이산 웨이브렛 변환을 이용한 유효 음성 추출을 위한 머징 알고리즘

(A Merging Algorithm with the Discrete Wavelet Transform to Extract Valid Speech-Sounds)

김진옥[†] 황대준[††] 백한욱[†††] 정진현[††††]

(Jin Ok Kim) (Dae Jun Hwang) (Han Wook Paek) (Chin Hyun Chung)

요 약 데이타로부터 유효한 음성 데이타를 추출하는 것은 음성 인식분야에서 중요하다. 본 논문의 음성 추출 기술은 빠른 연산이 가능하며 음성의 전처리 과정에 적합한 이산 웨이브렛 변환을 사용하고 있으며, 이산 웨이브렛 변환의 복수 해상도 해석 특징을 이용한 머징 알고리즘으로 유효한 음성을 추출하고 노이즈 제거를 동시에 구현한다. 머징 알고리즘은 음성만으로도 처리 매개변수를 결정할 수 있고 또한 시스템 잡음에 대하여서도 독립적이기 때문에, 유효 음성을 추출하는데 매우 효과적이다. 그리고 머징 알고리즘은 시스템 잡음에 대한 적응 특성을 갖고 탁월한 노이즈 분리 특성을 갖는다.

키워드 : 머징 알고리즘, 복수 해상도 해석, 이산 웨이브렛 변환, 신호대 잡음비, 유효 음성, 퓨리어 변환

Abstract A valid speech-sound block can be classified to provide important information for speech recognition. The classification of the speech-sound block comes from the MRA(multi-resolution analysis) property of the DWT(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 needs some numerical methods for an adaptive DWT implementation and 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 SNR(signal-to-noise ratio).

Key words : Merging algorithm, MRA(multi-resolution analysis), DWT(discrete wavelet transform), SNR(signal-to-noise ratio), valid speech-sound, Fourier transform

1. Introduction

In the case of building a speech recognition system, we spend a lot of time in tuning up the pre- or post-process of an imported speech because the quality of the pre- or post-process is sometimes dependent on the range of the speech that is coming in or out. Even a valid speech may overlap with some high frequency noises. The Fourier transform method to extract valid speech takes a lot of processing time and is limited in the frequency analysis. Since a zero-crossing method is very sensitive to external noises, it could be ineffective [1] [2] [3].

In order to obtain reliable and valid speech, the MRA property is proposed because it can track and reconstruct the unvoiced phonemes in speech. The merging algorithm [4] extracts valid speech data, especially the unvoiced speech-sound blocks that consider position and frequency range. When

[†] 학생회원 : 성균관대학교 전기전자및컴퓨터공학부
jinny@ece.skku.ac.kr

[††] 종신회원 : 성균관대학교 전기전자및컴퓨터공학부 교수
djhwang@skku.ac.kr

[†††] 비 회 원 : American-Panel Corporation 연구원
jaco811@hotmail.com

[††††] 정 회 원 : 광운대학교 정보제어공학과 교수
chung@daisy.kwangwoon.ac.kr

논문접수 : 2001년 6월 19일
심사완료 : 2002년 2월 26일

extracting a valid speech-sound block, much work has to be devoted to the search of the frequency range included in the voiced/unvoiced speech and in each of its positions. However, the simultaneous analysis of the frequency and time(position) can hardly be obtained by the Fourier transform [5] [6]. Extracting data from the desired frequency range of the original signal by the DWT involves considering the denoising effect and the compression effect on the speech signal [7] [8] [9] [10]. Thus, the DWT is used for simultaneous analysis and for a decrease in its computational amount.

The merging algorithm is therefore proposed to discriminate between valid phonemes and silence.

2. Discrete Wavelet Transform

In general, a wavelet is a small wave which has its energy concentrated in time. It can be used to give a tool for the analysis of transient, nonstationary, or time-varying phenomena. A wavelet still maintains an oscillating wavelike characteristic but also has the ability to allow simultaneous time and frequency analysis with a flexible mathematical foundation [7] [11]. The two-dimensional parameters are achieved from a function called "the generating wavelet" or "mother wavelet" [12], Ψ(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. This parameter 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 [13] [14]. 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 $f(t)$. The MRA property of the DWT is well defined in the implementation [15], which is formulated by requiring a nesting of the spanned spaces as

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots \subset L^2$$

or

$$V_j \subset V_{j+1}$$

for all $j \in *$ with

$$V_{-\infty} = \{0\}, \quad V_\infty = L^2$$

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

3. Merging Algorithm

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 [14] [16] [17]. 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;

- 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. However, if a system is constructed in real fields, denoising high frequency noises is included. The merging algorithm is proposed to focus on denoising and the extraction of the unvoiced speech-sound block.

The silence-discrimination method that uses energy and zero-crossing is useful in the case of an extremely high SNR(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. The detailed algorithm is described below:

Step 1: Discrete Wavelet Transform(Daubechies-6)

Our interests are concentrated on the coefficients spread in the MRA domain processed by the DWT. The data in the assumed frequency range(100 ~ 3000 Hz) will be extracted by the thresholding process with a proper value. In step 1, the valid speech data spread over that range are classified by the DWT and weighted with a proper value at each frequency band. The Daubechies-6 wavelet is applied to avoid interferences from the neighbor-band wavelet packets in the reconstruction.

Step 2: Filtering

For the extraction of the valid speech-sound block, the windowing AMDF(average magnitude difference function) [18] [19] is used as a filter to diminish the ripples and contours in the signals. The equation used to implement the filter 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.

Step 3: Thresholding

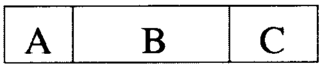
The result obtained in step 2 is thresholded to the adjustable value which is made from many trials. Needless to say, it is processed from a valid speech-sound block.

Step 4: Merging the valid speech-sound block

The input, speech data, is purified for the discrimination of valid and unvalid speech-sound blocks with the MRA. Several processing facts are merged to extract the valid speech-sound block according to the rules proposed in this paper. To merge the valid phoneme block, the following rules are necessary because of the experimental results:

- A stand-alone block which consists of less than 300 samples is not valid.
- A block that is between the valid blocks and consists of less than 300 samples can be included in the valid blocks.

Figure 1 shows an array of phonemes.



A : Unvoiced Speech
B : Voiced Speech
C : Silence Range

Figure 1 The position relation of each phoneme

The signal is merged with the information and the pre-defined rules. In general, a person produces speech at an average rate of about 10 phonemes per second. Therefore, for classification purposes, at least 1000 samples are needed in a sampling frequency of 11025 Hz. Except for the detail classification, the minimum size of a valid block frame is suggested at 300 ~ 500 samples. To determine the valid block, one should consider the energy and position of each frame simultaneously. Since the valid block is extracted by the merging rules described, its valid speech-sound block can be classified.

4. Experiment Results

The merging algorithm is implemented to get the sample data through a microphone within the sampling rate of 11025 Hz. To describe the DWT's extraction performance of the desired frequency range, Fig. 2 shows the denoising(filtering a band range's data) effect of the DWT.

$$\mathrm{SNR} = \frac{E[x^2(n)]}{E[e^2(n)]} = \frac{\sum_n x^2(n)}{\sum_n e^2(n)}$$

Table 1 Denoising SNR

Original Signal	0.962391
Denoised Signal + 6 kHz Sin	254.214

Table 1 shows that the noises have no interference following the extraction of the desired band, if the desired frequency range is added.

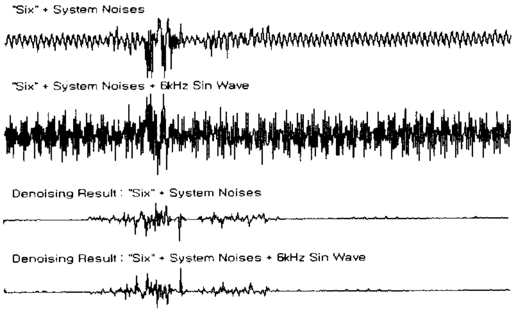


Figure 2 Denoising results

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.

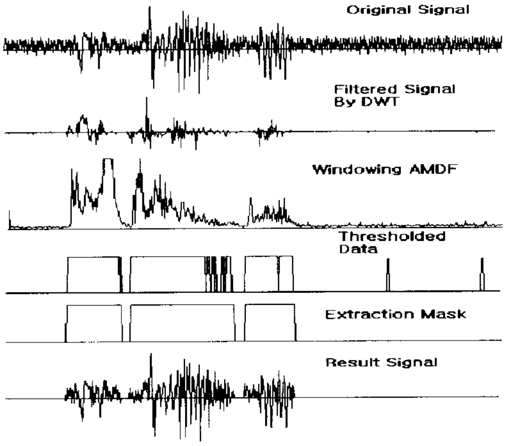


Figure 3 Each of the extraction processes

Figures 4 and 5 show that the original signals compounded with 80 Hz and 6000 Hz sin wave are processed by the merging algorithm. They also show that the merging algorithm is not disturbed by an unexpected system interference.

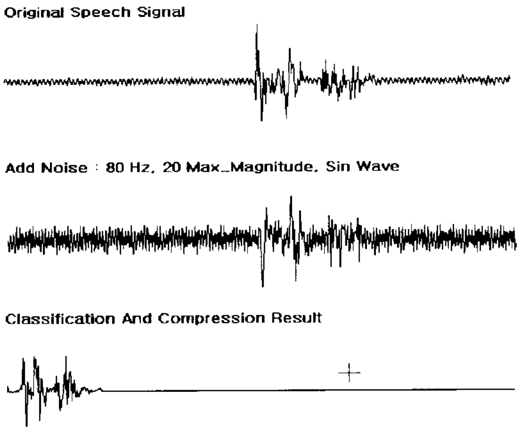


Figure 4 The result from the signal added noise: 80 Hz sin wave.

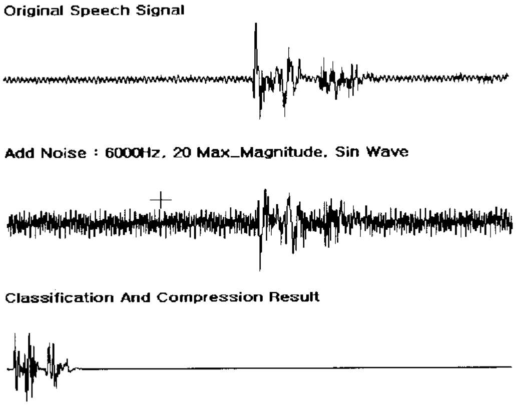


Figure 5 The result from the signal added noise: 6000 Hz sin wave.

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.

Table 2 Property comparison

Items	Zero-crossing & energy consideration	Merging algorithm
System dependency	Higher	Lower
Spot-noise	Non-adaptive	Adaptive
Signal analysis	None	DWT's MRA
Unvoiced phoneme	Dependent on system	Dependent on the frequency range

5. Conclusion

In general, high frequency noises included in a normal speech stream are difficult to remove from the speech stream. Because an unvoiced phoneme seems like a high frequency noise, it may be removed during denoising. A low frequency noise(hum noise), on the other hand, may come from a circuitry imbalance, a wrongly designed ground point in PCB, or imbalance among the parts mounted on a board. This experiment results show that the merging algorithm is very robust against external effects. Since the merging algorithm proposed in this paper is based on the MRA with the DWT, its computation seems complicated. However, because the basic computation of the DWT is processed by convolution, it can be done more quickly by the pipeline processing of convolution. Since the other methods must decide the processing parameters of system noises and voices, they can hardly tune themselves. The merging algorithm is useful for extracting valid speech-sounds since it can decide the processing parameters relating to voices only and is independent of system noises. The merging algorithm has an adaptive feature for arbitrary system noises and an excellent denoising SNR.

References

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

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

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

[4] Jin Ok Kim, Dae Jun Hwang, Han Wook Paek, 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, May 2001, pp. 67-70, IEEE.

[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] S. K. Mitra, *Digital Signal Processing: A Computer-Based Approach*, McGraw-Hill, New York, 2nd edition, 2000.

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

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

[16] 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.

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

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

[19] Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, 1993.

김진옥

1989년 성균관대학교(문학사). 1998년 성균관대학교 대학원 정보통신공학과(공학석사). 1998년 ~ 현재 성균관대학교 전기전자 및 컴퓨터공학부(박사과정). 1992년 ~ 1994년 ㈜현대전자산업 정보통신본부. 1994년 ~ 1999년 ㈜현대정보기술 인터넷사업본부 과장. 1999년 ~ 2000년 ㈜온세통신 온라인사업 팀장. 2000년 ~ 2001년 ㈜유로코넷 기술담당 이사. 관심분야는 Multimedia, Image Processing, Biometrics, Data Mining, Recognition

황대준

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. 관심분야는 멀티미디어, 원격교육, 병렬처리, 가상교육, 지적재산권 보호 시스템

백한욱

1994년 광운대학교 제어계측공학과(공학사). 2000년 광운대학교 제어계측공학과(공학석사). 1994년 ~ 1996년 한국전자(KEC) 전자기기사업부 연구원. 1996년 ~ 1998년 LG전자 생산기술센터 연구원. 2001년 ~ 현재 American-Panel Corporation in USA 연구원. 관심분야는 VHDL, Recognition, Biometrics, Embedded System

정진현

1981년 연세대학교 전기공학과(공학사). 1983년 연세대학교 대학원 전기공학과(공학석사). 1990년 Rensselaer Polytechnic Institute(Ph.D). 1991년 ~ 현재 광운대학교 정보제어공학과 교수. 관심분야는 DSP, VHDL, CIM, Network, Intelligent Control, Recognition, Biometrics, Embedded System