이산 웨이브렛 변환을 이용한 유효 음성 추출에 관한 연구

김 진 옥[†] · 황 대 준[††] · 백 한 욱[†††] · 정 진 현[††††]

요 약

유효한 무성음이 시스템 노이즈와 합성됐을 경우 유효한 무성음 추출에 많은 어려움이 있으나 본 논문에서는 유효한 무성음 추출에 있어 이산 웨이브렛 변환을 이용한 신호 해석 내용을 기반으로 주파수와 그 위치를 블록별로 머징 규칙으로 유효 여부를 결정하기 때문에 노이즈가 많은 환경에서도 유효한 무성음 추출이 가능하다. 머징 알고리즘은 음성만으로도 처리 매개변수를 결정할 수 있고 시스템 잡음에 대하여서도 독립적이기 때문에 유효한 음성을 추출하는데 매우 효과적이다. 실험 결과를 통하여 유효한 음성 추출 처리 과정에서 보다 향상된 결과를 보이고 있으며 특히 고주파 노이즈에 대한 강한 적응력을 제시하고 시스템 구현에도 용이한 시스템 튜닝을 가능케 한다.

A Study on Extracting Valid Speech Sounds by the Discrete Wavelet Transform

Jin Ok Kim[†] · Dae Joon Hwang[††] · Han Wook Baek[†††] · Chin Hyun Chung[††††]

ABSTRACT

The classification of the speech-sound block comes from the multi-resolution analysis property of the discrete wavelet transform, which is used to reduce the computational time for the pre-processing of speech recognition. The merging algorithm is proposed to extract valid speech-sounds in terms of position and frequency range. It performs unvoiced/voiced classification and denoising. Since the merging algorithm can decide the processing parameters relating to voices only and is independent of system noises, it is useful for extracting valid speech-sounds. The merging algorithm has an adaptive feature for arbitrary system noises and an excellent denoising signal-to-noise ratio and a useful system tuning for the system implementation.

키워드 : 음성추출, 복수 해상도, 이산 웨이브렛 변환, 머징 알고리즘, 노이즈 분리 특성

1. Introduction

The merging algorithm[1] extracts valid speech data, specially focused on the extraction of unvoiced speech-sound blocks with the consideration of its position and frequency range. This information is supplied by the multi-resolution analysis[2, 3]. Since the position of unvoiced phonemes[4] found in speech can be used to reconstruct it, the discrete wavelet transform plays an important role. In the extraction of a valid speech-sound block, a lot of works are devoted to search the frequency range included in the voiced/unvoiced speech and each of its positions. But the simultaneous analysis of the frequency and time can hardly be obtained with the Fourier transform. Thus, the discrete wavelet transform is used for its simultaneous analysis and for a decrease in its computational amount [5-7]. To extract data on the desired frequency range of the original signal by the discrete wavelet transform is to take the denoising effect and the compressional effect on the speech signal resource. The merging algorithm is proposed to discriminate between valid phonemes and silence ranges[8, 9].

† 정 회 원 : 성균관대학교 대학원 전기전자및컴퓨터공학부
†† 정 회 원 : 성균관대학교 전기전자및컴퓨터공학부 교수
††† 준 회 원 : American-Panel Corporation in USA 연구원
†††† 정 회 원 : 광운대학교 정보제어공학과 교수
논문접수 : 2001년 6월 9일, 심사완료 : 2002년 1월 31일

2. Discrete Wavelet Transform

The wavelet transform has the advantages of a fast

computation and its localization. It extracts the frequency contents of the signal similar to the Fourier transform but it relates the frequency domain with the time domain. The two-dimensional parameters are achieved from a function called the generating wavelet or mother wavelet[10], $\psi(t)$, by

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k),$$
$$\varphi_{j,k}(t) = 2^{j/2}\varphi(2^j t - k),$$

where □($j, k \in$□) is the set of all integers and the factor $2^{j/2}$ maintains a constant norm. The parameters of the time or space location by k and the frequency or scale (actually the logarithm of scale) by j turn out to be extraordinarily effective[11]. The goal is to generate a set of expansion functions so that any signal $L^2(R)$ can be represented by the series

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

where the two-dimensional set of coefficients $a_{j,k}$ is called the discrete wavelet transform of $f(t)$. If $\varphi_{j,k}(t)$ and $\psi_{j,k}(t)$ are orthogonal, the j level scaling coefficients are found by taking the inner product

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

The analysis tree, in terms of the scaling coefficients, calculates the discrete wavelet transform down to a lower resolution. If $f(t) = V_j$, it can be expressed as

$$f(t) = \sum_k c_{J_0}(k)\varphi_{J_0,k}(t) + \sum_k \sum_{j=J_0}^{J-1} d_j(k)\psi_{j,k}(t).$$

The coefficients of the expansion functions give us more useful information about the signal by the expansion[12]. Since the most of the coefficients are zero or very small, the sparse representation is very useful in applications for statistical estimation and detection, data compression, nonlinear noise reduction, and fast algorithm.

3. Extraction of Valid Speech-Sounds

A band in which the vowels or voiced sounds are dominant in the speech signal is selected for analysis. The statistical results of many vowels of adult males and females indicate that the first formant frequency does not exist below approximately 100 Hz[13, 14]. However, the unvoiced sound is spread over all frequencies as noise. Thus, when searching for valid unvoiced speech sounds, one can make the following assumptions[15-17] ;

- The energy of noise is less than the valid voiced sound.
- The valid unvoiced sound wraps the voiced sound.
- The valid unvoiced sound spreads over the band that is less than about 3 kHz.

In general, a speech sound obtained by a microphone includes less noise than a valid unvoiced speech sound. The merging algorithm is proposed to focus on denoising and the extraction of the unvoiced speech-sound block. <Table 1> relates the wavelet coefficients with the according frequency band. In order to extract speech coefficients in band-limited frequency, the input speech data is analyzed by the discrete wavelet transform.

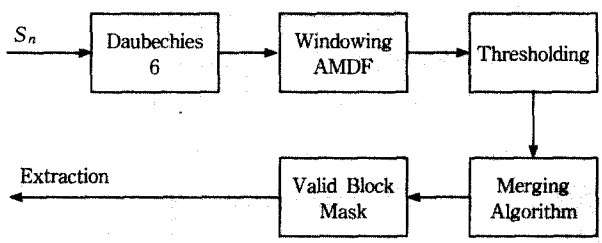
〈Table 1〉 Number of coefficients for each resolution

Frequency range in Hz	Number of coefficients
2,756~5,512	512
1,378~2,756	256
689~1,378	128
344~689	64
172~344	32
86~172	16
43~86	8
21~43	4
10~21	2
0~10	1

The silence-discrimination method that uses energy and zero-crossing is useful in the case of an extremely high signal-to-noise ratio. However, such ideal conditions are not practical for most application environments in which a neighboring device occasionally generates high frequency noises. The merging algorithm is used in the extraction process with a position array of each phoneme and a higher frequency of unvoiced speech than of voiced speech. Several processes are dedicated to the extraction of the valid block in order to merge each phoneme block. These are processed in terms of the phoneme-block to increase the discrimination property.

(Figure 1) shows a blockdiagram of the merging al-

gorithm ; Our interests are concentrated in coefficient's data spread over in a multi-resolution analysis domain with the discrete wavelet transform, the data in the assumed frequency range and the data extracted by thresholding a limited value. The valid speech data spread over the assumed frequency range is weighted by the discrete wavelet transform to classify the valid speech-sound block. While Daubechies-4 wavelets is used widely, Daubechies-6 wavelet is applied because of the pre-implemental results. It is a bi-orthogonal wavelet basis function, which can avoid interferences of the neighbor-band wavelet packets in reconstruction.



(Figure 1) Processing diagram

For the extraction of the valid speech-sound block, the windowing AMDF (average magnitude difference function) is used as a filter to diminish the ripples and contours in the signals. To implement the filter, its equation is defined as

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

where β is a normalizing coefficient and p is the block size. The windowing AMDF is applied to generate the basic resources of the merging process. It can filter the transformed data when considering a valid speech-sound block and preparing the thresholding process.

The input speech data is purified for the discrimination of valid and unvalid speech-sound blocks with the multi-resolution analysis and several processing facts are merged to extract the valid speech-sound block by the rules proposed in this paper. To merge the valid phoneme block, the following rules are needed because of the experimental results : A stand-alone block which consists of less than 300 samples is not valid and a block that is between the valid blocks and consists of less than 300 samples can be included in the valid blocks.

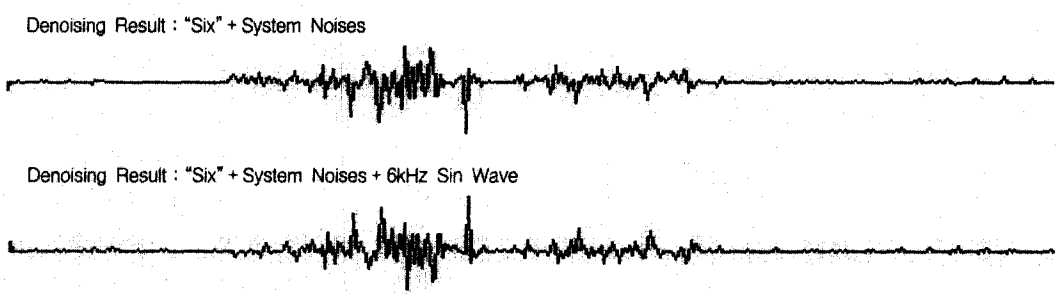
The signal is merged with the information and the predefined rules. In general, a person produces speech at an average rate of about 10 phonemes per second. Therefore, for classification, at least 1000 samples are needed in a sampling frequency of 11,025 Hz. But for the detail classification, the minimum size of a valid block frame is suggested at 300~500 samples. To determine the valid block, we must consider simultaneously the energy and position of each frame. Since the valid block is extracted by the merging rules described, its valid speech-sound block can be classified.

4. Experiment and Result

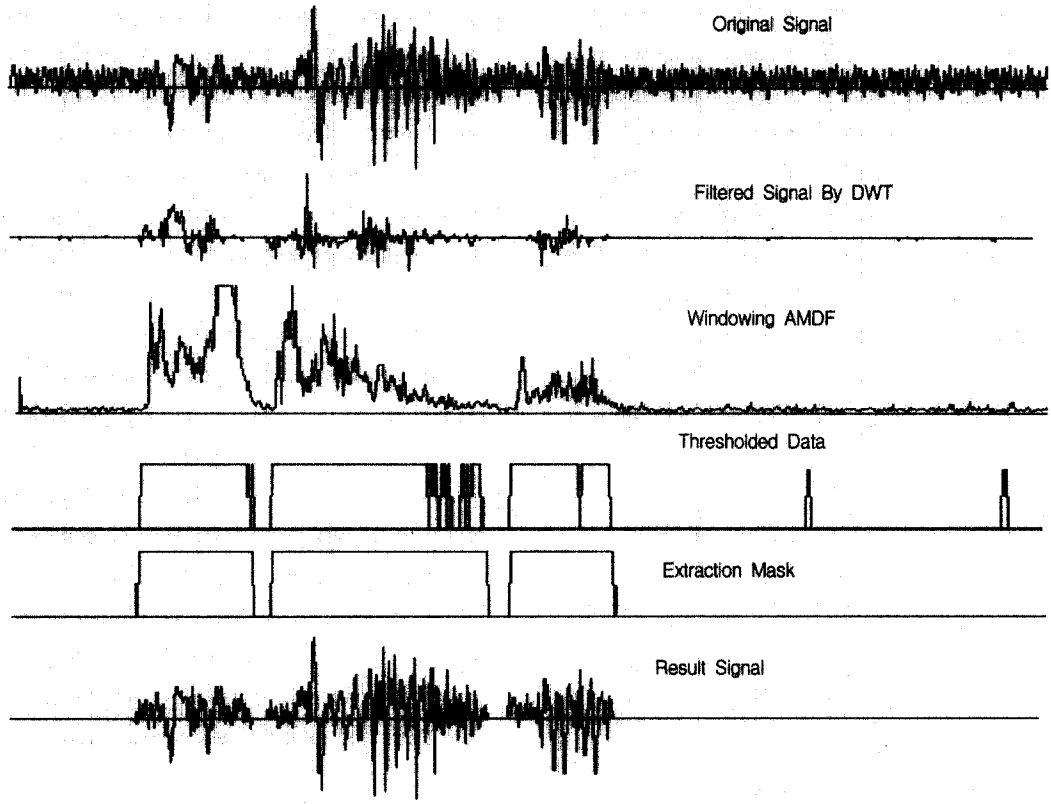
The merging algorithm is implemented to get the sample data through a microphone within the sampling rate of 11,025 Hz.

Describing the discrete wavelet transform's extraction performance of the desired frequency range, (Figure 2) shows the denoising effect of the discrete wavelet transform.

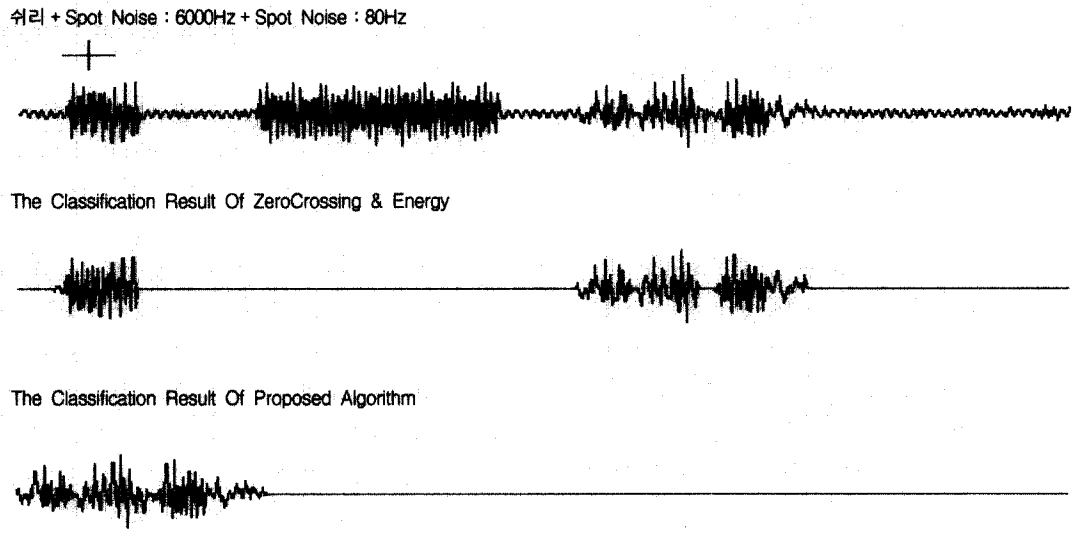
(Figure 3) shows that the merging algorithm has an adaptive feature for arbitrary system noises. The original signal includes the high frequency system noises, whereas in the result signal, the merging algorithm shows an improved extraction performance, especially when denoising the higher frequency noise.



(Figure 2) Denoising



(Figure 3) Extraction processes



(Figure 4) Classification signal with spot-noise.

〈Table 2〉 Property comparison

Specifications	Merging algorithm	Zero-crossing & energy consideration
Speech sound starting with a voiced	91%	91%
Speech sound starting with an unvoiced	89%	87%
Speech sound with spot-noise	90%	30%

(Figure 4) shows the classified and compressed signals in speech signals with spot-noises.

<Table 2> shows a comparison of the merging algorithm with the "Zero-Crossing & Energy Consideration." The merging algorithm is independent of system noises and it has an adaptive feature for spot noises.

5. Conclusion

Since the merging algorithm is based on the multi-resolution analysis with the discrete wavelet transform, its computation seems to be more complex, but, with the fact that the basic computation of the discrete wavelet transform is processed by convolution, it is processed more quickly by the pipeline processing of convolution. Since the other methods must decide the processing parameters with view-points of the consideration of system noises and voices, they hardly tune themselves.

However, the merging algorithm is more useful to extract valid speech-sound since it can decide the processing parameters only at the standpoint of voices and is independent of system noises. The merging algorithm has the adaptive feature for arbitrary system noises and the excellent de-noising signal-to-noise ratio.

References

[1] Jin Ok Kim, Dae Joon Hwang, Han Wook Baek, and Chin Hyun Chung, "*An application of the merging algorithm with the discrete transform to extract valid speech-sound,*" in IEEE VIMS 2001, Budapest, Hungary, pp.67-70, IEEE, May, 2001.

[2] Raghuveer M. Rao and Ajit S. Bopardikar, *Wavelet Transforms : Introduction to Theory and Applications*, Addision Wesley, Readming, MA., 1998.

[3] James S. Walker, *A primer on Wavelets for Their Scientific Applications*, CRC Press, Boca Ration, FL., 1999.

[4] Randy Goldberg and Lance Riek, *A Practical Handbook of Speech Coders*, CRC Press, Boca Ratin, FL., 2000.

[5] Jaideva C. Goswami and Andrew K. Chan, *Fundamentals of Wavelets : Theory, Algorithms and Applications*, John Wiley & Sons, New York, 1999.

[6] Anthony Teolis, *Computational Signal Processing with Wavelets*, Springer Verlag, New York, 1998.

[7] C. Sidney Burrus, Ramesh A. Gopinath, and Hitao Guo, *Introduction to Wavelets and Wavelet Transforms : A primer*, Prentice Hall, New Jersey, 1997.

[8] T. L. Marzetta, "*A new interpretation for capon's maximum likelihood method of frequency wavenumber spectral estimation,*" IEEE Trans. Acoustics, Speech and Signal Processing, Vol.31, 1983.

[9] John R. Deller, John H. L. Hansen, and John G. Proakis, *Discrete-Time Processing of Speech Signals (IEEE press Classic Reissue)*, IEEE Press, New York, 2000.

[10] D. L. Donoho, "*Denosing by soft-thresholding,*" IEEE Trans. Information Theory, Vol.41, 1995.

[11] Agostino Abbate, Casimer M. Decusatis, and Pankaj K. Das, *Wavelets and Subband : Fundamentals and Applications*, Birkhauser, Stuttgart, Germany, 2001.

[12] R. Todd Ogden, *Essential Wavelets for Statistical Applications and Data Analysis*, Springer Verlag, New York, 1996.

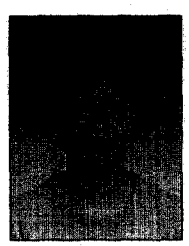
[13] Thomas W. Parsons, *Voice and Speech Processing*, McGraw-Hill, New York, 1986.

[14] Sadaoki Furui, *Digital Speech Processing, Synthesis and Recognition*, Marcel Dekker, New York, 2nd edition, 2001.

[15] Dan Jurasfky, James H. Martin, Keith Vander Linden, and Daniel Jurafsky, *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice Hall, New Jersey, 2000.

[16] Nelson Morgan and Ben Gold, *Speech and Audio Signal Processing : Processing and Perception of Speech and Music*, John Wiley & Sons, New York, 1999.

[17] Lawrence R. Rabiner and Ronald W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, New Jersey, 1978.



김 진 옥

e-mail : jinny@ece.skku.ac.kr

1985년 성균관대학교(문학사)

1998년 성균관대학교 대학원 정보통신공학과(공학석사)

1998년~현재 성균관대학교 전기전자및컴퓨터공학부(박사과정)

1992년~1994년 ㈜현대전자산업 정보통신본부

1994년~1999년 ㈜현대정보기술 인터넷사업본부 과장

1999년~2000년 ㈜온세통신 온라인사업 팀장

2000년~2001년 ㈜유로코넷 기술담당 이사

관심분야 : Multimedia, Image Processing, Biometrics, Data Mining, Recognition

황 대 준

e-mail : djhwang@skku.ac.kr

1978년 경북대학교 컴퓨터공학과(공학사)

1981년 서울대학교 컴퓨터과학과(이학석사)

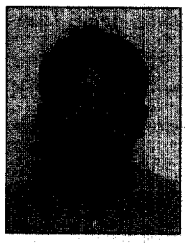
1986년 서울대학교 컴퓨터과학과(이학박사)

1981년~1987년 한남대학교 전자계산학과 교수

1990년~1991년 미국 MIT 컴퓨터과학연구소 연구교수

1987년~현재 성균관대학교 전기전자 및 컴퓨터공학부 교수

관심분야 : 멀티미디어, 원격교육, 병렬처리, 가상교육, 지적재산권 보호 시스템



백 한 욱

e-mail : jaco811@hotmail.com

1994년 광운대학교 제어계측공학과(공학사)

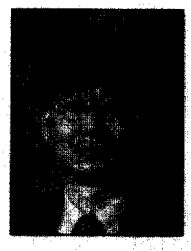
2000년 광운대학교 제어계측공학과 (공학석사)

1994년~1996년 한국전자 (KEC) 전자기기 사업부 연구원

1996년~1998년 LG전자 생산기술센터 연구원

2001년~현재 American-Panel Corporation in USA 연구원

관심분야 : VHDL, Recognition, Biometrics, Embedded System



정 진 현

e-mail : chung@daisy.kwangwoon.ac.kr

1981년 연세대학교 전기공학과(공학사)

1983년 연세대학교 대학원 전기공학과 (공학석사)

1990년 Rensselaer Polytechnic Institute (Ph.D)

1991년~현재 광운대학교 정보제어공학과 교수

관심분야 : DSP, VHDL, CIM, Network, Intelligent Control, Recognition, Biometrics,Embedded System