

## 웨이브렛 변환을 이용한 음성신호의 유성음/무성음/묵음 분류\*

Voiced/Unvoiced/Silence Classification of Speech Signal  
Using Wavelet Transform

손영호 · 배건성\*\*

(Young-Ho Son · Keun-Sung Bae)

## ABSTRACT

Speech signals are, depending on the characteristics of waveform, classified as voiced sound, unvoiced sound, and silence. Voiced sound, produced by an air flow generated by the vibration of the vocal cords, is quasi-periodic, while unvoiced sound, produced by a turbulent air flow passed through some constriction in the vocal tract, is noise-like. Silence represents the ambient noise signal during the absence of speech. The need for deciding whether a given segment of a speech waveform should be classified as voiced, unvoiced, or silence has arisen in many speech analysis systems. In this paper, a voiced/unvoiced/silence classification algorithm using spectral change in the wavelet transformed signal is proposed and then, experimental results are demonstrated with our discussions.

**Keywords:** voiced sound, unvoiced sound, sound classification, wavelet transform

## I. 서론

음성신호는 음성이 존재하지 않는 배경잡음에 해당하는 묵음(silence)구간과 음성이 존재하는 실제의 음성신호 구간으로 분류된다. 음성신호는 인간의 발성기관에서 2개의 얇은 막으로 구성된 성대(vocal folds)의 운동에 의해 발생하는 공기의 흐름이 성도(vocal tract)를 지나면서 변조되어 공기압의 파동형태로 나타나는 것이다. 이러한 음성신호는 성대를 통과한 공기흐름의 성질에 따라 유성음(voiced

---

\* 이 연구는 한국과학재단의 핵심전문연구비 지원에 의한 것임.

\*\* 경북대학교 전자·전기공학부

sound)과 무성음(unvoiced sound)으로 나눌 수 있다. 일반적으로 유성음은 허파에서 방출되는 공기가 성문을 통과하면서 성대를 규칙적으로 진동하게 함으로써 만들어지는 준주기적인 펄스형태의 공기흐름에 의하여 생성된다. 이로 인하여 유성음은 파형이 준주기적이며 무성음에 비해 상대적으로 저주파 대역에 에너지가 집중되어 나타난다. 반면에 무성음의 경우는 성대의 진동 없이 성도의 수축된 지점을 통과하는 거친(turbulent) 공기흐름에 의하여 생성되어 잡음처럼 파형이 불규칙적이고 상대적으로 고주파 대역에 에너지가 집중되어 나타난다[1]. 음성신호처리에서 이와 같은 유성음/무성음/목음의 정확한 분류는 피치검출, 음성검출, 음성합성, 음성 세그멘테이션 등의 음성분석 및 음성인식 분야에서 중요하게 다루어지는 문제이며 실제로 음성신호의 유성음/무성음/목음 분류를 위한 다양한 방법들이 제안되어져 왔다. 이 방법들은 크게 두 종류로 구분될 수 있다. 첫째는 피치분석과 관련한 음성신호의 유성음/무성음/목음 분류 방법이고, 둘째는 일종의 형태인식(pattern recognition) 방법으로 해당 음성구간을 대상으로 에너지, 영교차율 등의 다양한 파라미터 값들을 검출한 후 이를 토대로 한 분류방법이다[2~5]. 전자의 경우는 해당 음성구간에서 음성신호의 주기성을 파라미터로 검출하여 유성음/무성음/목음 분류를 하는 방법으로 음성분류를 위하여 피치검출을 전제로 함으로써 음성분류에 많은 시간이 걸리게 된다. 또 피치검출을 위하여 30~40 ms 길이의 음성 단위로 분석함으로써 유/무성음이 혼재된 구간에서 정확한 분류에 어려움이 있다. 후자의 경우는 이러한 단점을 보완하여 유성음/무성음/목음의 분류에서 피치검출을 제외하고 대신에 다양한 다른 파라미터들을 사용함으로써 피치를 파라미터로 사용하는 경우에 비해 상대적으로 짧은 음성구간을 대상으로 한 분석이 가능하여 보다 신뢰성 있는 음성신호의 유성음/무성음/목음 분류가 가능하다[4,5].

본 논문에서는 시간 및 주파수 영역에서 동시에 음성신호의 국부적인 특성을 잘 반영하는 웨이브렛 변환을 이용한 음성신호의 유성음/무성음/목음 분류 알고리즘을 연구하였다. 제안하는 알고리즘에서는 음성신호의 웨이브렛 변환된 신호에서 스펙트럼상의 변화를 검출하기에 적합한 MLR(maximum likelihood ratio)[6,7] 값을 이용하여 음성분류 파라미터로 사용하였다. 본 논문의 구성은 다음과 같다. 먼저 2장에서는 웨이브렛 변환에 대하여 소개하며 3장에서는 MLR을 이용한 유성음/무성음/목음 분류 파라미터와 알고리즘에 대하여 설명한다. 그리고 4장에서는 실험한 결과를 수동으로 검출한 각 구간의 경계와 비교 분석하여 5장에서 결론을 맺는다.

## II. 웨이브렛 변환(Wavelet Transform)

웨이브렛 이론은 응용수학에서 처음 소개된 후 최근 컴퓨터비전 분야 등에서 연구되어 온 다중 해상도 표현과 연관성이 있음이 밝혀진 신호해석 방법이다. 일반적으로 웨이브렛 변환은 모 웨이브렛(mother wavelet)의 확장(dilation) 및 축소

(contraction) 등의 과정을 통하여 특정 주파수 대역의 신호 성분을 분리하여 해석하는 것을 가능하게 한다. 특히 고주파 대역에서는 축소된 웨이브렛을 이용하여 시간영역에서의 분해능을 개선시키고, 저주파 대역에서는 확장된 웨이브렛을 이용하여 주파수 영역에서의 분해능을 향상시킬 수 있다. 실제 응용에 이용되는 이산 웨이브렛 변환식은 식 (1), (2)로 주어지며 이러한 이산 웨이브렛 변환 이론은 이산 신호의 subband 분할 방법과도 연관성이 존재한다. 한편 식 (2)의 이산 웨이브렛 변환식에서 계수 구현을 용이하게 하기 위하여  $a_0$ 를 2로 한 dyadic 웨이브렛 변환(DyWT: Dyadic Wavelet Transform)은 식 (3)과 같이 정의된다. 식 (1)~(3)에서  $d$ 는 이산변환을 가리키며,  $j$ 는 scale,  $k$ 는 translation,  $T$ 는 translation 간격을 나타낸다.

$$d_{j,k} = \int f(t) \varphi_{j,k}^*(t) dt \quad (1)$$

$$\varphi_{j,k}(t) = a_0^{-\frac{j}{2}} \varphi(a_0^{-j}t - kT) \quad (2)$$

$$d_{j,k} = \frac{1}{\sqrt{2^j}} \int f(t) \varphi^*\left(\frac{t}{2^j} - kT\right) dt \quad (3)$$

DyWT을 이용하여 신호를 분석하는 과정은 그림 1과 같이 트리(tree)형태로 구성된 필터뱅크의 출력과 같다. 여기서  $H_0$ 는 저역통과 필터이고,  $H_1$ 은 고역통과 필터이다. 입력 신호가 저역통과 필터와 고역통과 필터를 차례로 거치게 되면 한 번의 웨이브렛 변환이 이루어지며, 저역통과 필터를 통과한 신호에 대해 이러한 과정을 반복적으로 수행하여 웨이브렛 변환된 신호를 얻을 수 있다. 신호처리 관점에서 DyWT은 constant-Q, octave band 특성을 갖는 밴드패스 필터들의 뱅크 출력으로 볼 수 있다[8]. 본 연구에서는 식 (2)에서의 웨이브렛 함수로 quadratic spline[9] 웨이브렛을 사용하였으며 quadratic spline 웨이브렛에 사용된 필터의 계수는 표 1과 같다. 그림 2는 10 kHz 샘플링 주파수에서 본 연구에서 사용한 웨이브렛 함수의 스케일 2~5에 해당하는 주파수 응답을 나타낸 것으로 스케일이 증가할수록 주파수 영역에서 중심 주파수가 저주파 대역으로 접근함을 볼 수 있다. 이러한 스케일에 따른 웨이브렛 함수의 주파수 응답 특성은 음성신호 분류시 음성신호의 주파수 영역에서의 특성상 유성음의 경우는 높은 스케일의 신호에서, 무성음의 경우는 낮은 스케일의 신호에서 잘 검출될 수 있다는 것을 의미한다. 한편 본 연구에서는 웨이브렛 변환 시에 그림 1과 같이 웨이브렛 변환된 신호를 스케일 1에서부터 순차적으로 구하는 방법 대신에 각 스케일에 해당하는 웨이브렛 함수를 직접 구현하여 음성신호를 웨이브렛 변환함으로써 필요한 스케일의 신호를 직접 얻는 방법을 사용하였다.

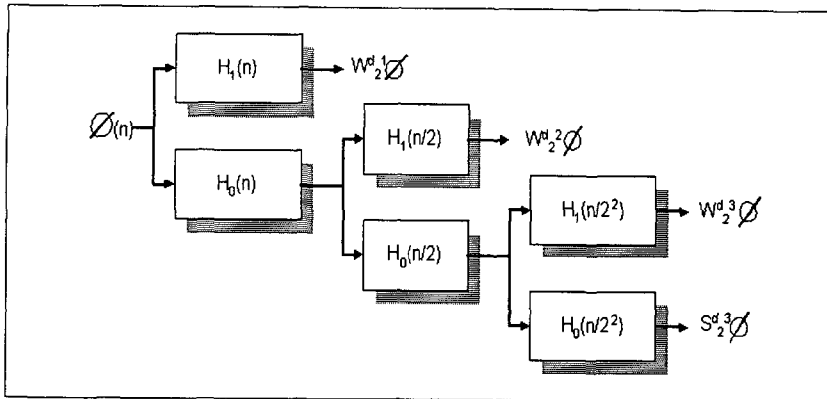


그림 1. 이산 웨이브렛 변환에 대한 dyadic 필터뱅크 구조

표 1. quadratic spline 웨이브렛의 필터 계수

n	$H_0$	$H_1$
-1	0.125	
0	0.375	-2.0
1	0.375	2.0
2	0.125	

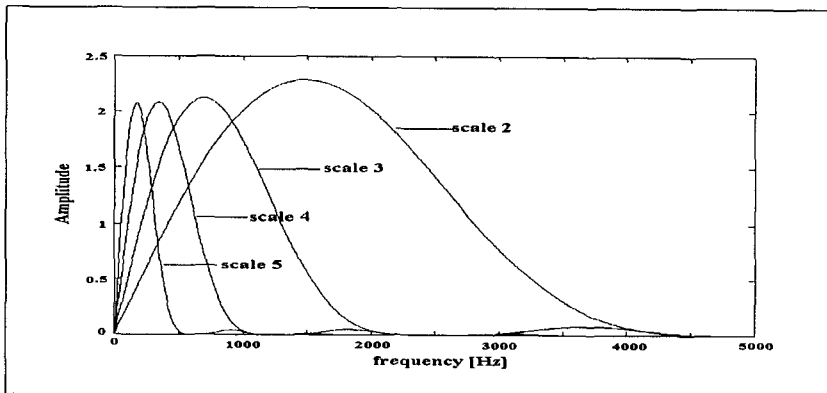


그림 2. quadratic spline 웨이브렛 함수의 주파수 특성

### III. 웨이브렛 변환을 이용한 음성신호의 분류

#### 3.1 음성분류 파라미터

음성신호의 해당 구간을 유성음/무성음/목음 등으로 분류할 때 어려운 점은 무성음을 유성음이나 목음구간과 구분하는 것이다. 특히·진폭이 작은 마찰음(fricative)이나 파열음(plosive)의 경우는 단순히 시간영역의 에너지나 영교차율만으로 배경잡음에 해당하는 목음구간과 정확히 구분하기 어렵다. 본 연구에서는 음성신호의 분류를 위하여 주파수 영역에서 신호의 변화를 잘 모델링하는 것으로 알려진 MLR(maximum likelihood ratio)을 웨이브렛 변환한 신호에 적용하였다. 각 스케일에 해당하는 프레임별 MLR에 대한 정의는 식 (4)와 같다. 식 (4)에서  $MLR_k$ 는 스케일 k에서의 MLR 값을 가리키고,  $\sigma_k^2$ 와  $\sigma_{noise}^2$ 는 각각 해당 프레임에 대하여 스케일 k에서 웨이브렛 변환한 신호의 분산과 배경잡음에 해당하는 목음구간의 분산을 나타낸다. 특히  $\sigma_{noise}^2$ 의 경우는 음성신호의 처음 몇 구간을 목음구간이라 가정하여 구한 분산의 평균값으로 한다. N은 MLR을 구할 때 동시에 분석되는 프레임 수로서 1로 설정하였다.

$$MLR_k = \frac{N}{2} \left| \ln \left( \frac{\sigma_k^2}{\sigma_{noise}^2} \right) - \left( \frac{\sigma_k^2}{\sigma_{noise}^2} \right) \right| \quad (4)$$

그림 3은 음성신호 /fish/의 음성파형과 스케일 1과 4에서의 MLR 값을 프레임 단위로 나타낸 것으로 웨이브렛 함수의 주파수 응답 특성에서 예상했던 것처럼 스케일 1에서는 유성음 /i/와 함께 상대적으로 무성음 /f/, /s/가 스케일 4에서는 유성음 /i/가 잘 나타나고 있음을 볼 수 있다. 본 연구에서는 MLR 값의 이러한 특성을 이용하여 유성음과 무성음을 각각 잘 반영하는 파라미터들을 동시에 이용하여 일차적으로 음성신호를 음성구간과 목음구간으로 분류한 뒤 검출된 음성구간을 최종적으로 유성음과 무성음으로 분류하였다. 음성구간을 검출하기 위한 파라미터로 식 (5)에 주어진 PS를 정의하였으며 유성음과 무성음을 구분하기 위한 파라미터로는 식 (6), (7)에 주어진  $PU_1$ ,  $PU_2$ 를 정의하였다. PS의 경우는 그림 3에서 나타난 것처럼 무성음을 잘 반영하는  $MLR_1$ 과 유성음을 잘 반영하는  $MLR_4$  값들을 더함으로써 유성음과 무성음의 조합으로 이루어진 음성구간을 목음구간과 분류하는 파라미터이며,  $PU_1$ 과  $PU_2$ 의 경우는 무성음이 유성음에 비해 상대적으로 고주파 대역에 에너지가 집중되어 있는 특성을 이용하여 무성음 구간에서 큰 값을 갖도록 하여 무성음을 유성음과 구분할 수 있게 한 파라미터이다. 식 (5), (6), (7)에서  $MLR_k$ 는 k번째 스케일에서의 MLR 값을 나타낸다.

$$PS = MLR_1 + MLR_4 \quad (5)$$

$$PU_1 = \frac{MLR_1}{MLR_4} \quad (6)$$

$$PU_2 = 100 \times \frac{MLR_1}{MLR_1 + MLR_2 + MLR_3} \quad (7)$$

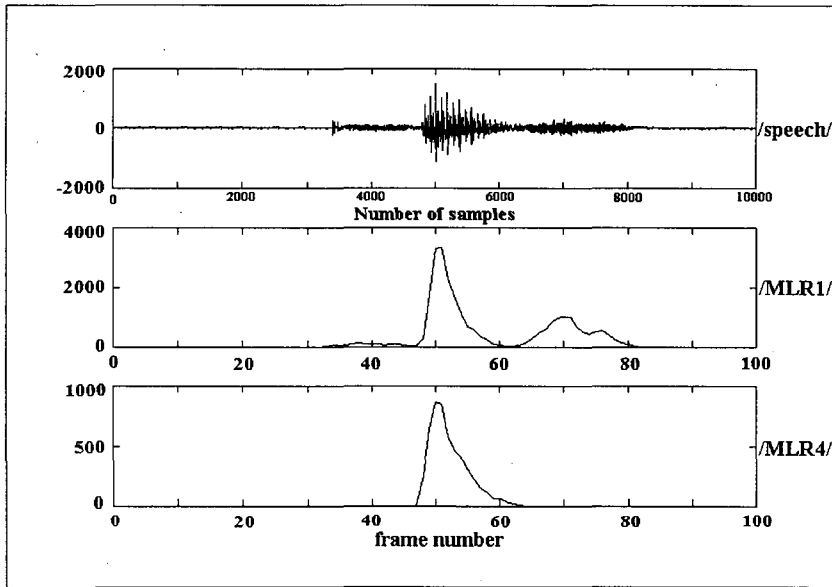


그림 3. 음성신호 /fish/에 대한 MLR 값의 변화

그림 4는 단어 /fish/를 대상으로 식 (5), (6), (7)에서 정의한 파라미터 값을 프레임 단위로 추출하여 나타낸 것이다. 이 결과에서 PS의 경우는 묵음구간에 비해 음성구간에서 충분히 큰 값을 가지게 되며, 또 무성음에 해당하는 /f/와 /ʃ/부분에서  $PU_1$ 과  $PU_2$  값이 유성음이나 묵음구간에 비해 상당히 크게 나타남을 확인할 수 있다.

### 3.2 웨이브렛 변환을 이용한 음성 분류 알고리즘

본 논문에서 제안하는 웨이브렛 변환을 이용한 음성신호의 유성음/무성음/묵음 분류 알고리즘은 다음과 같다.

STEP 1: 음성을 프레임 단위로 받아들여 스케일 1~4까지 웨이브렛 변환한다. 이때 분석 프레임의 길이는 20 ms로 하였으며, 각 스케일에서의 필터링으로 인해 생기는 지연(delay)을 보상해 주었다.

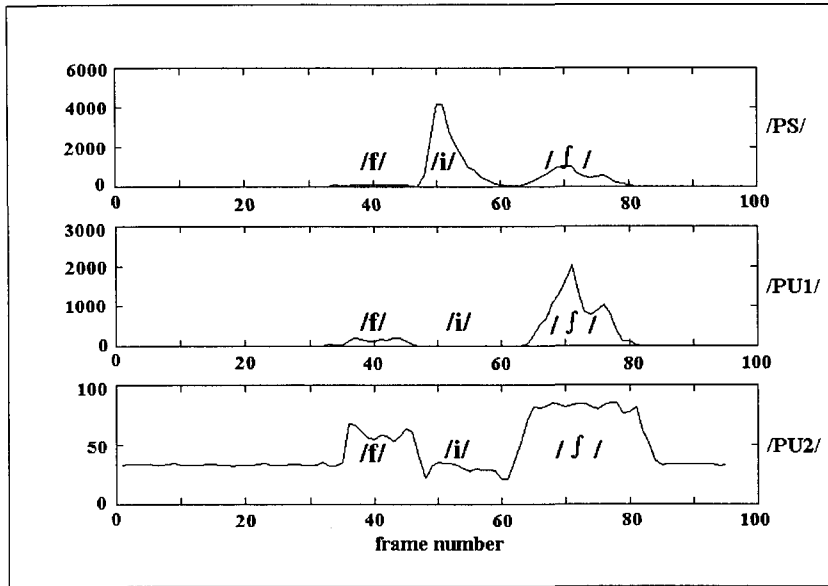


그림 4. 음성신호 /fish/에 대한 파라미터 PS, PU<sub>1</sub>, PU<sub>2</sub>의 변화

- STEP 2: Step 1의 출력신호를 이용하여 스케일별로 MLR 값을 계산하여 제안한 파라미터들을 계산한다. 한편 음성신호의 앞부분 150 ms에 대하여서는 묵음을 가정하여 각 파라미터들의 문턱값을 설정하는 단계를 거친다.
- STEP 3: 음성구간과 묵음구간을 구분하는 단계로 PS값이 설정된 문턱값 보다 클 경우 음성구간으로 간주하고 그렇지 않을 경우는 묵음구간으로 간주한다.
- STEP 4: 음성구간에서 유성음과 무성음을 구분하는 단계로 Step 3에서 음성구간으로 결정된 경우 PU<sub>1</sub>, PU<sub>2</sub>를 문턱값과 비교하여 문턱값보다 클 경우 무성음으로 간주하고 그렇지 않을 경우 유성음으로 간주한다.
- STEP 5: 한 프레임의 분석이 끝나면 다음 프레임으로 이동하여 Step 1~4를 반복한다. 이때 프레임의 이동 간격은 10 ms로 하였다.

본 연구에서는 각 파라미터에 대한 문턱값의 설정을 위하여 주어진 음성신호의 처음 150 ms 정도를 묵음구간이라 간주하고 이 구간을 대상으로 파라미터 PS, PU<sub>1</sub>, PU<sub>2</sub>에 대한 평균값을 구한 후 스케일 인자를 곱하여 최종적인 문턱값으로 설정하였다. 실험적으로 PS의 경우는 묵음구간 평균값의 5배를, PU<sub>1</sub>과 PU<sub>2</sub>의 경우는 각각 묵음구간 평균값의 10배와 1.5배를 문턱값으로 설정하였다. 한편 Step 3에서 검출한 음성구간의 길이가 100 ms 이내인 경우는 nonstationary한 잡음 등으로 인하여 묵음구간이 음성구간으로 잘못 분류된 것으로 간주하여 해당 구간을 묵음구간으로 처리하였다.

## IV. 실험 및 검토

본 연구에서 실험에 사용한 음성 데이터는 연구실 환경에서 남자 3, 여자 3명의 화자들로부터 5개의 문장을 대상으로 수집되어진 10 kHz로 샘플링 되고 16비트로 양자화된 음성들을 대상으로 하였다. 한편 제안한 알고리즘으로 검출한 유성음/무성음/목음구간의 경계들은 음성파형을 관찰하며 수동으로 검출한 경계들과 비교 제시함으로써 제안한 알고리즘의 성능을 비교하도록 하였다.

그림 5, 6은 단어 /fish/에 대한 유성음/무성음/목음 분류 결과를 제시한 것이다. 특히 그림 5의 경우는 검출된 음성구간에서  $PU_1$ 과  $PU_2$ 를 이용하여 분류한 유성음과 무성음의 경계를 무성음 /f/, /ʃ/를 중심으로 수동으로 분류한 실제 경계와 비교하여 나타낸 것이다. 그림에서 위쪽 방향의 점선은 수동으로 분류한 각 구간의 실제 경계를, 아래쪽 방향의 실선은 제안한 알고리즘으로 분류한 경계를 나타낸 것이다. 그림 6은 전체 음성에 대한 분류 결과를 나타낸 것으로 위쪽의 실선은 음성구간의 경계를, 아래쪽의 실선은 음성구간에서 유성음과 무성음의 경계를 나타내고 있다. 음소에 따라 다소의 차이는 있으나 분류된 경계의 대부분은 실제의 경계와 비교하여 50~100샘플내의 차이를 보였으며, 이 결과는 음성 분류시 200샘플 길이의 프레임 단위로 분석한 것을 감안할 때 제안한 알고리즘이 유성음/무성음/목음구간을 큰 오차 없이 신뢰성 있게 검출한 것으로 볼 수 있다.

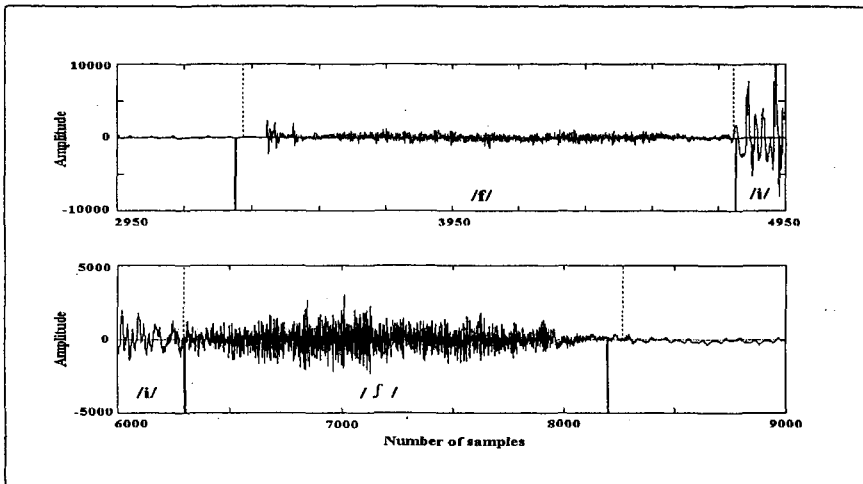


그림 5. 음소의 실제 경계와 검출된 경계와의 비교 /"fish"/



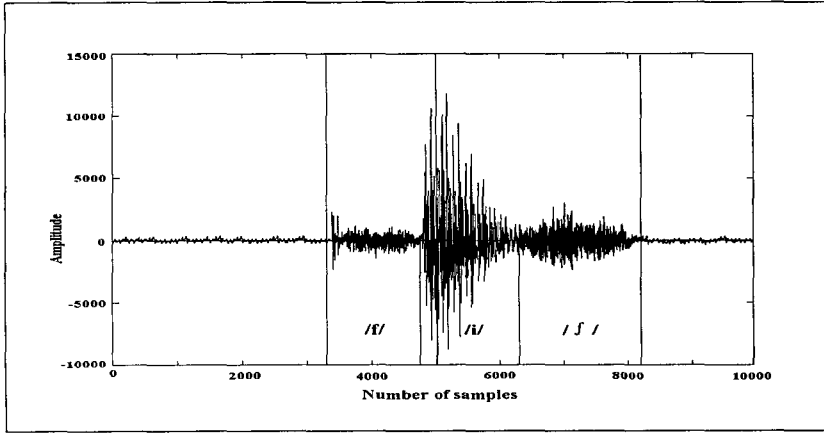
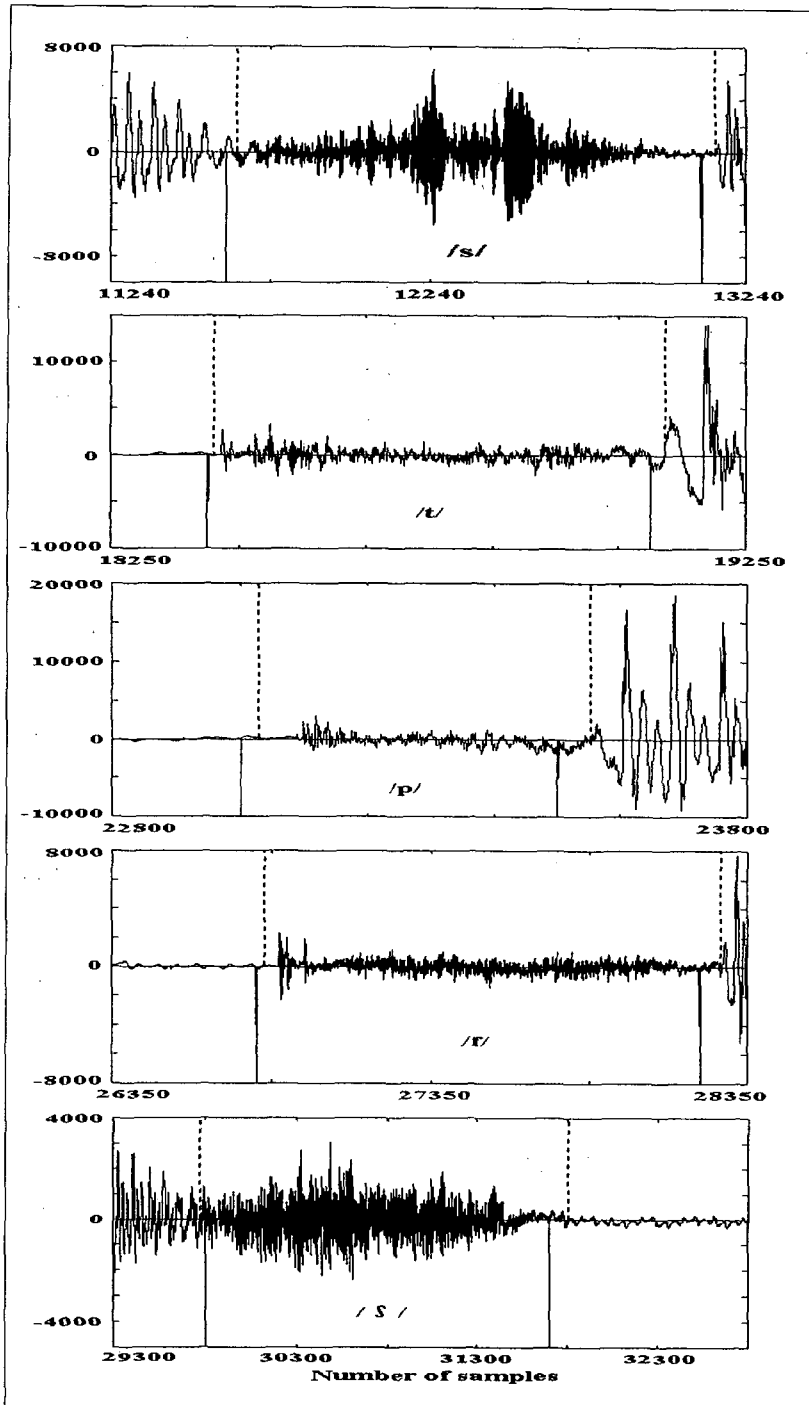


그림 6. 단어에 대한 유성음/무성음/묵음 분류 결과 /"fish"/

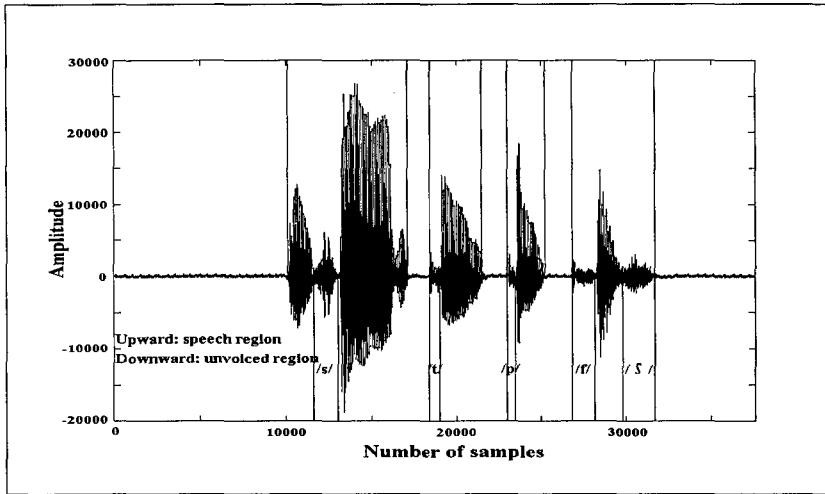
그림 7, 8은 제안한 알고리즘이 단어가 아닌 문장 내에서도 유성음/무성음/묵음 구간을 잘 분류할 수 있는지를 확인하기 위하여 "We saw the ten pink fish."라는 실제 문장을 대상으로 분류 실험한 결과를 나타낸 것이다. 그림 7에서는 문장 내에서 제안한 알고리즘으로 분류한 각 구간의 경계를 그림 5에서와 같이 무성음 /s/, /t/, /p/, /f/, /ʃ/ 등의 경계를 중심으로 수동으로 분류한 실제 경계와 비교하여 나타낸 것이다. 그림 8에서는 전체 음성신호에 대하여 분류한 경계를 나타내고 있다. 위쪽의 실선은 묵음구간과 음성구간을 분류한 경계이며, 아래쪽의 실선은 음성구간에서 유성음과 무성음 구간을 분류한 경계이다. 그림 9는 "Should we chase those cowboys?"라는 다른 문장에 대한 분류 결과를 그림 8과 같은 방법으로 제시하고 있다. 그림 7~9의 결과에서 볼 수 있는 것처럼 제안한 알고리즘이 문장을 대상으로 한 경우에도 단어를 대상으로 한 경우와 비슷한 결과를 보임을 확인할 수 있다.

그림 10은 그림 9에서 제시한 "Should we chase those cowboys?"라는 문장에서 발음상의 특성으로 주위 묵음구간과 비슷한 진폭을 가지는 마찰음 /ʃ/와 /ð/를 대상으로 제안한 알고리즘이 해당구간을 어느 정도 잘 분류할 수 있는지를 조사한 결과이다. 그림에서 실선은 제안한 알고리즘으로 검출한 경계이며 점선은 파형을 관찰하며 수동으로 검출한 실제의 경계에 해당한다. 그림에서 검출한 경계와 실제 경계의 차이를 볼 때 제안한 알고리즘이 배경잡음과 비슷한 진폭을 가지는 마찰음의 경우도 잘 분류함을 확인할 수 있다.



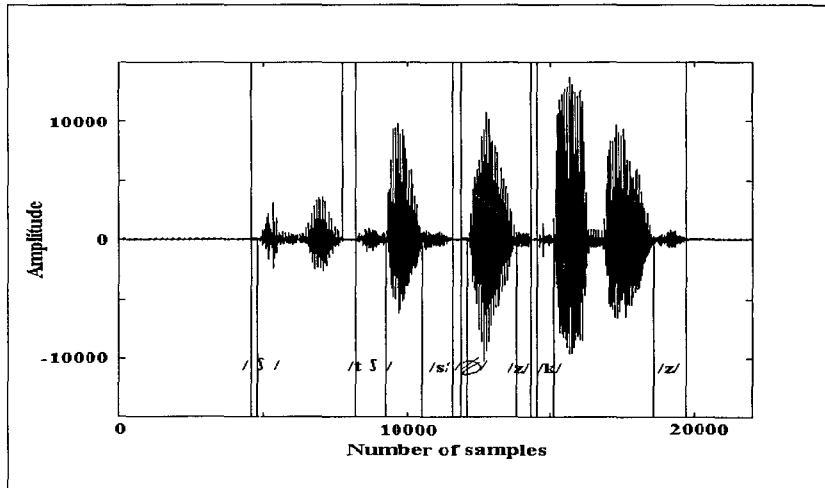
/"We saw the ten pink fish."/

그림 7. 음소의 실제 경계와 검출된 경계와의 비교



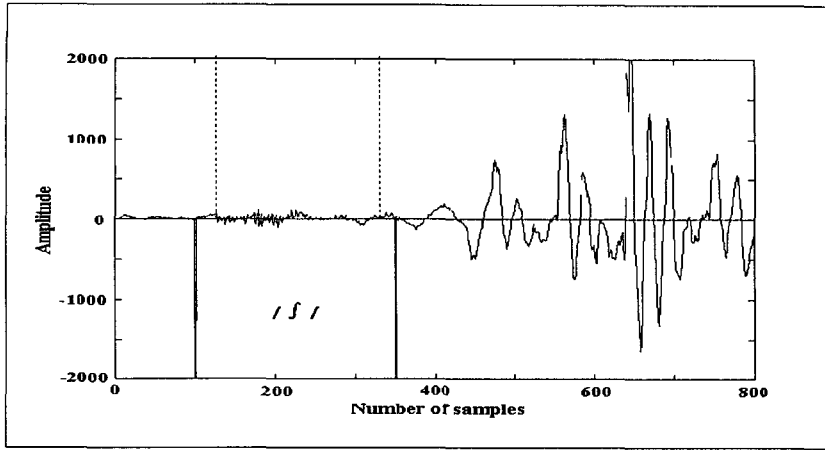
/"We saw the ten pink fish."/

그림 8. 문장에 대한 유성음/무성음/묵음 분류 결과

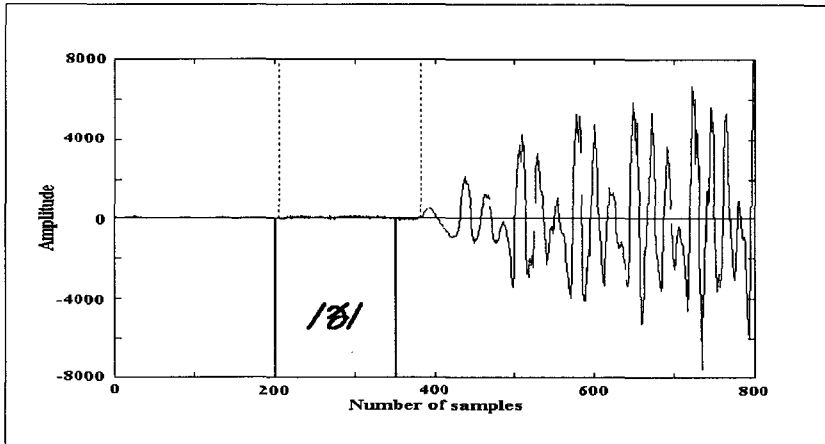


/"Should we chase those cowboys ?"/

그림 9. 문장에 대한 유성음/무성음/묵음 분류 결과(계속)



(a). should의 /s/



(b). those의 /ð/

그림 10. 배경잡음과 진폭이 비슷한 마찰음 분류 예

## V. 결 론

본 논문에서는 음성신호의 웨이블릿 변환된 신호에서 각 스케일별 MLR 값을 이용하여 음성신호의 주파수 영역에서의 특성을 반영하는 파라미터를 정의하고, 이를 이용하여 음성신호를 유성음/무성음/목음으로 분류하는 알고리즘을 제안하였다. 제안된 알고리즘으로 분류한 유성음/무성음/목음 구간들의 경계들은 음소들에 따라 다소의 차이는 있었으나 파형과 스펙트로그램(spectrogram)을 관찰하면서 수동으로

분류한 음소들의 실제 경계를 기준으로 대부분 50~100 샘플내의 차이를 보였다. 특히 진폭이 작은 마찰음이나 파열음의 경우에 있어서도 주변잡음으로부터 정확하게 분류될 수 있음을 확인할 수 있었다. 제안한 알고리즘으로 분류한 경계가 실제의 경계를 기준으로 이와 같은 차이를 보이는 것은 부분적으로는 프레임 단위의 분석에 기인하는 것으로 생각된다. 따라서 분석 프레임의 크기를 줄이거나 또는 프레임 단위 분석 후 분류된 유성음/무성음/목음 구간의 경계 프레임을 다시 세부구간으로 나누어 분석하는 등의 후처리 과정을 통해 제안된 알고리즘의 검출 성능을 개선할 수 있을 것으로 생각한다.

### 참 고 문 헌

- [1] Douglas O'Shaughnessy. 1987. "Speech Communication" Addison Wesley
- [2] Jean Rouat, Yong Chun Liu, Daniel Morissette. 1997. "A pitch determination and voiced/unvoiced decision algorithm for noisy speech." *Speech Communication* 21.
- [3] Hidefumi Kobatake. 1987. "Optimization of Voiced/Unvoiced Decisions in Nonstationary Noise Environments." *IEEE Trans on ASSP*, vol. 35, No.1.
- [4] Bishnu S.Atal, Lawrence R. Rabiner. 1976(6). "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition." *IEEE Trans on ASSP*, vol. 24, No. 3.
- [5] Leah J. Siegel. 1979(2). "A Procedure for Using Pattern Classification Techniques to Obtain a Voiced/Unvoiced Classifier." *IEEE Trans on ASSP*, vol. 27, No.1.
- [6] B. T Tan, R. Lang, Heiko Schroder, Andrew Spray. 1994. "Applying wavelet analysis to speech segmentation and classification." *SPIE Wavelet Applications*, vol. 2242.
- [7] V. Ralph Algazi, Kathy L. Brown, Michael J. Ready, David H. Irvine. 1993. "Transform Representation of the Spectra of Acoustic Speech Segments with Applications-I." *IEEE Trans Speech and Audio Processing*, vol. 1, No.2.
- [8] Olibier Rioul, Martin Vetterli. 1991(10). "Wavelets and Signal Processing." *IEEE SP Magazine*.
- [9] S. Mallat, W. L. Hwang. 1992. "Singularity Detection and Processing with Wavelets." *IEEE Trans Information Theory*, vol. 38, no.2.

접수일자 : '98. 9. 15.

게재결정 : '98. 10. 28.

## ▲ 손영호

대구광역시 북구 산격동 1370(우 : 702-010)

경북대학교 전자·전기공학부

Tel: (053) 940-8627, Fax: (053) 950-5527

e-mail: youngho@mmir11.kyungpook.ac.kr

## ▲ 배건성

대구광역시 북구 산격동 1370(우 : 702-010)

경북대학교 전자·전기공학부

Tel: (053) 950-5527, Fax: (053) 950-5527

e-mail: ksbae@ee.kyungpook.ac.kr