On the Extraction of the Valid Speech-Sound by the Merging Algorithm with the Discrete Wavelet Transform

Jin Ok Kim[1], Han Wook Paek[2], Chin Hyun Chung[2], Jun Hwang[3], and Woongjae Lee[3]

[1] School of Information and Communication Engineering, Sungkyunkwan University, 300, Chunchun-dong, Jangan-gu, Suwon, Kyunggi-do, 440-746, KOREA
jinny@ece.skku.ac.kr

[2] Department of Information and Control Engineering, Kwangwoon University, 447-1, Wolgye-dong, Nowon-gu, Seoul, 139-701, KOREA
chung@kw.ac.kr

[3] Division of Information and Communication Engineering, Seoul Women's University, 126, Kongnung2-dong, Nowon-gu, Seoul, 139-774, KOREA
wjlee@swu.ac.kr

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).

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].

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 is proposed to extract valid speech data, especially the unvoiced speech-sound blocks that consider position and frequency range.

P.M.A. Sloot et al. (Eds.): ICCS 2003, LNCS 2657, pp. 314–322, 2003.

When 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 [2] [3]. 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 [4] [5] [6] [7]. 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 [4] [8]. The two-dimensional parameters are achieved from a function called "the generating wavelet" or "mother wavelet" [9], $\psi(t)$ by

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

$$\varphi_{j,k}(t) = 2^{j/2}\varphi(2^j t - k), \quad j,k \in \mathbf{Z} \tag{2}$$

where $\mathbf{Z}$ 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 [10]. The goal is to generate a set of expansion functions so that any signal $L^2(R)$ can be represented by the series

$$f(t) = \sum_{j,k} a_{j,k} 2^{j/2}\psi(2^j t - k) \tag{3}$$

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 [11], which is formulated by requiring a nesting of the spanned spaces as

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots \subset L^2 \tag{4}$$

or

$$V_j \subset V_{j+1} \quad \text{for all} \quad j \in \mathbf{Z} \tag{5}$$

with

$$V_{-\infty} = \{0\}, \quad V_\infty = L^2 \tag{6}$$

$$f(t) = \sum_k c_{J_0}(k)\varphi_{J_0,k}(t) + \sum_k \sum_{j=J_0}^{J-1} d_j(k)\psi_{j,k}(t) \tag{7}$$

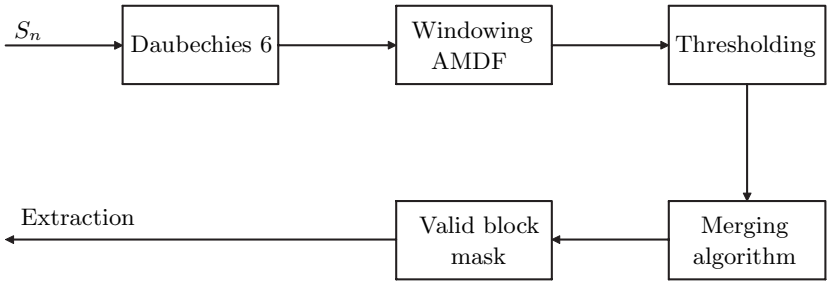


Fig. 1. Processing diagram

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 [10] [12] [13]. 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 [14] [15]. However, if a system is constructed in real fields, denoising high frequency noises is included [16]. 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.

Figure 1 shows a block diagram of the merging algorithm and its detailed algorithm is described in Algorithm 1.

Figure 2 shows an array of phonemes. 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

Algorithm 1: The Merging Algorithm

Data : speech signal with high frequency noises and unvoiced speech-sound block

Result : denoising and extraction of unvoiced speech-sound block

begin

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 $\sim$ 3,000 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.

2 **Filtering:** 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. The equation used to implement the filter is defined as

$$\gamma(n) = \beta \sum_{m=0}^{p} |x(n+m) - x(n+m+1)| \tag{8}$$

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.

3 **Thresholding:** The result obtained in step 2 is thresholded to the adjustable value which is made from many trials. It is processed from a valid speech-sound block.

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 consists of less than 300 samples can be included in the valid blocks.

end

least 1,000 samples are needed in a sampling frequency of 11,025 Hz. Except for the detail classification, the minimum size of a valid block frame is suggested at 300 $\sim$ 500 samples. To determine the valid block, we 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.

Table 1. Denoising SNR.

Original Signal	0.962391
Denoised Signal + 6 kHz Sin	254.214

4 Experiment Results

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

To describe the DWT's extraction performance of the desired frequency range, Fig. 3 shows the denoising(filtering a band range's data) effect of the DWT.

$$\mathbf{SNR} = \frac{E[x^2(n)]}{E[e^2(n)]} = \frac{\sum_n x^2(n)}{\sum_n e^2(n)} \tag{9}$$

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

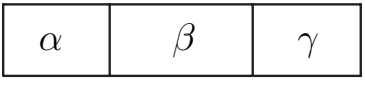
Figure 4 shows that the merging algorithm has an adaptive feature for arbitrary system noises. The original signal includes the high frequency system noises, whereas in the result signal, the merging algorithm shows an improved extraction performance, especially when denoising the higher frequency noise.

Figures 5 and 6 show that the original signals compounded with 80 Hz and 6,000 Hz sin wave are processed by the merging algorithm. They also show that the merging algorithm is not disturbed by an unexpected system interference.

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

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



α: unvoiced speech

β: voiced speech

γ: silence range

Fig. 2. The position array of each phoneme

(a) a speech signal of “six” added with system noises

(b) a speech signal of “six” added with system noises and 6kHz sine wave

(c) denoising the speech signal of “six” added with system noises

(d) denoising the speech signal of “six” added with system noises and 6kHz sine wave

Fig. 3. Denoising.

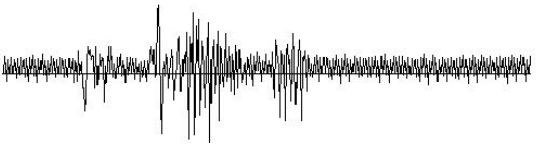
Table 2. Property comparison.

Item	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

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.

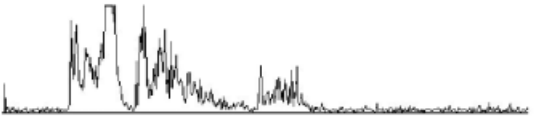
References

1. Goldberg, R., Riek, L.: A Practical Handbook of Speech Coders. CRC Press, Boca Raton, FL (2000)
2. Goswami, J.C., Chan, A.K.: Fundamentals of Wavelets: Theory, Algorithms and Applications. John Wiley & Sons, New York (1999)
3. Teolis, A.: Computational Signal Processing with Wavelets. Springer Verlag, New York (1998)
4. Burrus, C.S., Gopinath, R.A., Guo, H.: Introduction to Wavelets and Wavelet Transforms: A Primer. Prentice Hall, New Jersey (1997)
5. Marzetta, T.L.: A new interpretation for capon's maximum likelihood method of frequency–wavenumber spectral estimation. IEEE Trans. Acoustics, Speech, and Signal Processing **31** (1983)
6. Deller, J.R., Hansen, J.H.L., Proakis, J.G.: Discrete-Time Processing of Speech Signals. IEEE Press, New York (2000)
7. Donoho, D.L.: Denoising by soft-thresholding. IEEE Trans. Information Theory **41** (1995)
8. Abbate, A., Decusatis, C.M., Das, P.K.: Wavelets and Subband: Fundamentals and Applications. Birkhauser, Stuttgart, Germany (2001)
9. Ogden, R.T.: Essential Wavelets for Statistical Applications and Data Analysis. Springer Verlag, New York (1996)
10. Parsons, T.W.: Voice and Speech Processing. McGraw-Hill, New York (1986)
11. Furui, S.: Digital Speech Processing, Synthesis and Recognition. 2nd edn. Marcel Dekker, New York (2001)
12. Jurasfky, D., Martin, J.H., Linden, K.V., Jurafsky, D.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, New Jersey (2000)
13. Morgan, N., Gold, B.: Speech and Audio Signal Processing: Processing and Perception of Speech and Music. John Wiley & Sons, New York (1999)
14. Rabiner, L., Juang, B.H., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall, New Jersey (1993)
15. Huang, X., Acero, A., Hon, H.W., Reddy, R.: Spoken Language Processing. Prentice Hall, New Jersey (2001)
16. Quatieri, T.F.: Discrete-Time Speech Signal Processing. Prentice Hall, New Jersey (2001)

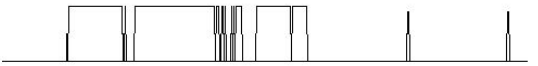


(a) a speech signal

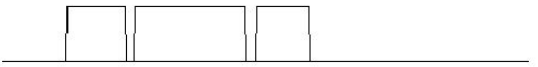
(b) a signal filtered by the DWT



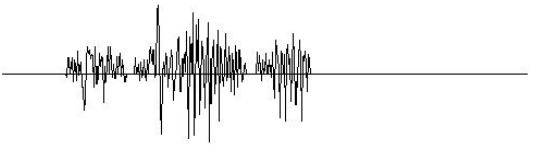
(c) a windowing AMDF



(d) a thresholding data



(e) an extraction mask

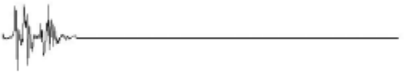


(f) an extracted signal

Fig. 4. Extraction.

(a) a speech signal

(b) a speech signal added with a noise of 80 Hz



(c) a speech signal classified and compressed by the merging algorithm

Fig. 5. Classification and compression from the signal added with a noise of 80 Hz

(a) a speech signal

(b) a speech signal added with a noise of 6 kHz

(c) a speech signal classified and compressed by the merging algorithm

Fig. 6. Classification and compression from the signal added with a noise of 6 kHz