Discrete Wavelet Transform을 이용한 음성 추출에 관한 연구

백한욱, 정진현
광운 대학교 제어계측 공학과

A Study Of The Meaningful Speech Sound Block Classification Based On The Discrete Wavelet Transform

Han Wook Baek , Chin Hyun Chung
Dept. of Control & Instrumentation Engineering, Kwangwoon Univ.

Abstract - The meaningful speech sound block classification provides very important information in the speech recognition. The following technique of the classification is based on the DWT (discrete wavelet transform), which will provide a more fast algorithm and a useful, compact solution for the pre-processing of speech recognition. The algorithm is implemented to the unvoiced/voiced classification and the denoising.

1. Introduction

In the meaningful speech sound block classification, the major key point is searching the frequency range that includes the voiced speech sound and the valid unvoiced speech sound. The extraction the frequency range′s data from the original signal by the DWT show the denoising effect and the compression effect of the original signal, which proceeds to the speech recognition.

2. Discrete Wavelet Transform

The wavelet transform has the advantages of the fast computation and its the localization. It extracts the frequency contents of the signal similar to the Fourier transform but it relates the frequency domain with the time domain. This link between the time and the frequency gives this transform a powerful characteristics for the determination of the boundaries of a frequency-band-defined signals such as the voiced and the unvoiced sounds in the speech signal. The voiced sound is a band limits sound and unvoiced sound is spread over all frequencies like the noise.

In general, the wavelet implied in the wavelet transform is a small wave from which many their waves are derived from the signal analyzed by translation and scaling(dilation) of the wavelet wave. The two-dimensional parameterization is achieved from the function called the generating wavelet or mother wavelet, $\psi(t)$ by

$$\psi_{j,k}(t) = 2^{j/2}\varphi(2t-k), \qquad j,k \in Z \tag{1}$$

where Z is the set of all integers and the factor $2^{j/2}$ maintains a constant norm independent of scale j. This parameterization of the time or space location by k and the frequency or scale (actually the logarithm of scale) by j turns out to be extraordinarily effective. This two-variable set of basis function is used in a similar to the short-time Fourier transform, the Gabor transform, or the Wigner distribution for time-frequency analysis. Our goal is to generate a set of expansion function such that any signal is $L^2(R)$ can be represented by the series

$$f(t) = \sum_{j,k} a_{j,k} 2^{j/2}\psi(2^j t-k) \tag{2}$$

or, using (1), as

$$f(t) = \sum_{j,k} a_{j,k}\psi_{j,k}(t) \tag{3}$$

where the two-dimensional set of coefficients $a_{j,k}$ is called the **discrete wavelet transform**(DWT) of $f(t)$. A more specific form indicating how the $a_{j,k}$′s are calculated can be written using inner products as

$$f(t) = \sum_{j,k} \langle \psi_{j,k}(t), f(t) \rangle \psi_{j,k}(t) \tag{4}$$

if the $\psi_{j,k}(t)$ form an orthogonal basis for the space of signals of interest. The inner product is usually defined as

$$\langle x(t), y(t) \rangle = \int x^*(t) y(t) dt \tag{5}$$

If $\varphi_{j,k}(t)$ and $\psi_{j,k}(t)$ are orthonormal or a tight frame, the j level scaling coefficients are found by taking the inner product

$$c_j(k) = \langle f(t), \varphi_{j,k}(t) \rangle = \int f(t) 2^{j/2}\varphi(2^j t-m) dt \tag{6}$$

which, can be written as

$$c_j(k) = \sum_m h(m-2k) \int f(t) 2^{(j+1)/2}\varphi(2^{j+1}t-m) dt \tag{7}$$

but the integral is the inner product with the scaling function at a scale of $j+1$ giving

$$c_j(k) = \sum_m h(m-2k) c_{j+1}(m) \tag{8}$$

The corresponding relationship for the wavelet coefficients is

$$d_j(k)=\sum_m h_1(m-2k)\,c_{j+1}(m) \qquad (9)$$

For any practical signal that is bandlimited, there will be an upper scale $j=J$, above which the wavelet coefficients, $d_j(k)$, are negligibly small. By starting with a high resolution description of a signal in terms of the scaling coefficients c_J, the analysis tree calculates the DWT down to as low a resolution, $j=j_0$, as desired by having $J\Rightarrow j_0$ if $f(t)\in V_j$, it can be expressed as

$$f(t)=\sum_k a_k\varphi(2^jt+k) \qquad (10)$$

Thus, for $f(t)\in V_J$, using (10) we have

$$f(t)=\sum_k c_{J(k)}\varphi_{J,k}(t)$$

$$=\sum_k c_{J-1}(k)\,\varphi_{J-1,k}(t)+\sum_k d_{J-1}\psi_{J-1,k}(t)$$

$$f(t)=\sum_k c_{J-2}(k)\,\varphi_{J-2,k}(t)+\sum_k\sum_{j=J-2}^{J-1} d_{j(k)}\psi_{j,k}(t) \qquad (11)$$

$$f(t)=\sum_k c_{J_0}(k)\,\varphi_{J_0,k}(t)+\sum_k\sum_{j=j_0}^{J-1} d_{j(k)}\psi_{j,k}(t) \qquad (12)$$

The goal of most expansion of a function or signal is to have the coefficients of the expansion $a_{j,k}$ give more useful information about the signal than is directly obvious from the signal itself.

A second goal is to have most of the coefficients be zero or very small. This is what is called a **sparse** representation and is extremely important in application for statistical estimation and detection, data compression, nonlinear noise reduction, and fast algorithm.

Although this expansion is call the discrete wavelet transform, it probably should be called a wavelet series since it is a series expansion which maps a function of a continuous variable into a sequence of coefficients much the same way the Fourier series does.

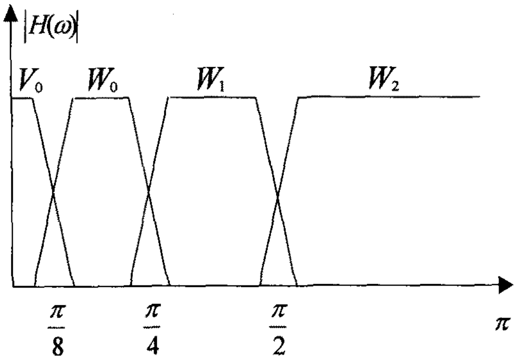

Fig 1. Frequency Bands for he Analysis Tree

3. Implementation Algorithm

A band of which the vowels or voiced sounds are dominant in the speech signal is selected for the analysis. The sampling rate was 11025 samples per second. The statistical results for many vowels of adult males and females indicates that the first formant frequency doesn't below 100hz approximately. But, unfortunately, the unvoiced sound is spread over all frequencies like the noise. Thus, for the searching the valid unvoiced speech sound , a assumption needs. It is that the noise's energy is less than the valid unvoiced sound's energy and the valid unvoiced sound's energy is spread over the band that is less than 3000hz. In general, fortunately, the speech sound obtained by the microphone in PC includes noise less than the valid unvoiced speech sound. With the assumptions and the experimental results, the implementation algorithm is described step by step in each the processing stage. The following table [1] relates the wavelet coefficients to the according frequency band.

Step1.
Discrete Wavelet Transform (Daubechies - 6)

In data extracted in MRA(Multi Resolution Analysis) by using the DWT, the data in the valid frequency range and the data extracted by thresholding to a limited value in data excepted in the valid frequency range will be reconstructed to the original signal. In this step, the meaningful speech data spread over the assumed range will be weighted by the DWT and the property included in the original signal that can classify the meaningful speech sound block will be increased.

Step2.
Filtering

For the extraction of the meaningful speech sound block, a filter is needed, which diminishes the ripples in the signal and contours the signal. This paper suggests the windowing AMDF(Average Magnitude Difference Function) as a filter. The equation implemented in the paper is defined as

$$\gamma(n)=\beta\sum_{m=0}^{k}|x(n+m)-x(n+m+1)| \qquad (13)$$

β is the normalizing number.

Step3.
Thresholding

The result in step 2 will be thresholded by the statistic method.

Step4.
Gathering the meaningful speech sound block

For gathering the meaningful speech sound block, the following rules is needed, which got from the experimental results.

1. A stand-alone block which is less than 300 samples is not meaningful.

2. A block that is between the meaningful blocks and is less than 300 samples can be included in the meaningful blocks.

In general, a man produces speech at an average rate of about 10 phonemes per second [3]. Therefore, for the classification, 1000 samples at least is needed. But, in paper, for the detail classification, the samples range is suggested at 300 ~ 500 samples.

4. Result

The Fig 2. can be obtained by the algorithm. It is good in the assumed pre-condition, but it is sensitive to the high-frequency condition. Therefore, a training algorithm will advance this system in the classification signal. Before the recognition process, this algorithm compresses the original speech into half of its index, and the compressed data includes the rich component of the original. The algorithm includes some denoising effects, which eliminates the high-frequency omitted in the assumed condition.

Frequency Range In Hz	Number of Coefficients
2756 - 5512	512
1378 - 2756	256
689 - 1378	128
344 - 689	64
172 - 344	32
86 - 172	16
43 - 86	8
21 - 43	4
10 - 21	2
0 - 10	1

Table 1. Frequency Range via Coefficients Number

References

[1] Strang/Nguyen, "Wavelets and Filter Banks", Wellesley-Cambridge Press

[2] C.Sidney Burrus, Ramesh A. Gopinath, and Haitao Guo, "Introduction to Wavelets and Wavelet Transforms", Prentice-Hall

[3] L.R. Rabiner / R.W. Schafer "Digital Processing of Speech Signals", Prentice-Hall

[4] Lawrence Rabiner/Biing-Hwang Juang, "Fundamentals of Speech Recognition", Pretice-Hall

[5] Thomas W.Parsons, "Voice and Speech Processing", McGraw-Hill

[6] John G. Proakis, "Digital Signal Processing", Prentice-Hall

[7] Kung, "VLSI Array Processors", Prentice-Hall

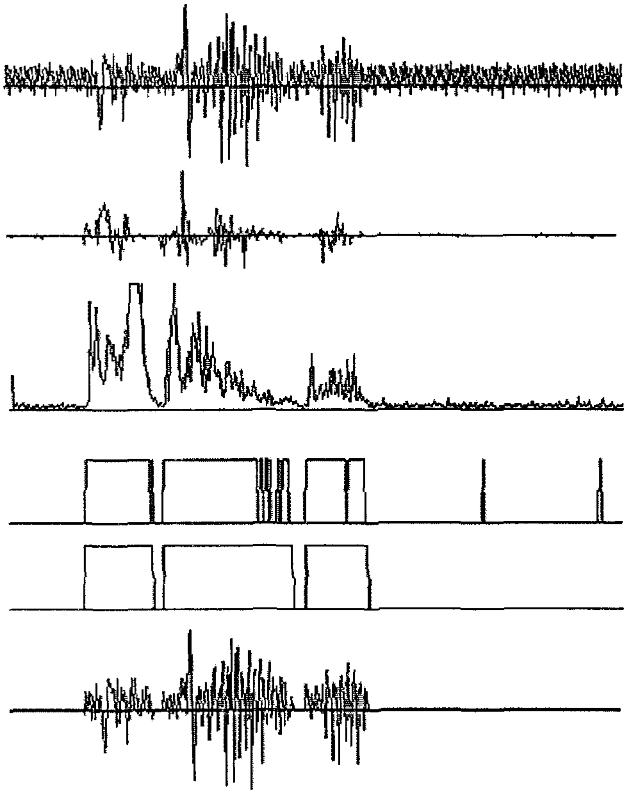

Fig 2 . Step Process Results