] Wakita, H. "Direct estimation of the vocal tract shape by inverse filtering of acoustic waveforms." IEEE Trans. Vol. AU-21, No. 5, pp. 417-427, Oct. 1973.
- [7] Nakajima, T., et al "Estimation of vocal tract area functions by adaptive inverse filtering methods." Bul. ETL, 34-4, p.50 (1973).
- [8] Oka, R. "Continuous words recognition by use of Continuous Dynamic Programming for pattern matching." Paper of Tech. Com. of Speech, A. S. J., S78-20 (1978-6).