Speech Feature Extraction Based on the Human Hearing Model

Kwang Woo Chung*, Paul Kim**, Kwang Seok Hong***
* Dept. of Operation-Mechatronics, KOREA Railroad College.
** R&D Center of SINDORICOH.
*** Dept. of Electronics Engineering, Sung Kyun Kwan University.
e-mail : kshong@yurim.skku.ac.kr

Abstract

In this paper, we propose the method that extracts the speech feature using the hearing model through signal processing techniques. The proposed method includes the following procedure ; normalization of the short-time speech block by its maximum value, multi-resolution analysis using the discrete wavelet transformation and re-synthesize using the discrete inverse wavelet transformation, differentiation after analysis and synthesis, full wave rectification and integration.

In order to verify the performance of the proposed speech feature in the speech recognition task, korean digit recognition experiments were carried out using both the DTW and the VQ-HMM. The results showed that, in the case of using DTW, the recognition rates were 99.79% and 90.33% for speaker-dependent and speaker-independent task respectively and, in the case of using VQ-HMM, the rate were 96.5% and 81.5% respectively. And it indicates that the proposed speech feature has the potential for use as a simple and efficient feature for recognition task.

I. Introduction

Human voice is transformed into electrical signal through ear, transferred to brain, and recognized. Many research papers on the modeling and applying of this procedure have been published.[1][2][3][4] In this paper, we apply the existing hearing model[3][4] to procedure of feature extraction using discrete Wavelet transformation, and propose a new method for speech feature extraction. The proposed method includes differential equation, discrete Wavelet transformation, and rectification.

In order to verify the usefulness of the proposed speech feature in speech recognition, korean digit recognition experiments were performed using DTW and

VQ-HMM algorithm. The results show relatively high recognition rate in comparison with simplicity of the algorithm.

II. Hearing Model and Discrete Wavelet Transform

1. The Structure and Operation of Ear

The internal structure of ear is shown in Fig 1, and main organs concerned with speech transfer are shown in Fig 2. The ear is largely composed of Outer Ear, Middle Ear, and Inner Ear. Among them, Inner Ear has very complicated structure, and corresponds to the part for voice frequency analyzation and feature extraction. Cochlear in Inner Ear plays very important role in speech recognition process. Its internal has Auditory organ called the Organ of Corti. The Organ of Corti is composed of Basilar Membrane, which shows the highest change responding to a special frequency of the transferred voice in the form of vibration, Hair Cells sensing mechanical position change and transforming it into electrical signal, and Cilia transferring the degree of position change to Hair Cells.[3][4] The voice recognition procedure is simply as follows.

Voice signal transferred in air vibration vibrates Eardrum via ear, and this vibration is transferred to Cochlear after amplified by Ossicles . The position change of Basilar Membrane , which is in Cochlear and plays the role of a kind of band-pass filter in that it responds to a specified frequency by the position, bends Hair Cells and nonlinearly controls the ion inflow to Hair Cells, so induces electrical signal to Hair Cells. While mechanical position change is transformed into electrical signal and goes through Auditory Nerves, various information is extracted and transferred to brain.[3][4]

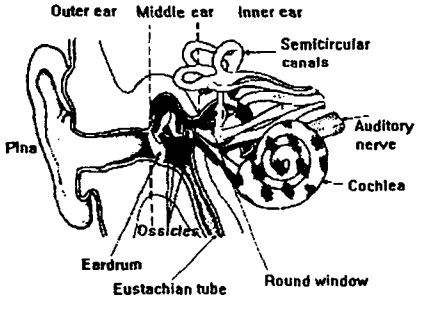


Fig 1. Ear structure

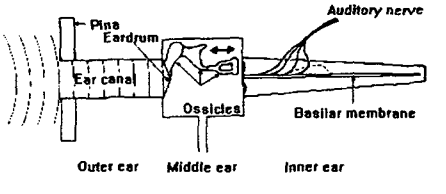


Fig 2. Simplified model of ear structure

2. Ossicles

Ossicles exists in Middle Ear, and it is a kind of mechanical transformation device concatenated with Malleus, Incus, Stapes in oder. The delicate vibration of the eardrum is amplified about 30 times and transferred to Cochlear in Inner Ear. Besides this function, Ossicles protects Inner Ear from very high voice or sharp change of pressure, and transfers pressure constantly. This is the preprocess, which keeps voice signal in constant level and amplifies voice in order to emit stabilized analyzation output.[3] Therefore any other appropriate method which can make high voice lower and low voice higher is good for this function. In this paper, we find a narrow band maximum and carry out normalization based on it as shown in Fig 3. As a result, stabilized outputs for too high or low inputs can be ensured when we extract feature using discrete Wavelet transform.

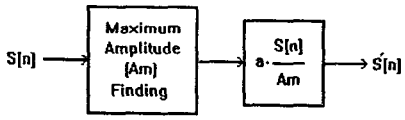


Fig 3. Block diagram of normalization process

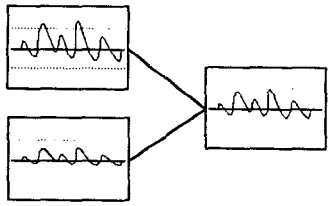


Fig 4. Functional diagram of normalization process

3. Basilar Membrane and Discrete Wavelet transform

Basilar Membrane exists in Cochlear, its base part near Oval window is narrow and hard ; the part nearer to the edge is more loose.[5] Due to its shape, the influence of voice frequency transferred through Ossicles shows the maximum position change in about 100Hz near the edge, in 10,000Hz or higher near the base part, and in 2,000Hz in the middle point of the edge and base part.[6]

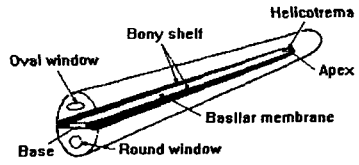
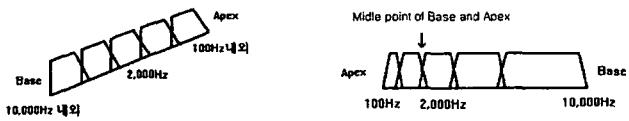


Fig 5. Basilar membrane



(a) (b)

Fig 6. Frequency response of basilar membrane

In this way, Basilar Membrane can act as a filter bank which analyzes each frequency band of voice signal into multi-resolution.[3] If we divide Basilar in Fig 5 in constant lengths of intervals, it is like Fig 6(a), and its frequency axis is assumed to be log-scaled. When the frequency axis of Fig 6(a) is linearized, it will seem to be Fig 6(b). This feature of Basilar tells us that human auditory system is sensitive to low frequency and is relatively less sensitive to high frequency. Also, when we analyze voice frequency, we can see that more information is concentrated in low frequency than in high frequency. Therefore, we can effectively analyze voice if we use high frequency-resolution filter for low frequency and low frequency-resolution filter for high frequency.

In this paper, we implemented this feature of Basilar by using discrete Wavelet transformation which carries out frequency multi-resolution analyzation and inverse discrete Wavelet transformation. Because of the feature that discrete Wavelet transformation carries out decimination continuously so information decreases when it goes to low frequency, we can perform speedy and effective Octave band filtering only by limited calculation. Mallat's Subband Coding Scheme, a kind of discrete Wavelet transformation, is shown in Fig 7.[8][9][10]

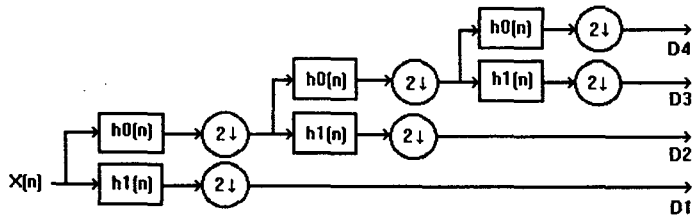


Fig 7. Discrete Wavelet Transform(DWT)

Where h_0, h_1 are Wavelet bases, and have the following equation.

$$h_1(n) = (-1)^{n+1} h_0(L-1-n) \qquad \text{(Eq. 1)}$$

where, L is the length of Wavelet Basis

When it is concerned with high-order Wavelet Basis, implementation speed is low but regularity is high.[9] So, it shows better filtering result. In this paper, we used Daubechies's 20-order Orthogonal Wavelet Basis with complete reconstructive condition. Inverse discrete Wavelet transformation, which synthesizes signal analyzed in multi-resolution, is show in Fig 8.

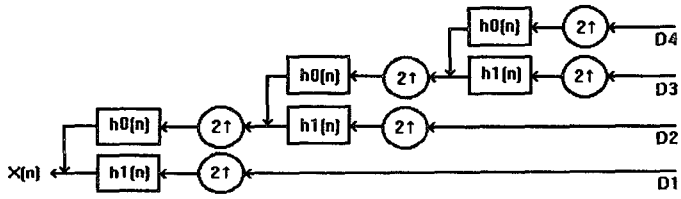


Fig 8. Inverse Discrete Wavelet Transform(IDWT)

DTW algorithm in itself has a continuous decimation process, so the total number of calculation is limited to an Equation like 2.[9] Therefore, 2 times or less calculation number of input samples is enough to effectively analyze signal into frequency multi-resolution.

$$C_{total} = C_0 + \frac{C_0}{2} + \frac{C_0}{4} + \ldots\ldots < 2C_0 \qquad \text{(Eq. 2)}$$

where, C_0 is the number of input signal samples.

The procedure of analysing and synthesizing voice signal into frequency multi-resolution using above mentioned discrete Wavelet transformation, and inverse discrete Wavelet transformation is mapped out in Fig 9.

Fig 9. Wavelet Filter bank

While the frequency multi-resolution analysis of voice signal is possible only by the method shown in Fig 7, differentiation is very difficult because the number of output samples in each frequency band differs due to decimination and hearing modeling is time-variant. So, in this paper, we obtain as many time-varying signals analyzed into frequency multi-resolution as the number of filter banks with the same samples.

The ideal result is shown in Fig 10 when Octave band pass filter is implemented by above mentioned method where the implementable number of filter banks is limited as like Equation 3 by the decimination feature of Tree algorithm.

The number of Filter Bank = $\log_2$ (The nunber of Analized Frame) + 1 (Eq. 3)

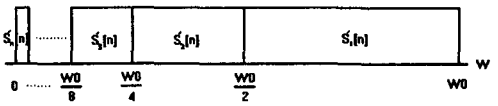
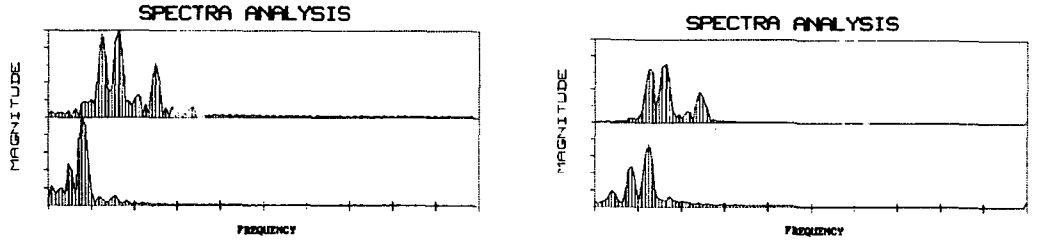


Fig 10. Ideal octave band division

A comparison between the filtering result of neighboring 2 frequency bands by using DTW-IDTW and that by Kaiser Window Method FIR filter is shown in Fig 11. In this figure, we can see that , when we use the method proposed in this paper, overlap between neighboring bands is very small and original signal is filtered very effectively.



(a) Two adjacent frequency band of filtered pronunciation /aa/ for DTW-IDWT.

b) The output of 65th order kaiser window FIR filter.

Fig 11. Comparison of DWT-IDWT filter and FIR filter

5. Fluid-Cilia Coupling

The stream of lymph happening to bacillar membrane's position change bent cilia

connected

with hair cell. The front phenomenon is modeled in differential form of bailar membrane filter outputs with time as Fig 12.[3][4]

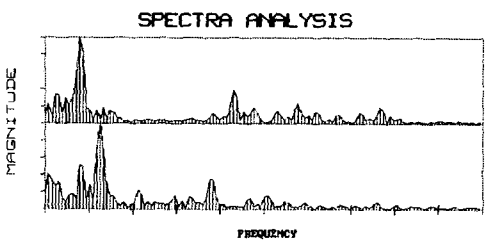
Fig 12. Block diagram of the Fluid-Cilia Coupling

Time-invariant Differentiation emphasis the property of high frequency and differentiation with discrete signal can materialized difference equation like Eq 4.

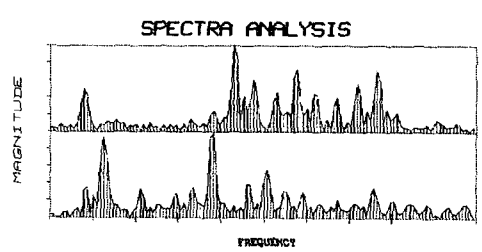
$$\dot{S}_n[n] = S'_n[n] - S'_n[n-1] \qquad \text{(Eq. 4)}$$

With proposed method, if it is applied to difference equation after filter bank output, frequency character of two pronunciation shown in Fig 13(a) clear up like Fig 13(b).

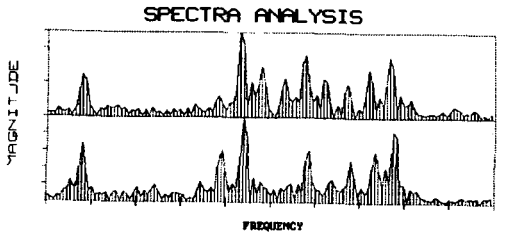
Even if another person is said, characteristics of same pronunciation /ih/ appear similarly with Fig 13(c). Using this method, improvement of speech recognition rate is expected.



(a) The performed output of DTW-IDWT for /ih/ and /eh/



(b) In case of differentiating output of fig (a)



(c) In case of differentiating pronunciation /ih/ for two speaker.

Fig 13. High-frequency emphasis of fluid-cilia coupling

6. Lateral Inhibitory Network(LIN)

In the Centural auditory system, various information, for example pitch, timbre and trait of time & frequency was extracted. The simple functions of LIN which exists in sensory system are used to model auditory system. LIN can guess the trait of structure and function by means of the output of the auditory nerve. These facilities are modeling three stages. First, differentiation in frequency side, which are emulating lateral interaction between LIN neurons. Second, half or full rectifation considering non-linearity of LIN neuron. Third, time-varying integration of outputs. Those procedures are presented in fig 14.

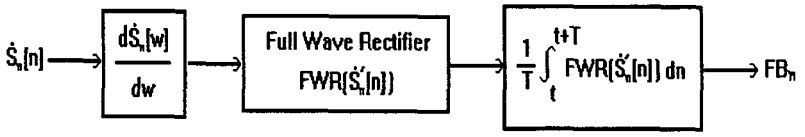


Fig 14. Modelling of the LIN

Time varying differentiation in the same manner, differentiation in frequency side formulated difference equation in expression 5

$$\dot{S}'_n[w] = \dot{S}_n[w] - \dot{S}_n[w-1] \qquad \text{(Eq. 5)}$$

In case of using discrete wavelet trnasformation, differentiation in frequency side

could be accomplished in applying difference equation to results which resoluted bandwidth using DWT. Time varying differentiation and differentiation in frequency side can be out of order by expression 6.

$$\frac{\partial}{\partial w}\left(\frac{\partial S'_n[n;w]}{\partial n}\right) = \frac{\partial}{\partial n}\left(\frac{\partial S'_n[n;w]}{\partial w}\right) \qquad \text{(Eq. 6)}$$

Final outputs in fig 14 are used for feature of speech recognition.

III. Expriments and results

Proposed feature extraction process for speech recognition is presented in fig15.

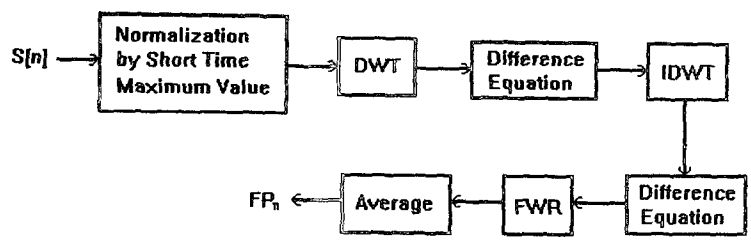


Fig 15. Block diagram of speech feature extraction process

Under the proposed method, the number of speech features are equal to that of practical filter banks and restricted like in expression 3. In this paper, 256 sample per 1 frame are used.

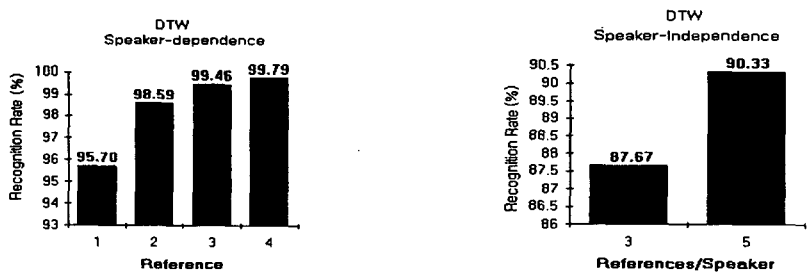
$$\log_2 256 + 1 = 9$$

Therefore, the compositions of filter bank are nine in total. Among the its, considering distortion output of low frequency are excepted. So, Eight outputs are used feature vector of speech signal. In order to test performance of proposed speech feature, DTW and VQ-HMM are used. Experiment conditions were presented in table 1.

Table 1. Conditions for experiments

Recognition Word	Korean isolated digit 0-9
Sampling frequency	16 KHz
Quantization level	16 bits
Speaker	Eight man in 20's
Utterence number per word	10 times
Total data number	800(8*10*10)
Recognizer	DTW, VQ-HMM

In case of DTW, the data used for reference was excepted in recognition test and the number of reference was variable in recognition test. The average of speaker-dependent and speaker-independent recognition results is presented in fig 16. In case of speaker-independent recognition, first and second speaker's data were used for reference. Also, the number of references is different according to speaker and the results are presented in fig 16(b). Similarly, in case of VQ-HMM, the data used for reference was excepted in recognition test. In case of speaker-dependent, experiments have two kinds. First, according to speaker, experiment was accomplished and 10 occurrences of each word by speaker were used. (method 1) It trained 5 occurrences of each word and the rest were used for recognition. Second, all every speaker were used in experiment.(method 2) It trained 5 occurrences of each word by 8 speakers, equally, 400 data and the rest, equally, 400 data were used for recognition. In case of speaker-independent, training set consisted of 400 data by 4 speakers and the rest, equally, 400 data by 4 speakers were used for recognition. Each result is presented in fig 17.



(a) (b)

Fig 16. The results of recognition experiments using DP matching for speaker-dependent and speaker-independent task

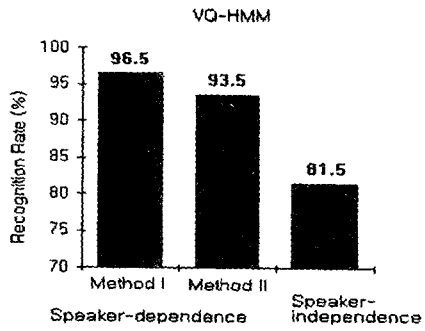


Fig 17. The results of recognition experiments using VQ-HMM for speaker-dependent and speaker-independent task

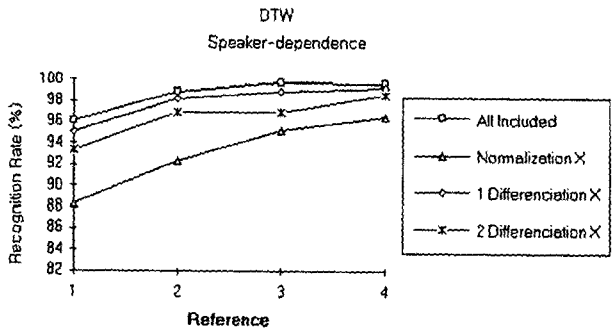


Fig 18. Effects of the hearing models on recognition rate

The effect of the hearing models on recognition rate are presented in fig 18. In case of application of normalization process, recognition rate was 5.48% higher on the average. In case of differentiation in frequency side, it's rate was 0.72% higher and with fluid-cilia coupling, was 2.13% higher. As the number of references decreased, the application of the hearing models was more effective.

The results using VQ-HMM were presented in table 2 and3. VQ level and the number of HMM state are variable in these tables. When VQ level was 32 and the number of HMM state was 6, the recognition rate for Korean digit, from 0 to 9, was highest.

Table 2. Recognition rate for speaker-dependent experiment

HMM state			3	4	5	6
VQ level	32	Method I	95.75 %	96 %	95.5 %	96.5 %
		Mathod II	90.5 %	91 %	92.5 %	93.5 %
	64	Mathod I	96 %	95.25 %	95 %	95.25 %
		Mathod II	92 %	92 %	92 %	92 %

Table 3. Recognition rate for speaker-independent experiment

HMM state		3	4	5	6
VQ level	32	80.75 %	78 %	80.25 %	81.5 %
	64	79 %	79.25 %	78.75 %	78.75 %

IV. Conclusion

In this paper, new algorithm of feature extraction for speech recognition was proposed. It used discrete Wavelet transformation and hearing model. To ascertain whether feature extraction using this algorithm is appropriate for speech recognition, DTW and VQ-HMM were used for recognition experiments. When using DTW, recognition rate was 99.79% at its highest with speaker-dependent and 90.33 at its highest whit speaker-independent. When using VQ-HMM, recognition rate was 96.5% with speaker-dependent and 81.5 with speaker-independent.

Considering the simplicity of realization and the small number of recognition parameters, performance of recognition was effective. Also, when applying each stage in DTW, recognition rate was investigated. So, the effect of the hearing models on recognition rate was checked and as a result of that, application of every stage for hearing model was useful to recognizer.

REFERENCES

[1] Jont B. Allen, "Cochlear Modeling", IEEE ASSP Magazine, January 1985

[2] Stephanie Seneff, *A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing*, Academic Press, 1988

[3] Xiaowei Yang, Kuansan Wang, Shihab A. Shamma, "Auditory Representation of Acoustic Signals", IEEE Trans. Information theory, Vol 38, No. 2, 1992.

[4] Kuansan Wang, Shihab A. Shamma, "Self-Normalization and Noise Robustness in Early Auditory Representations", IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 3, July 1994.

[5] Thomas D. Rossing, *The Science of Sound*, Addison-wesley Publishing Company, 1990.

[6] David Ottoson, *Physiology of the Nervous System*, Macmillan Press, 1983.

[7] Randy K. Young, *Wavelet Theory and Its Application*, Kluwer academic publishers, 1993.

[8] D.E. Newland, *An Introduction to Random Vibration, Spectral & Wvelet Analysis*, Longman Scientific & Technical, 1993.

[9] Olivier Rioul, Martin Vetterli, "Wavelets and Signal Processing", IEEE SP magazine, pp14-38, Oct., 1991.

[10] Mac A. Cody, "The Fast Wavelet Transform", Dr. Dobb's Journal, pp16-24, April, 1992.