【The invention】

[Discrete wavelet transform based multiple template-matching for speech recognition]

Inventor: Shinji Karasawa, Natori City, Japan

Patent Application No.: 2006-357183 Japan,
Date: Dec. 12, 2006

Keyword *Template matching, Haar wavelet transform, Speaker dependent speech recognition*

【Abstract】

This invention is method of speech recognition that consists of plural overlapped template-matching (TM). In order to economize calculations of TM, the data of low resolution given by Haar discrete wavelet transform (H-DWT) is used. Templates of wavelet coefficient (WC) on waveform are used for recognition of phoneme. A sum of WC in a scale (SWC) corresponds to a frequency component on the waveform. The templates of SWC on a waveform for short syllable are used for recognition of the syllable.

[Recognition on phoneme]

Three kinds of frame i.e. 64 of data (6.4msec), 128 (12.8msec) and 256 (25.6msec) were picked up from waveform where each segment begins at each peak of wave, and each segment of wave is transformed wavelet coefficients (WC) and the number of WC can be decreased up to 15 by the multiple resolution characteristics of WT.

A sum of absolute value of WC in each scale (SWC) indicates frequency components on the scale. That is, frequency distribution obtained from distribution of the SWC. The principal constituents of voices of a male are observed at scale on 6.4msec (156Hz), 3.2msec (312.5Hz), 1.6msec (625Hz), 0.8msec (1.25 kHz), and 0.4msec (2.5 kHz).

Since the ratio of SWC (the 3.2msec band)/SWC (the 1.6msec band) and SWC (the 6.4msec band)/SWC (the 1.6msec band) became a node (value=1) at the transition region of phoneme, segmentation of phoneme in continuous speech is able to find out.

[Recognition of syllable]

As for recognition of syllable, data size for a frame increases. Data on 204.8msec for a frame are shifted every 6.4msec were transferred to a set of WC and sum of absolute value of WC in each scale (SWC). (SWC) is used as a feature vector of TM on syllable.

[Overlapped TM for recognition of voice]

Overlapped TM is used for an integration of results. Any kind of TM is possible in a brain. A block diagram of layered TM system for presenting voice recognizer is shown in Fig.1.

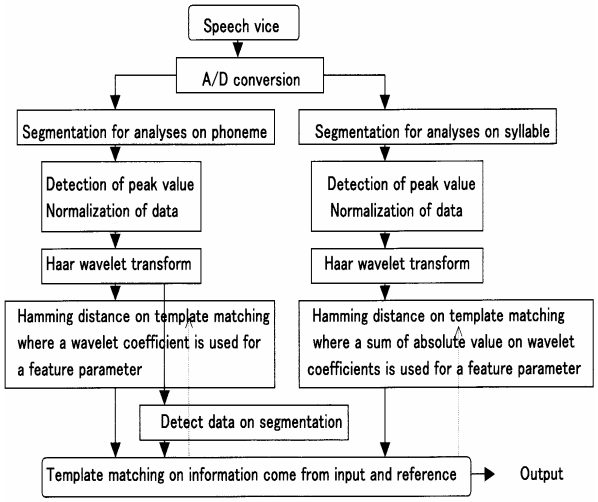


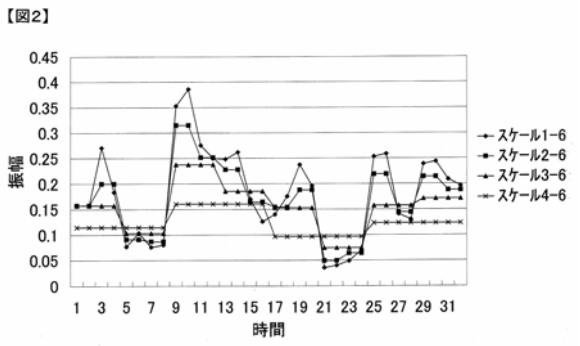
Fig.1 A speech recognition as an application of multiple-template matching

[Multi-resolution characteristics of H-DWT]

The main processing of speech recognizer is a TM. The calculations described in this document were carried out by means of Visual Basic for Application as a macro of Excel. The results are available to design a compact voice decoder.

Data with optimum resolution for feature vectors are obtained by H-DWT.

Fig.2 shows curves given by inverse wavelet transformation on the data in which higher resolution wavelets are omitted.



【Fig.2Multi-resolution characteristics of H-DWT

【Summary】

[Background of the invention]

There are many statistics-based speech recognition systems. These traditional systems are to make use of universality on voice recognition and to construct speaker independent voice recognition system.

However, mobile computer and mobile phone have become popular. The compact speaker dependent voice decoder gets the demand.

The intelligent faculties of our brain are carried out by means of the faculty of neurons. That is

the impulsive template matching. The basic intelligence has been implemented heuristically. When we use language, multiple overlapped impulsive activities occur in our brain. The meaning of the reaction depends on the circumstance and situation.

[Segmentation on vibration of vocal cord]

A waveform of vowel possesses a segment due to vibration of vocal cord. The pitch on one segment is detected by the time difference between nearest neighboring peak points. A starting point of waveform for a template is given by the peak.

Fig.3 shows waveform of Japanese vowels [a, i, u, e, o] uttered continuously by a male.

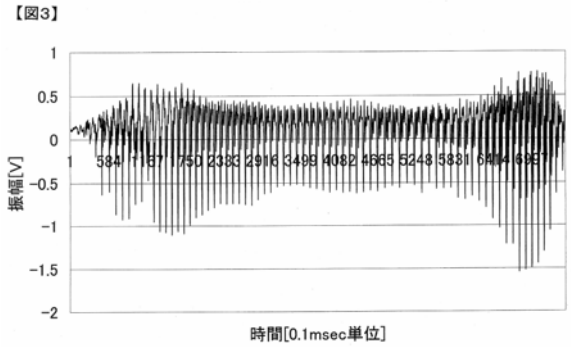


Fig.3 Wave form of Japanese vowels a;[あ], i;[い], u;[う], e;[え], o;[お]

[Time variations obtained by H-DWT]

One processing unit of waveform is translated to a set of wavelet representations of WC. The scaling function in H-DWT corresponds to a time slit. The values of WC in each scale are arranged according to the progress of time.

[Frequency characteristics in H-DWT]

A sum of absolute value of WC in each scale (SWC) indicates frequency components of the scale. Then frequency distribution obtained from distribution of SWC. The principal constituents of speech voices of a male are time scale on 6.4msec (156Hz), 3.2msec (312.5Hz), 1.6msec (625Hz), 0.8msec (1.25 kHz), 0.4msec (2.5 kHz).

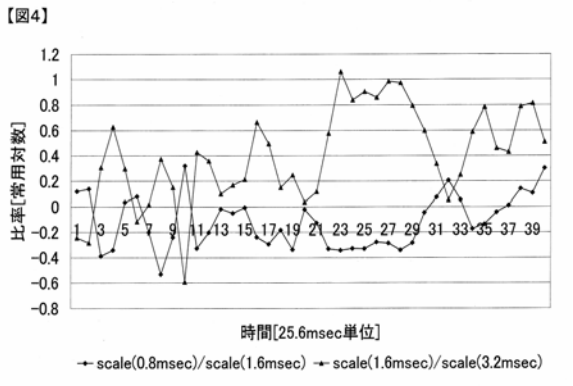
[Segmentation for waveform in H-DWT]

Segmentation of phoneme in continuous speech can be checked according the following facts.

Fig.4 shows the ratio of SWC (the 3.2msec band)/SWC (the 1.6msec band) and SWC (the 6.4msec band)/SWC (the 1.6msec band). The ratios on SWC form a node in the transition region of vowel.

The short period indicated by cross point of 3rd and 4th corresponds to non-vocalic utterance of [i].

The manner of articulation on [i] is the same but the sound in this case is fricative that is generated in oral cavity.

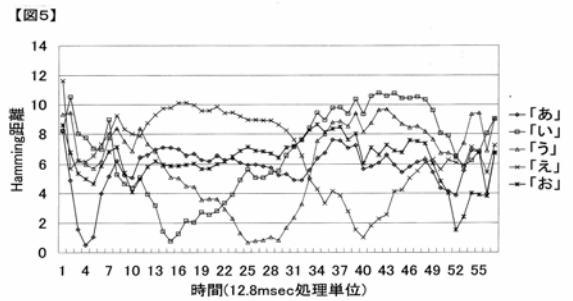


【Fig.4】Fig 4 Ratio on principal SWC on the wave shown in Fig.2

[TM on WC for discrimination of phonemes]

The maximum value of amplitude in every frame is normalized as is [1]. The sampled digital data on waveform are transformed to the same amount of sets on WC, i.e. 128 points of sampling rate of 0.1m sec data are transferred to 127 point of WC and one average value. The representations of WC are used for TM.

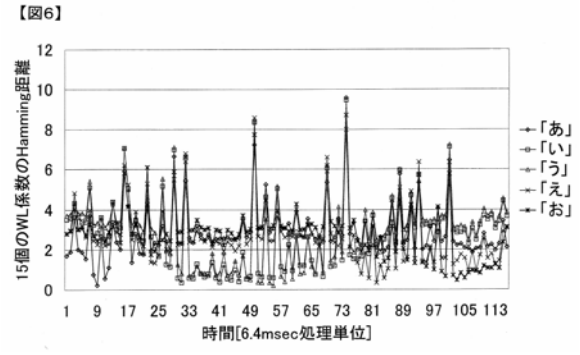
Fig.5 show Hamming distances on 128 points of WC. Although, the data on template and those to be referred are obtained from the same wave show in Fig.3, the segmentation is a little different, because starting point on template is assigned by manual and that on object is assigned automatically.



【Fig.5】Hamming distances of TM on WC of vowels a;[あ], i;[い], u;[う], e;[え], o;[お] where templates are obtained from the object.

[Effects of resolution for phoneme recognition]

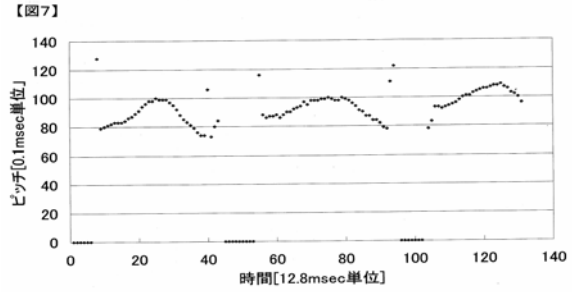
H-DWT is effective to economizes the calculations of TM reduce calculations. In order to compare the differences of resolution, Fig.6 shows Hamming distances on 15 points of WC on the same case shown in Fig.5. The number of data for a frame affects the sensitivity of TM.



【Fig.6】TM on 15 pieces of WC where resolution is decreased 1/8 from the case shown in Fig.5

[Recognition of phoneme on Japanese voice]
[Size of template and segmentation]
A set of many data is possible to possess many variations. The data-matching system that deals with large size of template needs many kinds of templates. The increase of constituents on a template increase calculations for the TM. A short part of waveform possesses similarities. That is, shorter segment of waveform is universal. It is better for the processing of TM that the size of template is small.

[Deviation on length of pitch]
Length of pitch varies during utterance, as shown in Fig. 7,



【Fig.7】Variation of pitch length on a-a;[あーあ], i-i;[いーい], u-u;[うーう]

[Analyses on vowel]
By using small size of waveforms as templates, those are picked up 6.4msec in the front part of wave with pitch 9.5msec in case of shown in Fig.7. Those 64 pieces of data are transferred to 15 pieces of WC. Continuously uttered vowel voices shown in Fig.6 are analyzed by TM method. Fig.8 shows the results of TM on WC of a;[あ], i;[い], u;[う], e;[え], o;[お] .

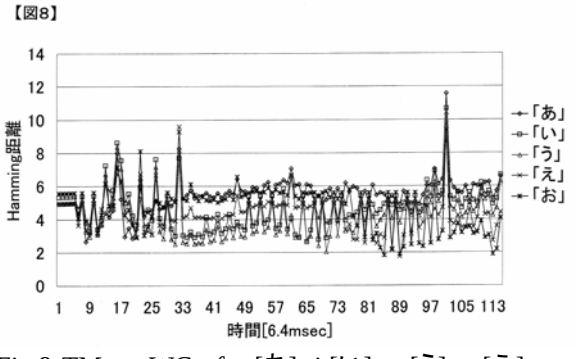
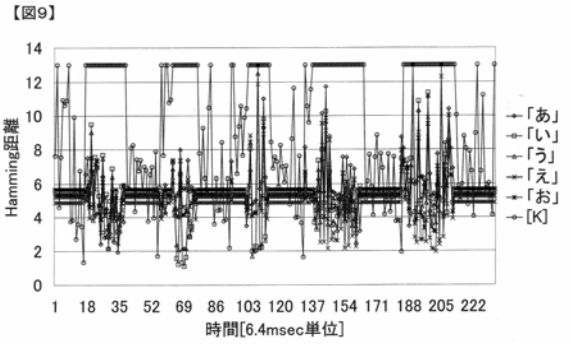


Fig.8 TM on WC of a;[あ], i;[い], u;[う], e;[え], o;[お]

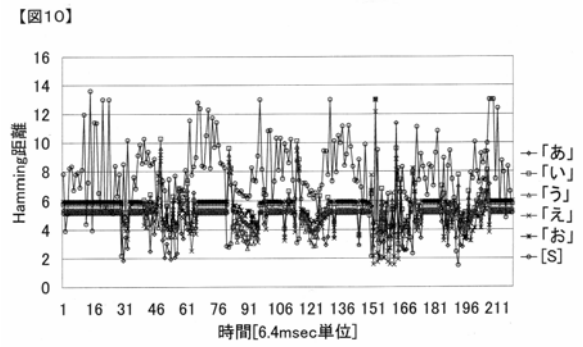
Here, [i] can not recognize in Fig.8. The reason of this phenomenon depends on the utterance between the case of Fig.6 and that of Fig.7. Voice [i] in continuous utterance of vowel shown in Fig.4. is different from that of vowel shown in Fig.7.

[Analyses on consonant]
Feature of consonant is characterized by the manner of utterance. The movements of utterance affect whole waveform. The feature of consonant must be described by whole syllable.
As a trial, beginning part of each syllable is used as a template of consonant, because almost all end part of syllable on Japanese [mora] is vowel. That is, template of consonant [K] is picked up in front of each Japanese mora as a trial.
Fig.9 shows the results of TM. Here, wave form of Japanese mora (i.e, ka;[か], ki;[き], ku;[く], ke;[け], ko;[こ]) uttered by a male, where sampling period is 0.1msec. Those 64 pieces of data are transferred to 15 pieces of WC. Here, the templates on vowels are picked up in the case of shown in Fig.7 and that of [K] is picked up from in front of waveform of [ki].



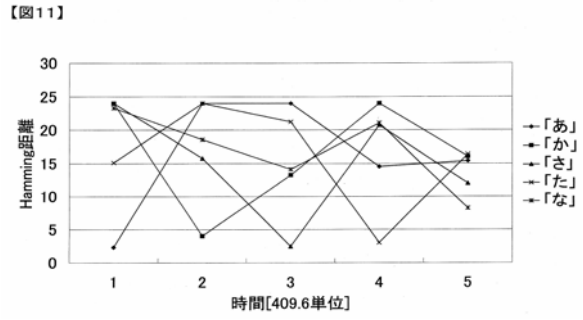
【Fig.9】Hamming distances of TM on ka;[か], ki;[き], ku;[く], ke;[け], ko;[こ], where 15 pieces of WC are referred as a feature vector.

Fig.10 shows the results of TM on sa;[さ], shi;[し], su;[す], se;[せ], so;[そ]. These results are obtained the same procedures. But, template of [S is picked up from in front of waveform of [shi].



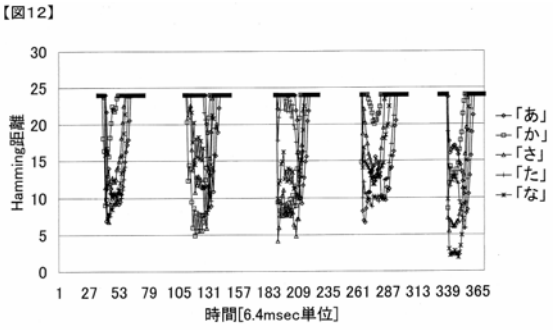
【Fig.10】Hamming distances of TM on sa;[さ], shi;[し], su;[す], se;[せ], so;[そ] where 15 pieces of WC are referred as a feature vector.

[Recognition of syllable on Japanese voice]
[Segmentation for template]
Syllables in a voice were analyzed by means of 409.6 msec of template as long syllables. Here, SWC were as used as a feature vector of a frame for Japanese mora of a;[あ], ka;[か], sa;[さ], ta;[た], na;[な].
The value of SWC corresponds to that of frequency constituent in a Japanese mora. Fig.11 shows hamming distances among frames where the waveforms for template are obtained from the wave to be analyzed. But SWC on long syllable uttered different time can not distinguish the syllables.



【Fig.11】Hamming distances of TM on SWC where long syllable is used as template (Template of syllables are obtained from the wave to be checked)

[Short syllable is fit for template]
A short syllable has universalities more than long syllable. The sorter syllable is useful for template. 204.8msec waveforms of Japanese mora a;[あ], ka;[か], sa;[さ], ta;[た], na;[な] are picked up for templates on syllable. Those waveforms were transformed into WC, and SWC are obtained from WC. Fig.12 shows the hamming distance on TM of SWC.



【Fog.12】Recognition of long syllable of Japanese mora i.e. a;[あ], ka;[か], sa;[さ], ta;[た], na;[な]. Here, short syllables are used as templates,

【Items clamed】

【Claim 1】
Recognition of phoneme through the template matching where differences on discreet wavelet coefficients are calculated for the evaluation of the template matching. Here the processing unit of waveform for phoneme is smaller than pitch length of waveform, and the peak value is set to 1 for normalization.

【Claim 2】
Recognition of syllable through the template matching in which the differences on sum of absolute value of discreet wavelet coefficients in each scale is calculated, where size of template of syllable is smaller than that of syllable to be recognized and the peak value in every segment is set to 1.

【Claim 3】
Segmentation in continuation voice for phoneme recognition by utilizing the fact that ratio of sum of absolute value on discreet wavelet coefficients on principal scale becomes near 1at transition of phoneme.

【Claim 4】
Recognition of speech by using distances on the template matching in which the data on template matching is the results on claim1, claim2 and claim3.

【Claim 5】
The software system or electronic circuit those make use of the function of claim1, claim2, claim3 and claim4.