Use of Haar Wavelet Transform Based Multiple Template Matching for Analyses of Speech Voice

Shinji Karasawa and Hiroshi Sakuraba
Miyagi National College of Technology
E-mail: shinji-karasawa@cup.ocn.ne.jp sakuraba@miyagi-ct.ac.jp

Abstract

Technologies of wavelet transformation were used in JPEG2000 and those will be available for CODEC. Pivotal reminders for voice recognition were investigated by using multi-resolution of Haar wavelet representation (H-WR). Template of a phoneme differs from that of a syllable. Optimum accuracy of the feature depends on segmentation of template-matching (TM) analyses.

64 components of Haar wavelet coefficients (H-WC) for recognition of a phoneme are able to decrease to 15 components with lower frequency. Here, each set of data begins at peak value in each pitch. Sampling frequency is 10 kHz. The period of segment for a phoneme is 6.4msec. Segmentation of phoneme in speech can be checked by using the fact that ratio (r) between SWC (sum of absolute value of WC in a scale) becomes r=1, at a transition.

SWC is available as a constituent in vector quantization for a syllable. Short syllables are decoded by means of 8 pieces of SWC, here the SWC was obtained from a set of data of 1024 pieces on a syllable (sampling frequency is 5 kHz, period of extraction for a syllable is 204.8msec).

Keywords: *Data-compression, Haar discrete wavelet transform, CODEC, Template matching.*

1. Introduction

A compact speaker dependent voice decoder has got demand, because mobile computer and mobile phone have spread widely. On the other hand, it is possible to analyze voice by using template matching (TM). The results are available to design voice decoder or CODEC.

The information is recognized by activities of new cortex in a brain [1]. The basic intelligence is implemented through activities in a real world individually [2]. When each cell acts in order to supply the demand of surroundings, the group is organized. Meaning of each action depends on the circumstance and the situation. The action of neuron is ignited through TM.

Although there are many reports on speech recognition, the statistics-based speech recognition has been developed in the field of technology. There are reports on the use of wavelet transforms in speech recognition. Those are, statistical model-based voice activity detection algorithm in the wavelet domain [3], an application of the merging algorithm with the discrete wavelet transform to extract valid speech-sound [4]. Thai phoneme segmentation using discrete wavelet transform [5], wavelet based feature extraction for phoneme recognition [6], uses of wavelet transforms in phoneme recognition [7]. Those traditional studies have been carried out to construct speaker independent voice recognition system. On the other hand, uses of H-WT for non-statistics treatments of voice are reported in this paper.

2. Multi-resolution characteristics of H-WC

The scaling function of H-WT corresponds to a time slit. Fig.1.shows 64 directions of original wave form and (N-1) = 63 directions of H-WC as an example of H-WT.

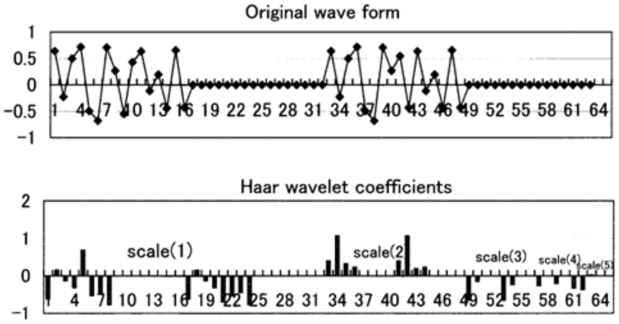

Figure 1. Multi-resolution characteristics of H-DWT

A data compression is realized through omitting of H-WC in lower scales. Various curves shown in Fig.2 are given by inverse wavelet transformation of the H-WC.

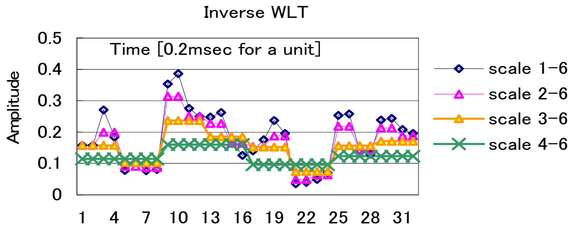

Figure 2. Multi-resolution characteristics of H-DWT

2.1. Frequency distributions expressed by sum of absolute values of Haar wavelet coefficients

A segment of waveform for a TM is transferred to one set of wavelet representations of WC.

Fig.3 shows waveform of voice [A B C] uttered by a male.

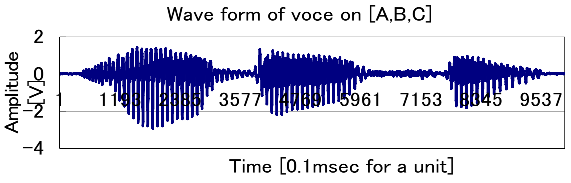

Figure 3. Wave form of voice uttered as A, B, C.

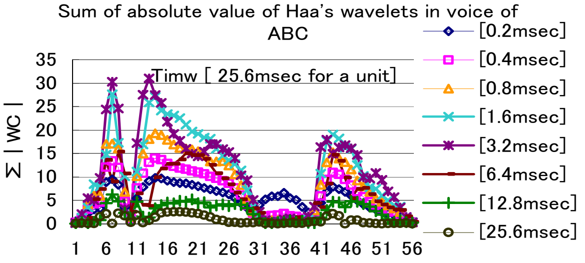

Figure 4. Frequency constituents obtained from SWC

Fig.4 shows SWC, those were obtained from the wave shown in Fig.3. Those curves indicate frequency characteristics. Principal constituents of speech voices are 3.2msec (312.5Hz), 1.6msec (625Hz), 0.8msec (1.25kHz), 0.4msec (2.5kHz), 6.4msec (156Hz).

2.2. Detection of signal on transition of phoneme

Fig.5 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 (r) form a node (r=1) in the transition region of vowel.

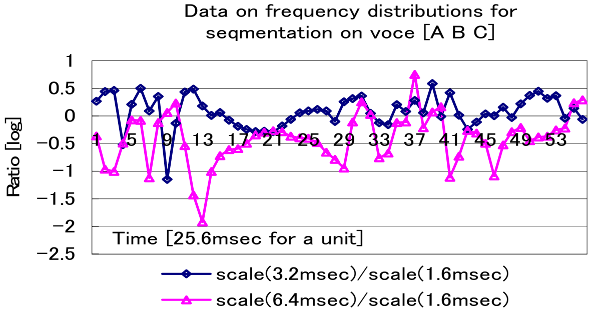

Figure 5. Ratio on principal SWC on the wave shown in Fig.3

Transition on phoneme in continuous speech is detected by the ratio among principal SWC as shown in Fig.5.

3. The template matching that consists of WC for phoneme representation

Fig.6, Fig.8 show wave form of vowel [a-i-u-e-o] and mora [ka ki ku ke kou] uttered by a Japanese male.

The data on waveform are transformed to sets of WC, and the sets of WC are used for TM. Fig.7, Fig.9 show results on TM where templates are obtained from wave itself. Here, the data are picked up from peak in each pitch, sampling frequency is 10kHz. The maximum value of amplitude in every frame is normalized as V(max)=1.

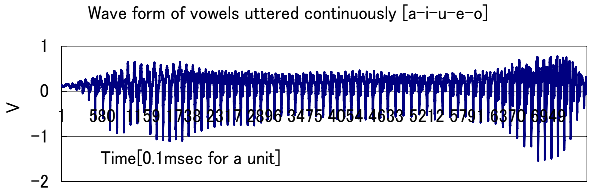

Figure 6. Wave form of vowels [a-i-u-e-o] uttered continuously

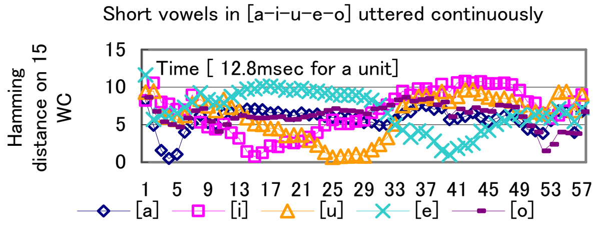

Figure 7. TM on WC between [a-i-u-e-o] and vowels

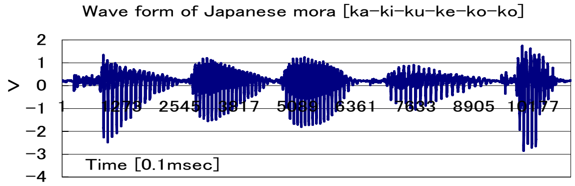

Figure 8. Wave form of Japanese mora [ka ki ku ke ko]

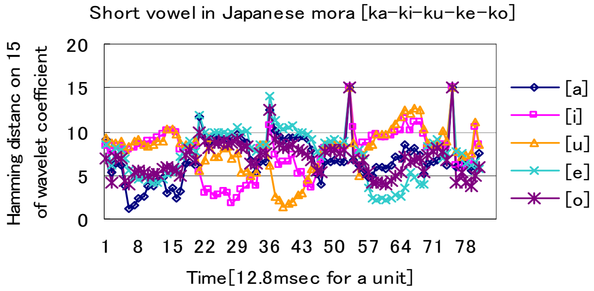

Figure 9. TM on WC between mora [ka, ki, ku, ke, ko] and vowel [a,I,u,e,o]

Here, 128 pieces of data for a frame were transferred to 15 pieces of WC. H-WR is effective to economize the calculations of TM calculations.

3.1. Shorter segmentation for a template

The shorter segmentation of TM provides the more universal template, but it loses sharpness of selectivity. Fig.10 shows TM on 15 pieces of WC those data obtained from 64 pieces (6.4msec) of sampled data.

It becomes clear by comparing Fig.9 and Fig.10 that segmentation of H-WC for a frame on phoneme is able to decrease from 128 (12.8msec) to 64 (6.4msec).

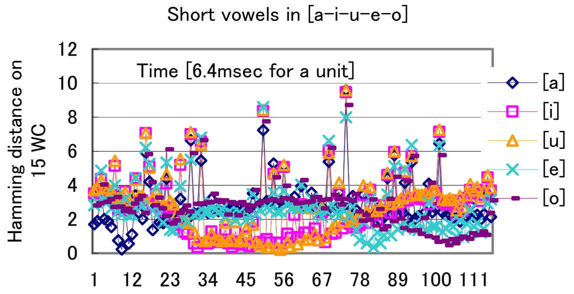

Figure 10. TM on WC between [a-i-u-e-o] and vowels

3.2. Variations of length of pitch on long vowel

Length of pitch of long vowel varies during utterance as shown in Fig.11. This fact causes a difficulty for TM.

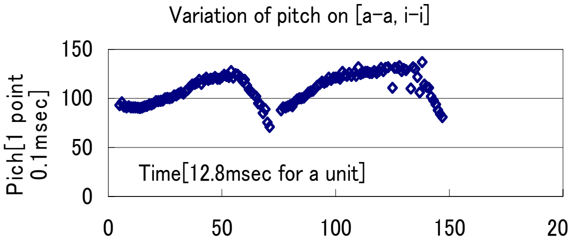

Figure 2. Variation of pitch length on [a-a, i-i]

4. TM on WC for analyses of phonemes

Phonemes in Japanese mora were investigated by using vowels uttered continuously as templates. Here, the period of template is selected 6.4msec in order to get universal templates of vowels. Templates in this section are picked up from continuous vowels with pitch of 9.5msec, of which parts are indicated in Fig.11. Fig.12 and Fig.13 shows the result of TM analyses on phonemes. There, inputs to be referred obtained from each peak in a pitch and 64 pieces of data (6.4msec) were used for a frame, and those were transferred to 15 pieces of WC.

4.1. Detection of different pronunciation by TM

There are differences of utterance on the same phonetic symbol. TM reveals such differences of pronunciations. Japanese insert a break for utterance of short vowel [i] in a continuous utterance. Such fact was indicated in TM, as shown in Fig.12.

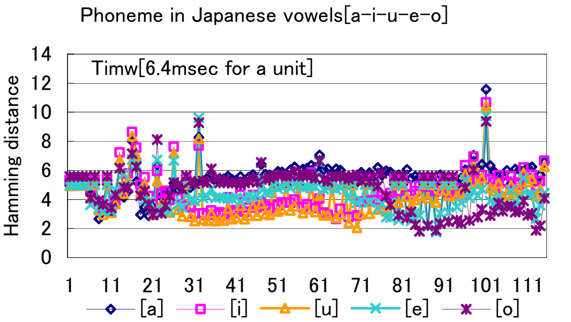

Figure 12. TM of WC on [a-i-u-e-o], where input voice [i] is uttered as a short break

4.2. Difficulty of recognition of consonant by TM

The feature of consonant is represented by whole of utterance. It is not represented only at front part of utterance. As a trial of Fig.13, a candidate waveform of consonant of [K] was picked up in front of Japanese mora. Although TM changes by utterance, minimum of distance of TM possesses universalities. Then, we can say from the results that voice recognition is possible.

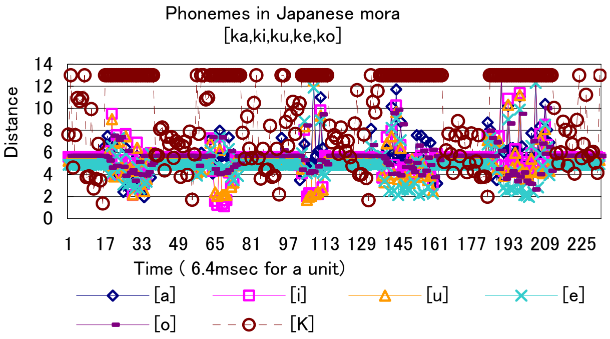

Figure 13. TM of WC between [ka,ki.ku.k.kou] and [a,i,u,e,o,K]

5. TM by SWC for analyses on short syllables

Waveforms of 204.8msec are picked up for templates on a short syllable and those were transformed into WC. And, SWC that corresponds to that of frequency constituent was used as a feature vector of a frame.

Fig.14 shows the hamming distance on TM of SWC.

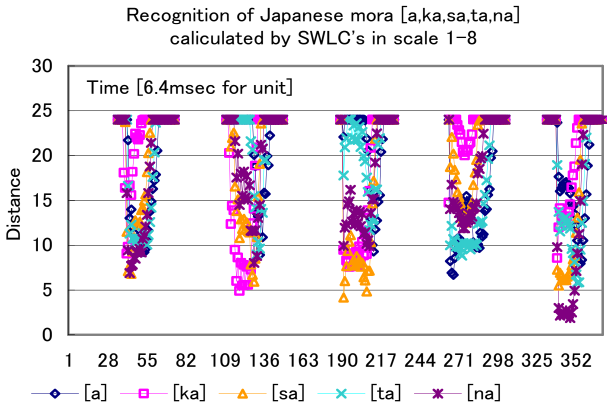

Figure 14. Recognition by means of short syllables as templates

SWC template that made of long syllable of 409.6 msec is not able to distinguish short syllables. Template of short syllable has universalities more than that of long syllable.

5.1. Combination of overlapping recognition

There are more sentences than words, and there are more words than phonemes. So, results of TM on shorter segmentation should be used to omit improbable nominee in larger segmentation. That is, analyses of phoneme are used to economize the analyses of syllable. Any kind of TM is possible in a brain. Digital technologies in a finite state machine are available to construct such speech recognition system. A block diagram of multiple TM system for a voice decoder is shown in Fig.19.

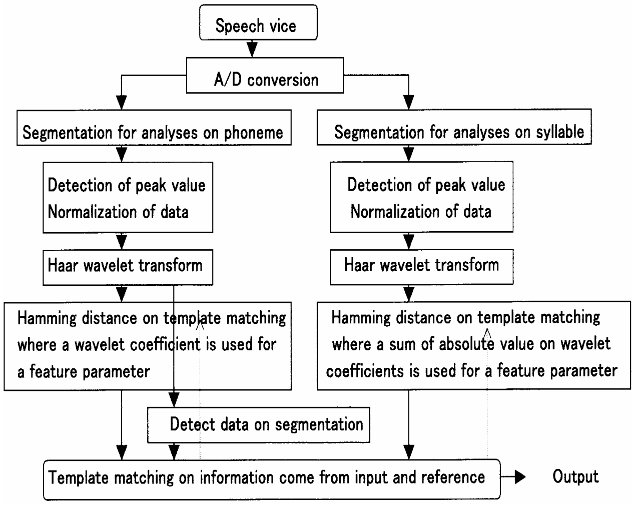
Speech vice
A/D conversion
Segmentation for analyses on phoneme
Segmentation for analyses on syllable
Detection of peak value
Normalization of data
Detection of peak value
Normalization of data
Haar wavelet transform
Haar wavelet transform
Hamming distance on template matching where a wavelet coefficient is used for a feature parameter
Hamming distance on template matching where a sum of absolute value on wavelet coefficients is used for a feature parameter
Detect data on segmentation
Template matching on information come from input and reference
Output

Figure 15. A speech recognition as an application of Multiple TM

6. Conclusions

An ultimate data-compression is recognition. The faculty is implemented through experience of activity. Moreover, TM is a powerful tool to analyze cognitive function. TM depends on the segmentation of template. We can detect the sign on segmentation of a frame. The pattern of low resolution given by H-WT economizes calculations of TM. Simple calculation of H-WT is attractive, and flexibility of H-WT on data compression for speech recognition is attractive. There is the possibility that H-WT can be used as a tool for data compression for a portable telephone.

7. References

S. Karasawa, "Brain Mechanism on Understanding of Information Explained by Concept of Activity", *IEICE Technical Report*, TL2007-2, ISSN 0913-5685. 2007.

S. Karasawa, "Attributes of Language Use Explained by Activities of Neuron", *IEICE Technical Report*, TL2006-11, ISSN 0913-5685, 2006, pp.31-36.

Y.C. Lee, S.S.Ahn, "Statistical Model-Based VAD Algorithm with Wavelet Transform", *Proc. IEICE Transaction on Fundamentals of Electronics, Communications and Computer Sciences*, E89-A (6) 2006, pp.1594-1600.

J.O. Kim, et al. "On the Extraction of the Valid Speech-Sound by the Merging Algorithm with the Discrete Wavelet Transform", *Inter. Conference on Computational Science*, 2003, pp.619-628.

B.Thipakom, B. Kaewkamnerdpong, "Thai Phoneme Segmentation using Discrete Wavelet Transform", *International Journal of Smart Engineering System Design*, Vol 5, No.4, 2003, 389-399.

C.J.Long, S.Datta, "Wavelet Based Feature Extraction for Phoneme Recognition", *Inter. Conference on Spoken Language Processing*, 1996.

B.T.Tan, M.Fu, A.Spray, F.Dermody, "The Use of Wavelet Transforms in Phoneme Recognition", *Inter. Conference on Spoken Language Processing*, 1996.

[1],[2]URL;http://www17.ocn.ne.jp/~shinji-k/index.htm
The calculations in this report were carried out by means of Visual Basic for Application as a macro of Excel.
Software on Haar wavelet transform was referred to Y. Okawa "Mathematical processing for abstraction of characteristics on wave from, ISBN4-7898-3715-7,pp.66".
Software on A/D conversion was referred to Y. Okawa "Excel for data processing, ISBN4-7775-1010-7,pp.182".